ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΔΙΑΧΕΙΡΙΣΗΣ ΚΑΙ ΒΕΛΤΙΣΤΟΥ
ΣΧΕΔΙΑΣΜΟΥ ΔΙΚΤΥΩΝ ΤΗΛΕΜΑΤΙΚΗΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ»

Μελέτη Βιολογικών Δικτύων με χρήση Ενσωμάτωσης σε χώρους Υπερβολικής Γεωμετρίας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χαρά Β. Μαστρόκαλου

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής, Ε.Μ.Π.

Αθήνα, Οκτώβρης 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ
ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΔΙΑΧΕΙΡΙΣΗΣ ΚΑΙ ΒΕΛΤΙΣΤΟΥ
ΣΧΕΔΙΑΣΜΟΥ ΔΙΚΤΥΩΝ ΤΗΛΕΜΑΤΙΚΗΣ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ»

Μελέτη Βιολογικών Δικτύων με χρήση Ενσωμάτωσης σε χώρους Υπερβολικής Γεωμετρίας

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χαρά Β. Μαστρόκαλου

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής, ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Οκτωβρίου 2021

.....	.....	.....
Συμεών Παπαβασιλείου	Θεοδώρα Βαρβαρίγου	Βασίλειος Καρυώτης
Καθηγητής, Ε.Μ.Π.	Καθηγήτρια, Ε.Μ.Π.	Αναπληρωτής Καθηγητής, Ιόνιο Πανεπιστήμιο

Αθήνα, Οκτώβρης 2021

.....

Χαρά Β. Μαστρόκαλου
Πτυχιούχος Βιολογίας ΕΚΠΑ

Abstract

The objective of this thesis is the comparative study of graph embedding algorithms in the Hyperbolic Space. In this context, the protein-protein interactions network of *Homo sapiens* is also studied. Real complex networks, among which, one may encounter many biological, social and technological networks, take advantage of graphs' topology as the means for their visualization, analysis and modeling. Some often displayed commonalities of such complex networks are a scale-free distribution, small-world behavior, heterogeneity and self-similarity. The heterogeneity between the elements of a network implies the existence of some kind of taxonomy, as well a classification between groups and subgroups of those elements, that approximates a hierarchical tree-like structure. Another well known domain with hierarchical organization is the space of Hyperbolic Geometry. From that shared metric structure arises the fact that Hyperbolic Geometry underlies the complex networks, while at the same time, networks generated in the Hyperbolic Space display heterogeneity and strong clustering as a result of its negative curvature. Several models attempted to reproduce the evolution of these networks, given the existence of an underlying Hyperbolic Geometry shaping their structure. One of these models, is the Popularity-Similarity Optimization (PSO) model, that includes and optimizes the trade-off between two measures of attractiveness: node popularity and similarity between nodes. The geometric interpretation of these measures condenses in the distance between nodes in the Hyperbolic Space, while their connection probability is captured as a decreasing function of that distance. Subsequently, many algorithms for network hyperbolic embedding, some of which will be considered in this thesis, rely their methodology on the aforementioned PSO model. The advantages and limitations of these algorithms will be examined, and their embedding results will be evaluated in artificial and real networks. Lastly, in the case of the *H.sapiens* protein-protein interactions (PPIs) network, the idea that the hyperbolic distance has a significant impact on the formation of edges between nodes will be examined using semantic similarity as a criterion, while the potential association between closeness in the Hyperbolic Space and functional relevance of proteins will also be assessed.

Keywords: Complex Networks, Hyperbolic Geometry, Network Embedding, Protein-Protein Interactions Network, Hyperbolic Space, Big Data

Περίληψη

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η συγκριτική μελέτη τεχνικών ενσωμάτωσης δικτύων στον Υπερβολικό Χώρο. Στο πλαίσιο αυτό μελετάται και το δίκτυο πρωτεϊνικών αλληλεπιδράσεων του *Homo sapiens*. Η τοπολογία των γράφων χρησιμοποιείται για την απεικόνιση, την ανάλυση και τη μοντελοποίηση των πραγματικών σύνθετων δικτύων, μεταξύ των οποίων συναντά κανείς βιολογικά, κοινωνικά και τεχνολογικά δίκτυα. Ορισμένα κοινά χαρακτηριστικά των σύνθετων δικτύων είναι η κατανομή βαθμού άνευ-κλίμακας, τα φαινόμενα μικρού κόσμου, η ετερογένεια και η αυτο-ομοιότητα. Η ετερογένεια μεταξύ των στοιχείων ενός δικτύου συνεπάγεται την ύπαρξη μιας μορφής ταξινόμησης των στοιχείων αυτών, μιας οργάνωσης τους, η οποία μπορεί να χαρακτηριστεί ως ιεραρχική δομή. Ένα πεδίο στο οποίο επίσης συναντάται αυτός ο τρόπος ιεραρχικής οργάνωσης είναι ο χώρος της Υπερβολικής Γεωμετρίας. Από την ύπαρξη αυτού του κοινού στοιχείου δομής μεταξύ τους, προκύπτει ότι η Υπερβολική Γεωμετρία είναι υποκείμενη των σύνθετων δικτύων, και ταυτοχρόνως ότι τα δίκτυα που δημιουργούνται εντός του Υπερβολικού Χώρου παρουσιάζουν χαρακτηριστικά ετερογένειας και ισχυρής ομαδοποίησης ως αποτέλεσμα της αρνητικής καμπυλότητας του. Αρκετά μοντέλα σχεδιάστηκαν για να μελετηθεί η δημιουργία και η εξέλιξη των σύνθετων δικτύων, βασισμένα στην ύπαρξη μιας υποκείμενης Υπερβολικής Γεωμετρίας που διαμορφώνει τη δομή αυτών. Ένα εξ' αυτών είναι το μοντέλο Βελτιστοποίησης Δημοτικότητας-Ομοιότητας. Σε αυτό περιλαμβάνονται δυο μετρικές, η δημοτικότητα των κόμβων και η ομοιότητα μεταξύ των κόμβων ενός δικτύου, τις οποίες το μοντέλο προσπαθεί να εξισορροπήσει κατά βέλτιστο τρόπο. Η γεωμετρική ερμηνεία των μετρικών αυτών αποτυπώνεται στην απόσταση μεταξύ των κόμβων στο υπερβολικό επίπεδο και στην απόδοση της πιθανότητας σύνδεσης αυτών, ως φθίνουσα συνάρτηση της υπερβολικής τους απόστασης. Πολλοί αλγόριθμοι για την ενσωμάτωση δικτύων στον Υπερβολικό Χώρο βασίζουν τη μεθοδολογία τους στο προαναφερθέν μοντέλο Βελτιστοποίησης Δημοτικότητας-Ομοιότητας. Στα πλαίσια της παρούσας διπλωματικής εργασίας, εξετάζονται ορισμένοι από αυτούς τους αλγόριθμους. Παρουσιάζονται τα πλεονεκτήματα και οι αδυναμίες τους, ενώ τα αποτελέσματα ενσωμάτωσης δικτύων αξιολογούνται τόσο σε συνθετικά όσο και σε πραγματικά δίκτυα. Τέλος, στην περίπτωση του πρωτεϊνικού δικτύου αλληλεπιδράσεων του *H.sapiens*, η υπόθεση ότι η απόσταση μεταξύ κόμβων στον Υπερβολικό Χώρο έχει σημαντική επίπτωση στη δημιουργία δεσμών, ελέγχεται χρησιμοποιώντας ως κριτήριο τη σημασιολογική ομοιότητα, ενώ αξιολογείται και η ενδεχόμενη σχέση μεταξύ της υπερβολικής απόστασης και της λειτουργικής συσχέτισης πρωτεϊνών.

Λέξεις Κλειδιά: Σύνθετα Δίκτυα, Υπερβολική Γεωμετρία, Ενσωμάτωση Δικτύου, Δίκτυα Πρωτεϊνικών Αλληλεπιδράσεων, Υπερβολικός Χώρος, Μεγάλα Δεδομένα

Εκτεταμένη Περίληψη

Η τεχνολογική πρόοδος των τελευταίων ετών έχει οδηγήσει σε μια εξαιρετικά μεγάλη συλλογή δεδομένων από σχεδόν κάθε πτυχή της ζωής, δημιουργώντας συστήματα διαφορετικών διασυνδεδεμένων οντοτήτων. Η επιστήμη των Δεδομένων, και ιδιαίτερα εκείνη της ανάλυσης Μεγάλων Δεδομένων, συμβάλλει στη διαχείριση αυτών, συνεισφέροντας επιπλέον στην κατανόηση πολύπλοκων συστημάτων σε διάφορους επιστημονικούς κλάδους, καθώς και στην επίλυση σύνθετων ερωτημάτων και προκλήσεων, των οποίων η προσέγγιση ήταν αδύνατη πριν την εποχή των Μεγάλων Δεδομένων. Μια από τις σύγχρονες προκλήσεις στο χώρο της Βιολογίας είναι η κατανόηση της κυτταρικής λειτουργίας και των αιτιωδών εκείνων συμβάντων και μηχανισμών της, που οδηγούν στην εμφάνιση παθολογικών καταστάσεων και ασθενειών, και κατ' επέκταση ο εντοπισμός υποψήφιων στόχων και μηχανισμών για την αντιμετώπισή τους, είτε στο επίπεδο του πληθυσμού είτε στο επίπεδο της προσωποποιημένης θεραπείας. Στις προκλήσεις αυτές, η εφαρμογή μεθοδολογιών από τον κλάδο της ανάλυσης Μεγάλων Δεδομένων είναι απαραίτητη, τόσο για την ανάλυση του μεγάλου όγκου δεδομένων που παράγονται από τα σύγχρονα βιολογικά πειράματα (ομικά δεδομένα), όσο και για την ερμηνεία των αποτελεσμάτων με βάση την ολοένα αυξανόμενη πρότερη βιολογική γνώση, όπως αυτή οργανώνεται κατάλληλα σε βάσεις δεδομένων. Στο σύνολο τους οι διαφορετικοί τύποι ομικών δεδομένων, συμπεριλαμβανομένης της γονιδιωματικής, της μεταγραφομικής, της πρωτεομικής και της μεταγονιδιωματικής, αποτυπώνουν διαφορετικές πτυχές της κυτταρικής λειτουργικότητας. Καθώς αυξάνεται ο όγκος ομικών δεδομένων, αυξάνεται τόσο η πολυπλοκότητά τους όσο και η δυσκολία ανάλυσης και κατανόησης τους. Η ανάπτυξη κατάλληλων μαθηματικών μοντέλων και υπολογιστικών συστημάτων είναι σημαντική για τη μείωση της πολυπλοκότητας και τη διαχείριση των δεδομένων, με σκοπό να ξεπεραστούν υπάρχουσες τεχνικές προκλήσεις και να απαντηθούν επιστημονικά ερωτήματα που θα οδηγήσουν εν τέλει στην απόκτηση νέας γνώσης.

Η επιμέρους μελέτη και συγκέντρωση γνώσης για τα συστατικά στοιχεία ενός συστήματος είναι αναμφίβολα σημαντική, ωστόσο, η μελέτη αυτών και των ρόλων τους ως μέρη ενός ευρύτερου συστήματος αλληλεπιδρώντων στοιχείων είναι εξίσου αναγκαία. Στην περίπτωση των σύνθετων δικτύων, ο τρόπος οργάνωσης τους αποτυπώνει τη λειτουργία τους, ενώ ο χαρακτηρισμός τους ως σύνθετα αιτιολογείται από την αδυναμία πρόβλεψης της συλλογικής τους συμπεριφοράς από τα επιμέρους συστατικά τους. Η αναγνώριση των αρχών που διέπουν αυτά τα συστήματα μπορεί να βοηθήσει στην πρόβλεψη των διαταραχών και του αντίκτυπου που θα προκαλέσουν οι όποιες αλλαγές στα ίδια ή στο περιβάλλον τους. Η αναπαράσταση των ίδιων των στοιχείων και των μεταξύ τους αλληλεπιδράσεων αλλά και η διερεύνηση της οργάνωσης του σύνθετου συστήματος που απαρτίζουν, μπορεί να επιτευχθεί με τη χρήση γράφων (δικτύων). Στα πραγματικά πολύπλοκα συστήματα περιλαμβάνονται μεταξύ άλλων βιολογικά, κοινωνικά και τεχνολογικά δίκτυα, τα οποία μοιράζονται ορισμένα στοιχεία δομής και οργάνωσης που δεν θεωρούνται ούτε

κανονικά (πλέγματος) ούτε τυχαία. Στα χαρακτηριστικά αυτά περιλαμβάνονται η κατανομή βαθμού άνευ-κλίμακας, τα φαινόμενα μικρού κόσμου, τα στοιχεία κοινότητας, η ετερογένεια/ανομοιογένεια μεταξύ των κόμβων και η ιεραρχική δομή. Η ετερογένεια εντός του δικτύου συνεπάγεται την ύπαρξη μιας μορφής ταξινόμησης και οργάνωσης των στοιχείων σε επίπεδο συνόλων και υπο-συνόλων, η οποία περιγράφεται ως ιεραρχική δενδροειδής δομή. Η προσπάθεια μοντελοποίησης αυτών των δομών στα πλαίσια ενός Ευκλείδιου χώρου είναι αρκετά περιορισμένη και προβληματική, αφού οι αποστάσεις μεταξύ των στοιχείων παραμορφώνονται σε σημαντικό βαθμό. Σε αντίθεση, ο Υπερβολικός Χώρος θα μπορούσε να εκληφθεί ως ανάλογο των δενδροειδών δομών, όπου επιτρέπεται η ενσωμάτωση με μικρότερα σφάλματα παραμόρφωσης. Ταυτόχρονα, μοντέλα και αλγόριθμοι υποστηρίζουν την ύπαρξη της Υπερβολικής Γεωμετρίας ως υποκείμενη της δομής των σύνθετων δικτύων, διαμορφώνοντας την τοπολογία τους και ορίζοντας την πιθανότητα σύνδεσης δυο κόμβων ως εξαρτώμενη από την απόσταση μεταξύ τους στα πλαίσια αυτού του μετρικού χώρου. Στο χώρο της βιολογίας, οι ιεραρχικές αναπαραστάσεις έχουν χρησιμοποιηθεί με επιτυχία για την κατασκευή και ανάλυση φυλογενετικών και εξελικτικών δέντρων, μεταβολικών δικτύων, για τη χαρτογράφηση των νευρωνικών συνάψεων του εγκεφάλου και την απεικόνιση διαφορών σε κυτταρικό ή πρωτεϊνικό επίπεδο. Συνεπώς, στα βιολογικά αυτά συστήματα μπορεί να υποτεθεί η ύπαρξη υποκείμενης Υπερβολικής Γεωμετρίας χαμηλής διάστασης. Μέσω αυτής της γεωμετρικής προσέγγισης θα μπορούσαν να μελετηθούν και άλλα βιολογικά συστήματα, όπως τα πρωτεϊνικά δίκτυα. Μια τέτοια εφαρμογή θα βοηθούσε στην εκτίμηση της πιθανότητας αλληλεπίδρασης οποιουδήποτε ζεύγους πρωτεϊνών και κατ' επέκταση στον εντοπισμό πιθανών αλληλεπιδράσεων μεταξύ πρωτεϊνών που εμφανίζονται κοντά στον Υπερβολικό Χώρο. Τα δεδομένα αυτά θα μπορούσαν να αξιολογηθούν σε τομείς όπως η προσομοίωση γεγονότων κυτταρικής σηματοδότησης, η ανακατασκευή μονοπατιών μεταγωγής σήματος και η μελέτη των επιπτώσεων των διαταραχών σε πρωτεϊνικές οδούς επικοινωνίας. Στη συνέχεια παρουσιάζονται συνοπτικά η διάρθρωση της εργασίας και το αντικείμενο του κάθε κεφαλαίου αυτής.

Στο Κεφάλαιο 1 δίνεται το συγκείμενο και η συμβολή της παρούσας εργασίας καθώς και μια συνοπτική περιγραφή της διάρθρωσης της.

Στο Κεφάλαιο 2 πραγματοποιείται μια σύντομη εισαγωγή στη Θεωρία των Γράφων, καθώς αποτελεί τον κλάδο των Διακριτών Μαθηματικών που χρησιμοποιήθηκε για τη μοντελοποίηση σχέσεων των στοιχείων και την οπτικοποίηση των δικτύων. Δίνονται ορισμοί βασικών εννοιών και μετρικών που συναντώνται στη Θεωρία Γράφων και ειδικά στην Ανάλυση Σύνθετων Δικτύων, όπως ο βαθμός κόμβου, η κατεύθυνση και το βάρος ακμών, ο πίνακας γειτνίασης, το ελάχιστο μήκος μονοπατιού, η κατανομή βαθμού, η κεντρικότητα βαθμού, εγγύτητας και η ενδιαμεσική κεντρικότητα. Επίσης, περιγράφονται τα πιο χαρακτηριστικά μοντέλα σύνθετων δικτύων, συμπεριλαμβανομένου του μοντέλου τυχαίου γράφου Erdős-Rényi, το μοντέλου μικρού κόσμου Watts-Strogatz, αλλά και το μοντέλο Barabási-Albert. Για το καθένα εξ αυτών δίνονται οι ιδιότητες τους όσον αφορά στις μετρικές κεντρικότητας βαθμού, συντελεστή

ομαδοποίησης και μέσου μήκους μονοπατιού, αλλά και οι μαθηματικές σχέσεις που διέπουν τον τρόπο δημιουργίας τους.

Στο Κεφάλαιο 3 γίνεται μια ανασκόπηση των βασικών εννοιών της Υπερβολικής Γεωμετρίας, της χρησιμότητάς της στην ανάλυση δεδομένων μεγάλης κλίμακας και στην ενσωμάτωση σύνθετων δικτύων στον Υπερβολικό Γεωμετρικό Χώρο. Με τον όρο ενσωμάτωση αναφέρεται η προβολή κάθε κόμβου του δικτύου στον Υπερβολικό Χώρο με τον ορισμό συγκεκριμένων συντεταγμένων. Η Υπερβολική Γεωμετρία (ή αλλιώς γεωμετρία του Lobachevsky) είναι μια μη-Ευκλείδεια γεωμετρία σταθερής αρνητικής καμπυλότητας, όπου αξιωματικά πλέον, από σημείο εκτός ευθείας άγονται άπειρες παράλληλες ευθείες. Δυο κοινά μοντέλα αναπαράστασης του Υπερβολικού Χώρου, τα οποία είναι χρήσιμα για τις ανάγκες της παρούσας εργασίας, είναι το μοντέλο δίσκου του Poincaré και το μοντέλο του Υπερβολοειδούς. Η ύπαρξη μετασχηματισμού από το ένα μοντέλο στο άλλο, με ταυτόχρονη διατήρηση όλων των γεωμετρικών ιδιοτήτων του χώρου, τα καθιστά ισομετρικά. Υποθέτοντας ότι η δομή και η διαμόρφωση της τοπολογία των Σύνθετων Δικτύων βασίζεται στην Υπερβολική Γεωμετρία, υπάρχουν βιβλιογραφικά δεδομένα που υποστηρίζουν ότι στοιχεία των σύνθετων δικτύων που τα χαρακτηρίζουν, όπως η ετερογενής κατανομή βαθμού, η ισχυρή ομαδοποίηση και ο σχηματισμός κοινοτήτων, απορρέουν φυσικά από την αρνητική καμπυλότητα και τις μετρικές ιδιότητες της υποκείμενης Υπερβολικής Γεωμετρίας. Έτσι, θεωρείται ότι οι υπερβολικές αποστάσεις μεταξύ των κόμβων ελέγχουν την πιθανότητα σύνδεσης αυτών. Αντιστρόφως, ένα δίκτυο που παρουσιάζει μετρική δομή και η κατανομή βαθμού του είναι ετερογενής, τότε το δίκτυο θα βασίζεται σε μοντέλα Υπερβολικής Γεωμετρίας.

Στο Κεφάλαιο 4, περιγράφονται τα μοντέλα και οι αλγόριθμοι που έχουν αναπτυχθεί για την ενσωμάτωση ενός δικτύου στον Υπερβολικό Χώρο, και θα χρησιμοποιηθούν στην παρούσα εργασία. Ένα τέτοιο μοντέλο δικτύων είναι το μοντέλο Βελτιστοποίησης Δημοτικότητας-Ομοιότητας (Popularity-Similarity Optimization, PSO). Χάρη στο μαθηματικό του πλαίσιο αναπαράγει επιτυχώς την κατανομή βαθμού άνευ-κλίμακας, τα φαινόμενα μικρού κόσμου και τον υψηλό συντελεστή ομαδοποίησης που χαρακτηρίζουν τα σύνθετα δίκτυα. Στο μοντέλο αυτό, κάθε κόμβος αποκτά πολικές συντεταγμένες λαμβάνοντας υπόψη πως όσο μεγαλύτερος ο βαθμός του κόμβου τόσο υψηλότερη η δημοτικότητα του και άρα τόσο μικρότερη αναμένεται η ακτινική του συντεταγμένη (τόσο πιο κοντά στο κέντρο η θέση αυτού), ενώ όσο μεγαλύτερη η ομοιότητα μεταξύ δυο κόμβων, τόσο μικρότερη η γωνιακή τους απόσταση. Η πιθανότητα δημιουργίας δεσμών καθορίζεται από την υπερβολική απόσταση μεταξύ του κάθε ζεύγους κόμβων.

Μια από τις μεθόδους που βασίζεται στη γενικευμένη εκδοχή του PSO μοντέλου, είναι ο αλγόριθμος ενσωμάτωσης HyperMap. Κατά την εκτέλεση του, το προς ενσωμάτωση δίκτυο αρχικά αποσυναρμολογείται σε ασύνδετους κόμβους. Έπειτα, σε κάθε κόμβο ανατίθεται μια ακτινική και μια γωνιακή συντεταγμένη, με στόχο τη μεγιστοποίηση μιας εκτιμήτριας μέγιστης πιθανοφάνειας, η οποία εκτιμά την πιθανότητα δημιουργίας μιας νέας σύνδεσης όταν, δεδομένου του ήδη σχηματισθέντος δικτύου και της ακτινικής του συντεταγμένης, ο κόμβος

έχει αυτή τη γωνιακή συντεταγμένη. Εν τέλει, ολόκληρο το δίκτυο ενσωματώνεται στον Υπερβολικό Χώρο. Το PSO μοντέλο χρησιμοποιείται και από τον αλγόριθμο LaBNE, ο οποίος βασίζει την ενσωμάτωση του δικτύου στο δίσκο Poincaré, σε μια μη-γραμμική μείωση διαστάσεων του Λαπλασιανού πίνακα. Ακολούθως, οι κόμβοι οργανώνονται ώστε οι ακτινικές συντεταγμένες τους να προσομοιάζουν με την κατάταξη βαθμού κόμβου και οι γωνιακές συντεταγμένες να λαμβάνονται μέσω της επίλυσης ενός προβλήματος ιδιοτιμής. Ο αλγόριθμος LaBNE εξαρτάται σε μεγάλο βαθμό από τις τοπολογικές πληροφορίες, με αποτέλεσμα να πετυχαίνει τιμές υψηλότερης ακρίβειας κυρίως σε περιπτώσεις δικτύων με υψηλό συντελεστή ομαδοποίησης και υψηλό μέσο βαθμό των κόμβων. Στη συνέχεια, εκμεταλλευόμενοι την εξαιρετική ταχύτητα και χαμηλή υπολογιστική πολυπλοκότητα του LaBNE, πραγματοποιείται μια βελτίωση της ενσωμάτωσης στον Υπερβολικό Χώρο μέσω της βελτιστοποίησης των γωνιακών συντεταγμένων που έχουν ήδη ανατεθεί. Για κάθε κόμβο του δικτύου εξετάζεται ορισμένος αριθμός νέων πιθανών γωνιακών συντεταγμένων, στο γωνιακό διάστημα που ορίζουν οι συντεταγμένες των δεύτερων γειτόνων του κόμβου, με στόχο την ελαχιστοποίηση της λογαριθμικής απώλειας. Το πλήθος των υποψήφιων θέσεων και των φορών επανάληψης όλης της διαδικασίας καθορίζεται από το χρήστη.

Ένας ακόμη αλγόριθμος που χρησιμοποιήθηκε στην εργασία είναι ο Rigel, ο οποίος κατά την ενσωμάτωση δεν έχει στόχο την προσαρμογή του γράφου στο PSO μοντέλο όπως οι προηγούμενοι, αλλά τη διατήρηση των αποστάσεων μεταξύ κόμβων στον Υπερβολικό Χώρο όσο το δυνατόν πιο κοντά στις αντίστοιχες γεωδαισικές διαδρομές που τους συνδέουν στο γράφο. Για τη λειτουργία του επιλέγονται κάποιοι αρχικοί κόμβοι ως "ορόσημα" με βάση τη μέγιστη τιμή κεντρικότητας βαθμού, οι οποίοι τοποθετούνται κατά τρόπο που οι μεταξύ τους αποστάσεις να είναι ίσες με τις αποστάσεις τους στο γράφο. Στη συνέχεια, στους υπόλοιπους κόμβους αναθέτονται συντεταγμένες βελτιστοποιώντας την απόσταση κάθε κόμβου από ένα υποσύνολο "ορόσημων" ώστε αυτή να προσεγγίζει την απόσταση τους στο γράφο. Η βελτιστοποίηση πραγματοποιείται χάρη στη μέθοδο Simplex του γραμμικού προγραμματισμού.

Στο Κεφάλαιο 5, παρουσιάζονται τα σύνθετα δίκτυα που χρησιμοποιήθηκαν (συνθετικά και πραγματικά), απαριθμούνται τα βήματα εκτέλεσης των αλγορίθμων ενσωμάτωσης και περιγράφονται οι μέθοδοι αξιολόγησης της ποιότητας και της ακρίβειας ενσωμάτωσης. Συγκεκριμένα, όσον αφορά στα δίκτυα που αναλύθηκαν στα πλαίσια της εργασίας αυτής, περιλαμβάνονται έξι δίκτυα που κατασκευάστηκαν με βάση το μοντέλο Βελτιστοποίησης Δημοτικότητας-Ομοιότητας και τρία δίκτυα πραγματικού κόσμου (ένα δίκτυο αυτόνομων συστημάτων, ένα κοινωνικό δίκτυο χρηστών ραδιοφωνικού σταθμού και το δίκτυο πρωτεϊνικών αλληλεπιδράσεων του *Homo sapiens*).

Ως μέθοδοι αξιολόγησης των αλγορίθμων ενσωμάτωσης χρησιμοποιήθηκαν η άπληστη δρομολόγηση και η πρόβλεψη δεσμών. Εάν η άπληστη δρομολόγηση εκμεταλλευόμενη τις συντεταγμένες των κόμβων εντός του γεωμετρικού χώρου, είναι αποτελεσματική, τότε το ενσωματωμένο δίκτυο θεωρείται βατό/προσπελάσιμο. Για την αξιολόγηση της αποτελεσματικότητας χρησιμοποιήθηκαν: το

ποσοστό των επιτυχώς ολοκληρωμένων μονοπατιών εντός του δικτύου και ο μέσος λόγος του μήκους ενός άπληστου μονοπατιού σε σχέση με το αντίστοιχο γεωδαισικό μονοπάτι στο γράφο.

Η πρόβλεψη δεσμών βασίζεται στο γεγονός ότι η μικρή υπερβολική απόσταση μεταξύ δυο κόμβων συνδέεται με μια υψηλή πιθανότητα ύπαρξης ενός μεταξύ τους δεσμού. Για την αξιολόγηση της προβλεπτικής ικανότητας ενός αλγορίθμου αφαιρέθηκε τυχαίο πλήθος δεσμών, υπολογίστηκαν οι υπερβολικές αποστάσεις μεταξύ των μη γειτονικών ζευγών κόμβων, και η τιμή αυτή χρησιμοποιήθηκε ως μέτρο της πρόβλεψης των δεσμών. Η αξιολόγηση του αλγορίθμου έγινε χάρη στην καμπύλη Ακρίβειας-Ανάκλησης, η οποία δημιουργήθηκε μέσω κινούμενου κατωφλιού επί της λίστας των πιθανών δεσμών ώστε να υπολογιστούν τα αντίστοιχα στατιστικά.

Μια ακόμη μετρική που ελέγχθηκε ήταν εκείνη της λογαριθμικής απώλειας, σύμφωνα με την οποία όσο μικρότερη η λογαριθμική απώλεια τόσο καλύτερη η ενσωμάτωση του δικτύου στο γεωμετρικό χώρο. Για κάθε αλγόριθμο, υπολογίστηκε η συνολική τιμή λογαριθμικής απώλειας, χρησιμοποιώντας τις συντεταγμένες που είχε αναθέσει στους κόμβους του δικτύου. Η αξιολόγηση των αλγορίθμων έγινε με σύγκριση των μεταξύ τους τιμών.

Τέλος, στην περίπτωση του πρωτεϊνικού δικτύου που ενσωματώθηκε με χρήση του αλγόριθμου Rigel, εφαρμόστηκε μια ακόμα μεθοδολογία προκειμένου να βρεθεί η δυνητική σχέση μεταξύ των αποστάσεων στον Υπερβολικό Χώρο και της λειτουργικής απόστασης των πρωτεϊνών-κόμβων του δικτύου. Συγκεκριμένα, κατασκευάστηκαν ροές εργασιών για να εξεταστεί εάν ο συγκεκριμένος αλγόριθμος δύναται να θέσει μικρότερες αποστάσεις μεταξύ πρωτεϊνών που εμπλέκονται στην ίδια μοριακή λειτουργία ή διαδικασία, σε σχέση με τυχαία επιλεγμένες πρωτεΐνες καθώς και να υπολογιστεί ο βαθμός συσχέτισης των λειτουργικών και υπερβολικών αποστάσεων ζευγών πρωτεϊνών. Για το σκοπό αυτό χρησιμοποιήθηκαν βάσεις βιοϊατρικών δεδομένων, όπου περιλαμβάνεται η λειτουργική επισήμανση των πρωτεϊνών, καθώς και μετρικές που εφαρμόζονται επί των σημασιολογικών αυτών σχημάτων.

Στο Κεφάλαιο 6 παρουσιάζονται τα αποτελέσματα των αλγορίθμων ενσωμάτωσης για κάθε σύνθετο δίκτυο, καθώς και η αξιολόγηση της απόδοσης αυτών.

Τέλος, στο Κεφάλαιο 7 συνοψίζονται τα συμπεράσματα της παρούσας εργασίας, και η πιθανή συνεισφορά της σε ανοιχτά ερευνητικά θέματα που θα μπορούσαν να αποτελέσουν μελλοντικές προεκτάσεις της.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Συμεών Παπαβασιλείου για την εμπιστοσύνη του και την ευκαρία που μου έδωσε να διερευνήσω ένα θέμα του ενδιαφέροντος μου συνδυάζοντας τους χώρους της Βιολογίας και της Ανάλυσης Δικτύων.

Ιδιαίτερα ευγνώμων είμαι στον Αναπληρωτή Καθηγητή κ. Βασίλειο Καρυώτη για την άψογη συνεργασία και την καθοδήγηση του, καθώς επίσης και στο Διδάκτορα Κωνσταντίνο Τσιτσεκλή, για τις συμβουλές και την αδιάκοπη βοήθεια που μου προσέφερε καθ' όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στην οικογένεια μου για όσα μου έχει προσφέρει, και στον άνθρωπο που με στηρίζει και με υπομένει ό,τι κι αν συμβεί.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Subject

In the past decades, the world has witnessed tremendous technological advances that are yielding an outstanding collection of data from almost every aspect of life, creating large systems of diverse interconnected entities. Data science, and especially Big Data analysis, contributes to the endeavor to utilize these data, improve our understanding of the world and find solutions to some prominent challenges. One such challenge is the understanding of biological phenomena and applying the newly acquired knowledge to many different levels of medicine. Each different type of omics data, including genomics, transcriptomics, proteomics and metagenomics, measure different aspects of cellular functionality. As the available amount of omics data grows, so does its complexity, and it becomes progressively harder to analyze, understand and draw conclusions about them. In addition, it is important to find the proper mathematical models that make the data manageable by computational analysis, in order to abstract these complex data systems. The liaison of biology, mathematics and computer science is a possible manner in which we will be able to overcome these challenges and reach the knowledge.

Undoubtedly, the reductionist approach of accumulating knowledge solely for the constituent elements is essential to our understanding of simple governing laws of individuals. However, it is just as important the study of these building blocks and their roles as part of a broader system of interacting components. Such systems display organization that reflects their function. Recognizing the governing principles of these systems will reveal the possible impact of

changes on themselves or their environment. Graphs (networks) are the means to represent elements and their interactions, and investigate the organization of complex systems. Real complex systems, including biological, social and technological ones, often having common features in their organization, such as their scale-free degree distributions and their small-world behavior, can be represented, analyzed and modeled by networks. Complex network analysis and studying of network models have become very applicable, with topics of interest including the understanding of dynamic or static processes, such as evolutionary pathways and patterns [1, 2], epidemics spreading [3, 4], detecting communities [5, 6] and predicting of missing, forthcoming or spurious interactions [7, 8]. Studying biological networks, such as protein-protein interactions network, is key to understand complex biological activities and systems and to identify biological functions.

Another fact about complex networks is their hierarchical structure with exponential expansion of possible states and its ability to be approximately represented by tree-like structures. Modelling these relationships in Euclidean spaces is quite limited and problematic, since pairwise distances distort substantially. On the other hand, Hyperbolic Space can be regarded as a continuous analog of trees allowing low-distortion embeddings [9]. At the same time, models and algorithms uphold the existence of a Hyperbolic Geometry underlying the structure of complex networks and shaping their topology, posing a distance-dependent connection probability between nodes in this metric space. In biology, hierarchical representations, such as phylogenetic trees have been successfully used in visualization of metabolic networks [10, 11], mapping of neural connections in the brain [12, 13] and depicting differences in a cellular or protein level [14, 15]. So the existence of hyperbolic metric as a low-dimensional geometry should be treated as a logical consequence. This geometric perspective could, furthermore, alleviate challenges in biology and medicine, as for example the prediction of protein-protein interactions, which corresponds to the identification of protein pairs that appear hyperbolically close to each other.

In the context of this thesis, the embedding of complex networks into the Hyperbolic Space using different algorithms was studied. The results were compared in terms of computational time, accuracy and total performance. Regards to the human interactome, a more biological question was posed, concerning proteins participating in a common biological mechanism and their inferred proximity in the Hyperbolic embedding Space.

1.2 Thesis Organization

This thesis is organized in seven chapters. Chapters 2 and 3 present the theoretical background, as well as the basic concepts and techniques being used. Chapter 4 poses the problem of network embedding in Hyperbolic Space and the algorithms that were adopted in order to implement this task, while chapters 5 and 6 describe in detail the experimental steps taken, the evaluation metrics as well the results of the embedding process in different types of networks. Finally, chapter 7 summarizes the conclusions of this thesis and suggests ideas for any future work.
In more details:

- Chapter 2 makes the necessary introduction to Graph Theory, a major branch of discrete mathematics used to model relationships between objects, analyze and visualize their networks. Here the reader will, also, get acquainted with the terminology and symbols of graphs being used in the rest of the thesis. In addition, the key features and metrics of complex networks will be described, as well their generation models, including the scale-free networks that will be the main object of the experiments in this thesis.

- Chapter 3 presents the basic axioms of Hyperbolic Geometry and its usefulness for Big Data analysis, and inspects two common representation models of embedded graph nodes in the hyperbolic plane.

- Chapter 4 describes the three embedding frameworks used in the present thesis, along with their embedding processes and representation models of the Hyperbolic Space, as well two versions of a network generation model that can be considered as prerequisites of the algorithms. Also, a novel optimization step that could be used to improve the embedding algorithm's accuracy, namely LaBNE, is outlined.

- In Chapter 5, the steps of the whole experimental process are described in detail accompanied with two downstream applications, greedy routing and link prediction, that are presented in conjunction with their role as evaluation measures of the quality and efficacy of the embedding. Also, in the case of human interactome an additional semantic similarity criterion

was implemented in order to detect potential association between hyperbolic distance and functional divergence of proteins.

- In Chapter 6, the corresponding experimental results of the different algorithms applied in the different networks are demonstrated, followed by the evaluation of these results.
- Finally, Chapter 7 summarizes the conclusions of this dissertation and provides directions for follow-up development, applications and future steps that could be taken.

Chapter 2

Graph Theory

2.1 Definitions and Features

Graph theory is a cognitive field of Discrete Mathematics, with applications in computer science, engineering, biology, chemistry, economics, sociology and the humanities, where it is used in the problem modeling and solving and the study of objects' relations when affinity or connections exists.

A graph is a mathematical object used to model relationships between graph entities. By definition, a graph $G = (V, E)$ is an ordered pair consisting of two finite sets, where $V = \{v_1, ..., v_n\}$ is a set of vertices and $E = \{e_1, ..., e_m\}$ a set of edges, such that every edge is associated to a tuple $\{u, v\}$ for vertices $u, v \in V$ which are called its endpoints. The order and size of a graph is the number of vertices, $|V|$, and the number of edges, $|E|$, respectively. A pair of vertices u, v are adjacent if $e = \{u, v\} \in E$. Node degree $deg(v)$ of a vertex v is the number of edges incident to node v in G. An edge connected at both ends to the same vertex is counted twice in node degree calculation and is called a loop. An edge can be directed or undirected and respectively, a graph can be characterized as directed if its edges do have a direction or undirected if none of its edges have a direction, capturing non reciprocal relations in the latter case. In directed networks, we distinguish between in-degree, representing the number of edges that point towards a particular node i, and out-degree, representing the number of edges originating from the node i towards other nodes. In this case, a node's total degree is the sum of in- and out-degree. The adjacency matrix of the graph G is a square matrix $A = [\alpha_{ij}]$ of size $V \times V$, which represents the edges of a graph, where $A_{ij} = 1$ indicates the presence of an edge from vertex i to vertex j

and $A_{ij} = 0$ means the two nodes are not connected. In weighted graphs, the edge between each connected pair of vertices is annotated with a weight w, and the corresponding value in the adjacency matrix is $A_{ij} = w$. For undirected graphs, the adjacency matrix is always symmetric, and the node degree of a node i equals the sum of the i-th row or column of the matrix. In directed graphs, row and column sums are respectively equivalent to the in- and out-degree of each node.

A path is an ordered sequence $\{v_0, v_1, ..., v_k\}$, such that $\{v_i, v_{i+1}\} \in E$ for $i = 0, ..., k$–1, and which does not contain any vertex more than once. A walk resembles a path except it has no restriction on the number of times a vertex can be visited, and a cycle is a path except that it starts and ends at the same vertex. The length of a path (or walk or cycle) is defined as the number of edges in it. Vertices $u, v \in V$ are characterized as connected in G if a uv-path exists in G. Also, $u, v \in V$ have distance k in G, if the minimum length of the uv-path in G is k. That shortest path is called a geodesic.

A graph G is connected, if all vertices are reachable from every other vertex. A connected component is an inclusion-maximal subgraph H of G, where for every pair of distinct vertices u, v in H there exists a uv-path in H. Maximal means that there is no other node in G such that it could be added to the subgraph H and all the existing nodes would remain connected. In a graph G, with more than one components, the subgraph H which contains the largest number of nodes, is called giant component. Visually, components of a graph G are its individual pieces that add up to make G (Fig. 2-1).

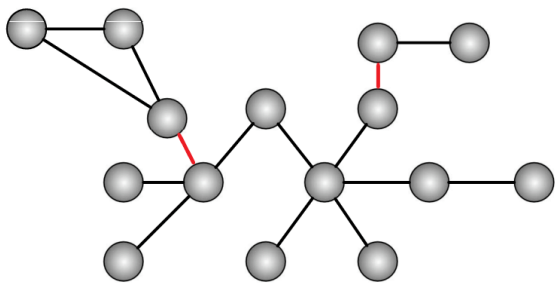

Figure 2-1: [16] A network with three components. The subgraphs of two and three nodes respectively, are the small components, and the larger subgraph in the middle, is the giant component. If the two red edges existed then the whole graph would be connected.

One of the most important features of the full-scale structure of a graph, but also a quite simple one information-wise, is the distribution of the network node degrees. The degree distribution of a network, $P(k)$, is defined as the fraction of nodes having degree k, or otherwise the probability of randomly selecting a node with degree k. In a graph with N nodes in total, and N_k nodes having degree k, the distribution is given by $P(k) = \frac{N_k}{N}$. Typically, in the simplest types of networks most nodes have similar degree. However, commonly, in real social or biological networks only a few nodes, referred to as hubs, have very large degree while the vast majority of nodes have relatively small degree [17–20]. This type of networks approximately follow a power-law degree distribution $P(k) \sim k^{-\gamma}$, where γ is a constant. Such networks are called scale-free and will be the main focus for the rest of the thesis.

A measure that determines the structural and topological importance of a node in a network as well as its influence on the other nodes, is the node centrality. Three of the most well-known aspects of centrality: degree, closeness, and betweenness. Those are used for the identification of nodes in prominent positions, of communication mediators towards remote nodes, of contributors to network scalability, along with vulnerable nodes in case of an attack. Given a node u in the network, degree centrality, $C_D(u) = \sum_{v=1}^{N-1} A_{uv}$, equals to the number of neighbors of node u and determines to what extent that node is connected the others in the network. The value of this metric can be normalized by dividing it by the number of all possible neighbors, $C_D(u) = \frac{1}{N-1} \sum_{v=1}^{N-1} A_{uv}$, comparing thereby the size of node's "local" neighborhood to the size of the whole graph. The degree centrality of node u can also be interpreted as its ability to interact with other nodes, but can not determine its topological position in the graph. Closeness centrality, $C_C(u) = \frac{N-1}{\sum_{v=1}^{N-1} d(v,u)}$, captures the "ease" that a walker could reach all the nodes in the network starting from node u, i.e., how "close" a node is to the others. This metric is based on the length of shortest paths. The smaller the graph-theoretic distances from one node to the rest of the graph, the more central this node is considered to be. In contrast to degree, this metric takes into account both direct (adjacent) and indirect (having a common contact or a linking path) connections of the node. Betweenness centrality, $C_B(u) = \sum_{v,w \in G} \frac{\sigma(v,w|u)}{\sigma(v,w)}$, highlights the "mediator" or "bridge" role of node u in the network by portraying the fraction of geodesic paths forced to pass through a specific node, i.e. how likely is that one will have to go through node u while navigating the graph to go from some node v to a node w. The

higher the value of betweenness centrality, the more necessary the node is in network navigability, and the more crucial it is in spreading phenomena. Also, in the case of removing a node with high betweenness centrality, there exists a high risk of breaking a connected component into two smaller ones.

2.2 Complex Networks and Graph Models

Both in technological and natural modern world, various networks are found with non-trivial topological properties (different from the ones met in simple networks) [21]. Those graphs are described as complex networks and are volatile, capable to evolve over time, while reorganizing their structure and fluctuating in size with the addition or elimination of nodes or edges and the edge-rewiring. Such networks are the World Wide Web, social networks as well biological networks. Due to the available computing power, nowadays it is possible to collect and process large-scale data from this type of networks, in order to extract their structural and functional features. Network modeling contributes to statistical analysis and visualization, allowing comparisons and interpretation of networks' creation, evolution and macroscopic behavior. The two main categories of models which simulate real networks' properties are the static and the dynamic models. Static models such as the Erdős-Rényi random graph model [22] and the Watts-Strogatz small world model [23], are characterized by a fixed number of nodes that remain immutable during graph's life. On the other hand, dynamic models change and evolve over time, creating a series of network snapshots, since new nodes are created and incorporated at each time step, forming connections based on the existing network structure. A typical example is the Barabási-Albert model [17].

In 1959, Paul Erdős and Alfréd Rényi, based on the creation of graphs with random edge formation, introduced the network representation model $G(N, M)$, with a fixed number of N vertices and a fixed number of M edges in the graph [22]. Creation of such a network is equivalent to a uniformly random choice from the collection of all graphs with N vertices and M edges. Contemporaneously and independently, Edgar Gilbert introduced the closely related model $G(N, p)$, with a fixed number of N vertices where each edge has a fixed probability of being present (p) or absent ($1 - p$), independently of the other edges [24]. When the number of vertices N tends to infinity, the two models are equivalent. The

behavior of random graphs is often studied in this context. In particular, by the law of large numbers any $G(N, p)$ graph will approximately have the expected number of edges $\binom{N}{2}p$. Hence, if $pN^2 \to \infty$ then $G(N, p)$ should behave similarly to $G(N, M)$ with $M = \binom{N}{2}p$ as M increases. It is worth mentioning that the structural properties of the network vary depending on the probability value p, and there is a critical threshold value that encourages some network properties to be displayed or hidden. For example, if $p \geq \frac{lnN}{N}$ the network is connected, while otherwise the network is partitioned into components that do not communicate with each other [22, 25]. Although simple and powerful, Erdős-Rényi (ER) graphs fail to describe two important properties of real-world networks, including the generation of local clustering and communities (low clustering coefficient due to the constant connection probability), as well the formation of hubs (degree distribution of ER graphs converges to Poisson rather than power-law).

In order to address the former problem, in 1998, Duncan Watts and Steven Strogatz proposed a model that combines the short average path length of the ER model with the clustering, by interpolating between a regular ring lattice and a random graph [23]. As found by Milgram's famous experiment [26], it takes an average of six consecutive steps (also known as "six degeee separation") to connect two randomly selected citizens of the United States. The main conclusion of the experiment is that even on very large networks, shortcuts do exist and nodes are able to locate these routes using only local information (searchable network). This behavior, referred to as "small-world" phenomenon, is ubiquitous in real world networks and is partially explained by Watts and Strogatz proposed model. To create such a network, N nodes are placed at first, and each node is connected to its k nearest neighbors. Afterwards, each edge is either reconnected or a new edge is added with probability p [23, 27], ensuring that short paths are formed between the nodes. The model parameter p controls the creation of networks inside the spectrum from regular ($p = 0$) to random ($p = 1$), while the case of an intermediate value provides the desired properties of small geodesic distances and high clustering (Fig. 2-2).

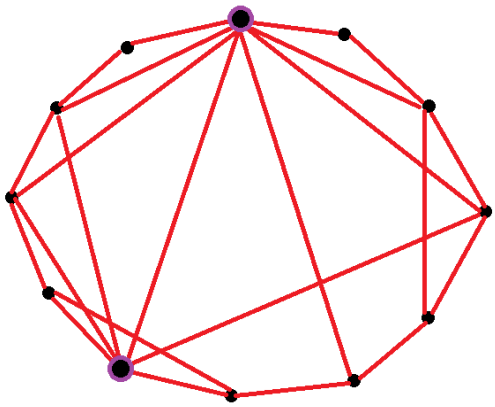

Figure 2-2: [28] Small-world network created by a ring lattice with a fraction of new edges added to it, resulting in many short connections accumulated in few highly connected hubs (highlighted as bigger), thus shortening the typical path length and increasing the local connectedness.

Although widely used, both Erdős-Rényi and Watts-Strogatz models suffer from a few weaknesses in comparison with real-world networks' properties. By definition, ER graphs are being created with a fixed number of nodes and a common, consistent probability of connection between all nodes of the network, which is not the case with real networks that are open systems capable of integrating new nodes throughout their evolution. Also, in biological or social networks new nodes follow a preferential connection behavior depending on the number of connections each existing node already has. Concerning the "small-world" approach, despite capturing the high clustering feature, it still remains a close system that also fails to describe the desired heterogeneous degree distribution.

The aforementioned deficiencies led Albert-László Barabási and Réka Albert [17], in 1999, to propose a model incorporating two important mechanisms: growth and preferential attachment. Growth implies a repeated entry of new nodes in the network that increases over time while preferential attachment implies the predilection of newly introduced nodes to link to the more connected nodes. Preferential attachment is an example of positive feedback in which an initial random deviation, i.e. the difference between the number of connections that the nodes have, is amplified automatically, resulting in a constant increase, thus "the rich getting richer".

Previous works that contributed to the "preferential attachment" concept include the studies of Herbert Simon and Derek de Solla Price. In 1955, Simon proposed a class of stochastic models leading to power-law distributions, in

order to explain the common features of his empirical observations from social, biological and economic phenomena [29]. Afterwards, Price applied the idea to the growth of networks, finding a proportionality between the number of citations of a publication and the new citations for that particular publication. This process was named "cumulative advantage" and used to produce a directed citation network governed by the adage "the rich get richer" [30]. Although, the mechanism of cumulative advantage could have been treated as the explanation of the observed power-law degree distribution, it was its rediscovery a few decades later, by Barabási and Albert that became known to the scientific community and gained the acceptance [17]. Barabási-Albert model has one important difference from Price's model, and therefore it can be considered as an undirected version of it, independently rediscovered and applied on the Internet [31, 32].

The model proposed by Barabási-Albert is not based on modifications of a given static topology, but instead generates a topology from scratch, resulting to a scale-free graph which captures the dynamic behavior of real-world networks. The term scale-free refers to the lack of degree scalability, i.e. the lack of uniformity in the degree distribution. Specifically, in scale-free networks, different groups of nodes present different degrees and that translates to different scale of connectivity and number of neighbors. The distribution of nodes' connections follows a power-law distribution, a decreasing function with scale invariance, in contrast to the Poisson distribution observed in random networks. In such, the distribution of connections is not "democratic", since there are few nodes that dominate the network and many others with a small number of connections.

The algorithmic steps to create a Barabási-Albert graph with m new edges per time step, are as follows:

- First, we start with m_0 nodes that are connected to each other in a completely random way. In this stage, we must ensure that each node has at least one edge.
- The network is developed by performing the following two steps:
    1. At each time step we add a new node with m edges (where $m \leq m_0$) that connect this newly introduced node to m nodes that already exist in the network (growth step).
    2. The probability p_i that one of the edges of the new node will be connected to the node i that already exists in the network depends

on its degree k_i and is given by the relation $p_i = \frac{k_i}{\sum_j k_j}$, where the sum in the denominator is computed over all the j nodes that already exist in the network at the time the new node is added (preferential attachment step).

- After t time steps, the algorithm generates a scale-free network with $t + m_0$ nodes and $m_0 + m * t$ edges.

The study of the model led to the conclusion that the final network has a power-law distribution with exponent $\gamma = 2.9 \pm 0.1$, independent of m, the only parameter of the model.

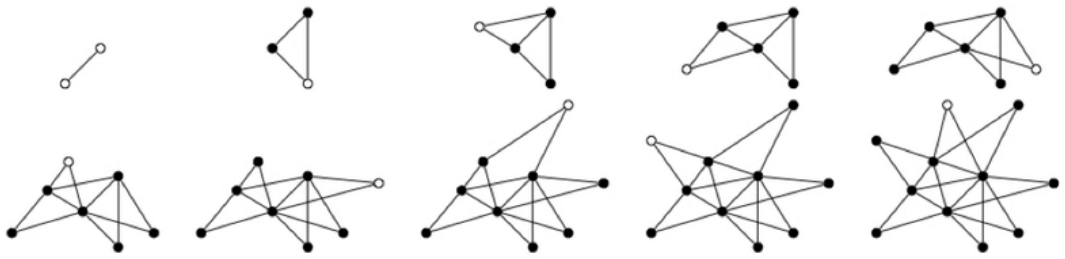

Figure 2-3: [33] Scale-free network evolution as proposed by Barabási-Albert model. At each iteration, a new node (marked as an empty circle) is introduced to the network, and attaches its two new edges using preferential attachment.

As shown in the above example of Fig. 2-3, we initially have $m_0 = 2$ nodes connected to each other and we add a new node in each step. The added node (marked as an empty circle) will always form $m = 2$ connections with the already existing nodes of the network at that time. Black circles are the interconnected nodes which constitute the formed network up to that time. After $t = 9$ iterations, we end up having $N = 9 + 2 = 11$ nodes and $1 + 2 * 9 = 19$ edges in the network.

Graphs generated by the Barabási-Albert model show significantly higher clustering coefficient as well as smaller average path length compared to an analogous random graph. With regards to small-world networks, although the average path length displays a logarithmic growth proportional to the number of nodes, the clustering coefficient decreases respectively, instead of remaining constant.

Finally, Barabási-Albert model belongs to the generative models, since scale-free networks' creation is controlled by specific mechanisms and not by chance. Another generative model is the copying model. Driven by the observation that

creators of new web pages on a topic tend to copy links from other relevant web pages, Jon M. Kleinberg and Ravi Kumar proposed a model in which every time a new node enters the network either chooses independently and uniformly at random k nodes to connect to, with probability β, or randomly chooses an existing node from whom copies a fraction of k edges, with probability $(1 - \beta)$ [34, 35]. This, also, results in networks with power-law degree distribution.

Chapter 3

Hyperbolic Geometry

3.1 History and Evolution

Euclid's Elements [36] is definitely the most famous mathematical work of classical antiquity, and also the world's oldest continuously used mathematical textbook. It has proven to be majorly contributory in the development of logic and modern science, and its logical rigor was not surpassed until the advent of non-Euclidean Geometry in the 19th century. It is a comprehensive collection of definitions, postulates, theorems (and their mathematical proofs). The books, thirteen in total, discuss perfect numbers and primes, Pythagoras' Theorem, the "golden ratio", contain formulas for calculating the volumes of solids, and finally cover plane and solid Euclidean Geometry. They are a masterful compilation of the geometric knowledge of earlier Greek mathematicians, and Euclid is credited with arranging all these in a logical manner, in order to establish that they comply with the five postulates of Euclidean Geometry. They are as follows:

1. A straight line segment can be drawn joining any two points.
2. Any straight line segment can be extended indefinitely in a straight line.
3. Given any straight lines segment, a circle can be drawn having the segment as radius and one endpoint as center.
4. All right angles are congruent.
5. If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the

two lines inevitably must intersect each other on that side if extended far enough.

The first four postulates form the "Absolute Geometry", while the last one is known as the Parallel Postulate. Many mathematicians attempted to prove it as a theorem or to prove the Euclid's Elements without using it, but to no avail. However, this process led independently Gauss (1777-1855), Lobatschewsky (1793-1856) and Bolyai (1802-1870) in developing the Hyperbolic Geometry, and Riemann (1826-1866) in developing the Elliptic Geometry. These entirely self-consistent "non-Euclidean Geometries" have been derived by using the Absolute Geometry along with various negations of the Parallel Postulate.

In 1899, Hilbert (1862-1943) published his book "Grundlangen der Geometrie" [37], axiomatizing and creating the foundation for a modern treatment of Euclidean Geometry. Hilbert's axiom system is constructed by:

- Axioms of Incidence
- Axioms of Order
- Axioms of Congruence
- Axioms of Continuity
- Axioms of Parallels

So, in Euclidean Geometry, instead of the Parallel Postulate, we can use the fifth Hilbert's axiom of Parallels, as firstly formulated by Playfair, stating that "Given a line and a point not on it, exactly one line parallel to the given line can be drawn through the point". In Elliptic Geometry (together with some minor adjustment in Axioms of Incidence and Order), we have that "Given a line and a point not on it, no lines parallel to the given line can be drawn through the point", while in Hyperbolic Geometry, we accept that "Given a line and a point not on it, infinite lines parallel to the given line can be drawn through the point" (Fig. 3-1).

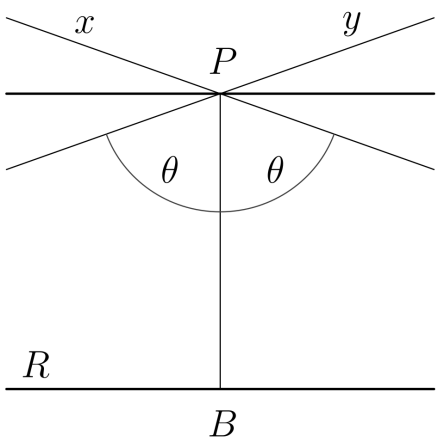

Figure 3-1: [38] Lines through point P are parallel (asymptotic) to line R. Both line R and point P belong to the same plane.

3.2 Fundamentals

Hyperbolic n-space, denoted H^n, is a n-dimensional Riemannian manifold that has a constant negative sectional Gaussian curvature and exhibits Hyperbolic Geometry. Usually a curvature value $K = -1$ is assumed. Hyperbolic Geometry has been shown to describe many aspects of our world, from olfaction [39], biological materials, natural elements [40, 41] and crystalline structures [42] to phylogenetic trees [43], Internet [44], special relativity and black holes [45, 46].

Unlike Euclidean Space, human perception cannot intuitively understand Hyperbolic Space, and in consequence various models for its representation have been proposed, such as the Beltrami–Klein model, the Poincaré disk model, the Poincaré half-plane model, the Hyperboloid model and the Hemisphere model [47–49]. Since all of them describe the same metric space, there is an available transformation from any model into the other, while retaining the geometric properties of the space, making the models isometric (Fig. 3-2). The Poincaré disk and the Hyperboloid, will be further analyzed, since they are the ones used in the present study to embed the nodes of a network into the coordinate space of hyperbolic plane.

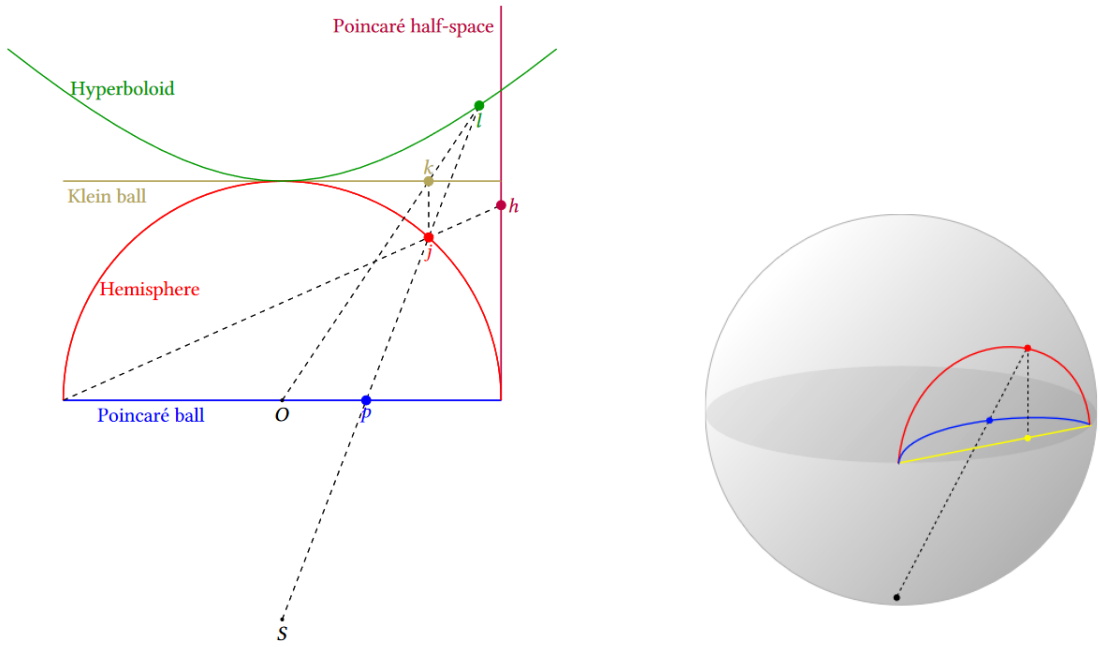

Figure 3-2: [49] **(A)** Relation between models of Hyperbolic Space. **(B)** Geodesics in Poincaré ball, Klein ball, and hemisphere models.

3.2.1 Hyperboloid Model

Hyperboloid model is also known as Minkowski or Lorentz model. It is a model of a n-dimensional Hyperbolic Space, $\mathbb{H}^n$, where points are represented on the forward (positive) sheet (Fig. 3-3) of a two-sheeted hyperboloid surface of $(n+1)$-dimensional Minkowski space, $\mathbb{R}^{n+1} = (x_0, x_1..., x_n)|x_i \in \mathbb{R}, \mathbb{I} = \{0, 1, ..., n\}$. The $x_0..., x_n$ points are such that satisfy the formula:

$$x_0^2 - x_1^2 - .... - x_n^2 = 1, x_0 > 0$$

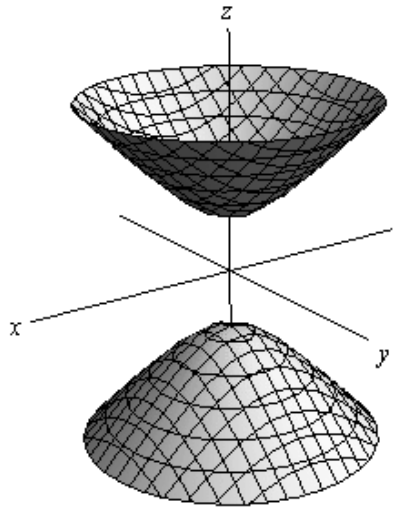

Figure 3-3: [50] Hyperboloid of Two Sheets, also known as "Two Opposing Facing Bowls". A Hyperboloid comes infinitely close to a conic surface. In three dimensions, the Hyperboloid of two sheets has the following form: $x_0^2 - x_1^2 - x_2^2 = 1$, where $x_0 > 0$ defines the so called "forward sheet" (the above part of the figure).

As expected, on a sphere there are many paths connecting two points, but the shortest one is called geodesic. A geodesic is created by the intersection of the hyperboloid surface $\mathbb{H}^n$ with the plane defined by the two points, that we want to connect, and the origin in $\mathbb{R}^{n+1}$. Geodesic is the generalization of a straight line into curved space, defined to be a curve where tangent vectors don't deform, in case of parallel transportation over it (Fig. 3-4).

Hyperbolic Space has a different distance metric from the Euclidean one, that is also different among its various models. In the case of the Hyperboloid model, the distance between two points x and y on $\mathbb{H}^n$ is described by the formula $d(x, y) = \text{arcosh}\, B(x, y)$ where arcosh is the inverse function of hyperbolic cosine and $B(x, y)$ corresponds to the indefinite bilinear form $B(x, y) = \sum_{i=1}^{n} x_i y_i - x_{n+1} y_{n+1}$, that endows the Minkowski space. Specifically, distance is given by the equation:

$$d(x, y) = \text{arcosh}\left(\sqrt{\left(1 + \sum_{i=1}^{n} x_i^2\right) * \left(1 + \sum_{i=1}^{n} y_i^2\right)} - \sum_{i=1}^{n} x_i y_i\right) * |K|$$

where K the space curvature.

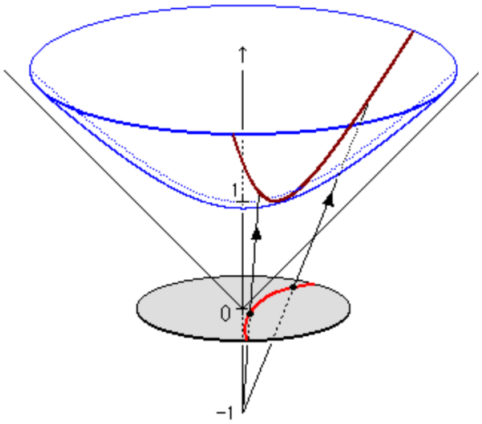

Figure 3-4: [51] Grey Poincaré disk model as a stereoscopic projection of the blue Hyperboloid model. The red geodesic in the Poincaré disk (arc along the unit circle) projects to the brown geodesic line on the blue Hyperboloid.

3.2.2 Poincaré Disk Model

Closely related to the Hyperboloid model is the Poincaré disk model that can be derived from a stereoscopic projection of the Hyperboloid from the focal point $(x_0 = -1, x_1 = 0, ..., x_n = 0)$ onto the unit circle of the $x_0 = 0$ plane (Fig. 3-4). The Poincaré disk model, also known as the conformal disk model, represents the infinite hyperbolic plane $\mathbb{H}^2$ inside the unit disk $\mathbb{D} = \{z \in \mathbb{R}^2 : \|z\| < 1\}$, where $\|\cdot\|$ denotes the $L2$ norm. The Euclidean circle $\mathbb{S}^1$, is the boundary of the unit disk $\partial\mathbb{D} = \{z \in \mathbb{R}^2 : \|z\| = 1\}$ or the boundary at infinity $\partial\mathbb{H}^2$. That circle represents the infinitely distant points of the hyperbolic plane, which are not belong to.

Poincaré disk model is conformal, meaning that euclidean angles between lines in the disk are equal to the corresponding hyperbolic angles, however, areas and distances are warped in it. In the Poincaré disk, hyperbolic geodesics, i.e., shortest paths between two points at the boundary $\partial\mathbb{D}$, appear either curved as arcs of Euclidean circles that intersect $\partial\mathbb{D}$ perpendicularly, or straight as disk diameters (Fig. 3-5 A). As depicted in Fig. 3-5 B, straight lines appear curved in the Poincaré disk, an observation justified by the relation $r_e = \tanh \frac{r_h}{2}$, which implies a distortion between r_e (euclidean distance from the disk center) and r_h (hyperbolic distance from the disk center). This distortion, is also the reason why hyperbolic distances grow exponentially towards the boundary $\partial\mathbb{D}$, and so why distances are shorter close to the disk center (and therefore faster to move through) than to the boundary. The function of the hyperbolic distance between

two points z_i, z_j in this model is given by the formula:

$$d(z_i, z_j) = \text{arcosh}\left(1 + 2\frac{\| z_i - z_j \|^2}{(1- \| z_i \|^2)(1- \| z_j \|^2)}\right)$$

where $\|\cdot\|$ represent the Euclidean distance ($L2$ norm).

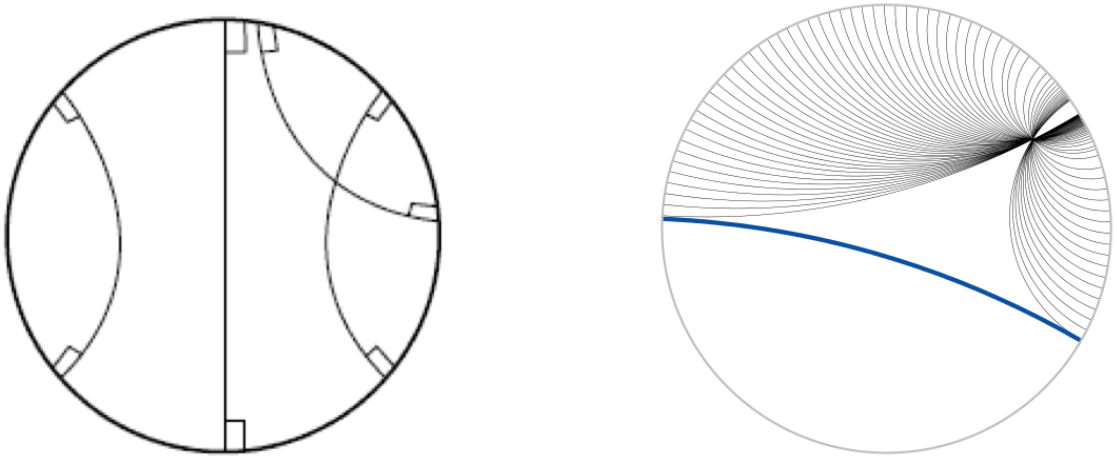

Figure 3-5: **(A)** Geodesics on the Poincaré disk, which include arcs meeting the edge of the disk at 90° (orthogonal) and diameters of the boundary circle. **(B)** [52] Straight lines in the Poincaré disk that pass through a given point, do not intersect and are parallel to the blue line.

3.2.3 Hyperbolic Space for Complex Networks Embedding

A distinctive advantage of the Hyperbolic Space is the property of "exponential scaling" instead of polynomial, with respect to the radial coordinate. This serves Big Data analysis and complex network embedding. Choosing the Hyperbolic over the Euclidean Space allows the integration of much more data in a much more "compact" area, like the Poincaré disk where the entire space is represented on the surface of the unit disk. So, Hyperbolic Space manages to condense more surface area within a given radius than flat or positively curved geometries.

Particularly, in a two-dimensional space $\mathbb{H}^2_\zeta$ with constant curvature $K = -\zeta^2 < 0, \zeta > 0$, the circumference C and the area A of a disk of radius r are:

$$C(r) = 2\pi \sinh(\zeta r)$$

$$A(r) = 2\pi(\cosh(\zeta r)–1)$$

Therefore, in case of a Poincaré disk ($\zeta = 1$), for small values of radius r (close to the disk center) the Hyperbolic Space appears flat, while for larger r, both C and A grow exponentially (asymptotically) with respect to r [9, 53], which makes it possible to locally approximate any Hyperbolic Geometry using Euclidean geometry. These values should be contrasted to the corresponding Euclidean quantities that expand polynomially, as $2\pi r$ and πr^2 respectively (for the two-dimensional case).

The aforementioned property of exponential space expansion, allows the Hyperbolic Space to fit and describe complex scale-free networks. In such networks, as described in section 2.2, the degree distribution follows a power-law, a relationship that can be modelled as $P(k) \propto k^{-\gamma}$, where $P(k)$ the fraction of nodes with k degree in the network, and γ a constant scaling parameter. Complex networks, and especially biological processes, pathways and protein complexes could be organized using hierarchical representations, such as dendrograms and Venn diagrams. These two perspectives are equivalent and connected through Hyperbolic Geometry. Namely, a branch of a dendrogram could be represented as a circle in a Venn diagram. Thus, the larger a circle in the Venn diagram the broader the group of entities it includes and the closer to the tree root the position of the corresponding branch in the dendogram. Also, the more two circles overlap, the more similar are the sets and the nodes and the smaller their hyperbolic distance. In case of partial instead of total overlapping, the resulting tree structure will contain a loop. However, as long as it approximates a tree hierarchy, it is negatively curved [54] and describable by Hyperbolic Geometry.

As stated by Krioukov *et al.* (2010) in [9], Hyperbolic Geometry exists underneath complex networks with properties like power-law degree distribution and community structure arising from the negative curvature of the space, while vice versa a scale-free network with a metric structure could be described by a Hyperbolic Geometry. An extension of this semantic connection between scale-free networks, tree-like structures and Hyperbolic Geometry, is the relation between the power-law degree distribution scaling exponent and the negative space curvature [9] as well the fact that both the branching factor of a tree and the curvature of the Hyperbolic Space are measures of how fast the space expands.

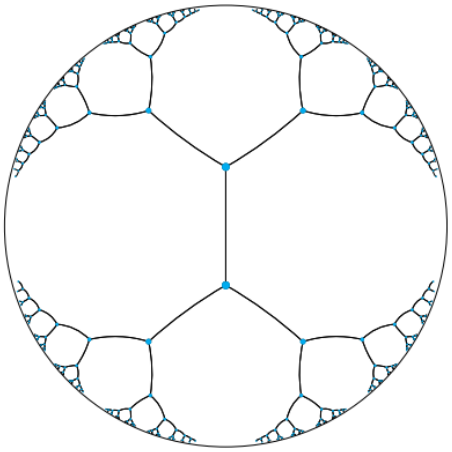

Figure 3-6: [55] Embedding of a regular tree (branching factor=2) in the Poincaré disk. Because of the negative curvature, distances between all points are actually equal, since they grow exponentially as you move toward the edge of the disk. Disk radius spanning is consistent with the network growth (more nodes) and the network hierarchy deepening.

In an n-ary tree (a tree with a branching factor n), the number of nodes at distance r from the root are $(n+1)n^{r-1}$, and the number of nodes at distance no more than r from the root are $\frac{(n+1)n^r-2}{n-1}$. Both of these quantities grow as n^r with r, thus, the metric structures of n-ary trees and $\mathbb{H}^2_\zeta$ are the same if $\zeta = ln(n)$, and can be considered as equivalent [9]. In conclusion, trees require an exponential space to grow, and this is what causes "crowding" effects in case of Euclidean embedding, while highlighting the adequacy of Hyperbolic Space (Fig. 3-6).

Chapter 4

Hyperbolic Space Embedding Algorithms

In order to examine the main embedding frameworks used in the present thesis, we have to define *network embedding*, so for this purpose, we use the definition given by Cvetkovski [56]: "Given a connected finite graph G with vertex set V, a hyperbolic embedding of G in $\mathbb{H}^d$ is a mapping $C(G) :\rightarrow \mathbb{H}^d$ that assigns to each vertex $v \in V$ a virtual coordinate $C(v)$." Network embedding aims to map and represent graph nodes into a low-dimensional latent space. This method should preserve the graph structure while reducing effectively sparsity and noise of the corresponding adjacency matrix. It can be used in graph analysis tasks, such as node classification, clustering, community detection, link prediction and visualization.

4.1 Laplacian-based Network Embedding Algorithm

Driven by the manifold unfolding problem, Belkin and Niyogi [57] proposed a geometrically inspired model, called Laplacian Eigenmaps (LE), for the representation of data lying in a low-dimensional manifold embedded in a higher-dimensional space. Noting that if matrix D is the degree matrix of the graph, that is $D = diag(\sum_j A_{ij})$, then the Laplacian matrix can be defined as the difference of the degree matrix and the adjacency matrix, $L = D–A$. Laplacian Eigenmaps algorithm, in contrast to other dimensionality reduction techniques such as Principal Component Analysis [58], considers the intrinsic data geometry and aims to preserve locality, being particularly stable to outliers and noise.

This model was the basis of the Laplacian-based Network Embedding (LaBNE) algorithm, proposed by Lobato *et al.* [59], where instead of the previous proximity preservation, the authors employed the eigen-decomposition of the Laplacian graph to infer hyperbolic coordinates and embed complex networks in the Poincaré disk model. In the original method [57], data points correspond to graph nodes and their in-between connections is ruled by neighbors proximity. In LaBNE, since actual euclidean distances between the pairs of nodes are not available, their heuristic estimations are used as similarity scores in an objective function that gives large penalties if two nodes with larger similarity are embedded far apart in the embedding space.

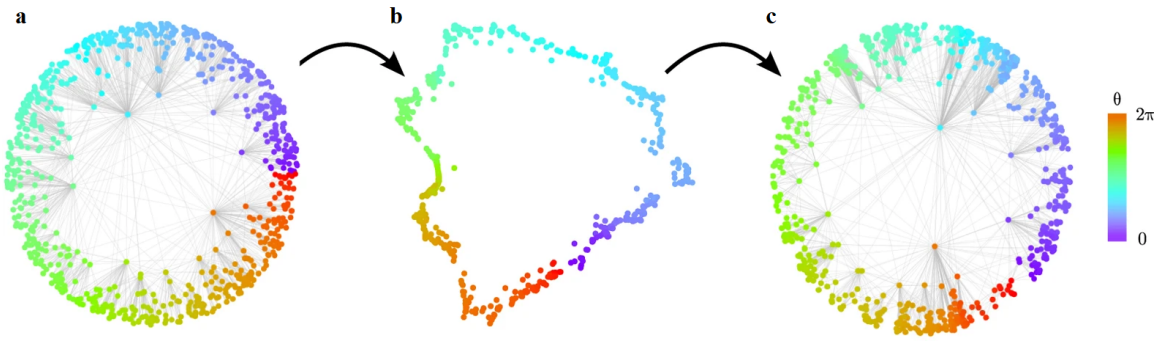

Figure 4-1: [59] Laplacian-based Network Embedding. A network generated with the PSO model (*a*) is embedded via LaBNE to the hyperbolic circle, revealing the angular coordinates of the nodes (*b*). Then, the radial coordinates are assigned, resembling the degree ranking (*c*). Although, the final depiction of the embedded network presents an angular rotation of the nodes in contrast to the initial PSO-generated graph, the distance-dependent connection probabilities remain consistent.

Embedding of the LE methods is performed on Euclidean Space, though the latent geometry of real complex networks is observed to be hyperbolic. Due to the conformal character of the Poincaré disk model (euclidean angles are equal to corresponding hyperbolic angular separation of nodes), the LaBNE algorithm is able to designate angular coordinates based on the inferred similarity subspace, while radial coordinates are typically assigned to resemble the degree ranking, accounting for their popularity. That popularity dimension is part of the Popularity-Similarity model, assumed to describe the network formation, making the embedding result model-dependent.

4.1.1 The PSO Model for Network Generation

Popularity-Similarity Optimization (PSO) model as described by Papadopoulos *et al.* [60], generates synthetic graphs which exhibit common structural and dynamical characteristics of real complex networks, such as strong clustering and scale-free degree distribution. The emergence of these properties is the outcome of an optimization process involving the trade-off between node popularity and similarity. These two, are both considered as measures of attractiveness, with the popularity reflecting the node's seniority and its ability to attract connections from other nodes over time, while similar nodes are very likely to connect, regardless of their rank. PSO model has a geometric interpretation in Hyperbolic Space, where the popularity-similarity trade-off is abstracted by the hyperbolic distance between nodes, leading to distance-dependent connection probabilities and link formation [9, 60]. Thus, hyperbolic embedding of a network reveals the value of the variables and parameters contributing to its topology (popularity and similarity in this model), facilitating the understanding of system's growth process.

In the proposed actively growing network model, nodes are introduced iteratively at logarithmically increasing distance from the origin of the native disk representation (radial coordinate) and at an uniformly random angular coordinate. At each step the new node connects to the pre-existing ones with a probability decreasing with the hyperbolic distance, taking into consideration the popularity of the older nodes and its similarity to the rest of the nodes. Apart from the final number of nodes N in the network, the parameters of the model can be listed as follows:

- $\zeta = \sqrt{-K}$, where $K < 0$ the curvature of the hyperbolic plane. It doesn't affect the properties of the generated network, apart from a simple rescaling of the hyperbolic distances. Usually is set to a constant, e.g. 1.
- m: the number of connections that each newly appearing node will form, corresponding to the half of the average degree, $k = 2m$.
- $\beta \in (0, 1]$: popularity fading parameter that controls nodes' drifting away from the center and determines the value of the exponent γ of the power-law degree distribution $P(k) \propto k^{-\gamma}$ of the network as $\gamma = 1 + \frac{1}{\beta}$.
- $T \geq 0$: temperature controls average clustering c of the network, which

gets its maximum value when $T = 0$, gradually decreases towards zero at $T = 1$ and becomes asymptotically equal to zero for $T > 1$.

Assuming an initially empty network in a hyperbolic plane of curvature $K = -1$, the algorithmic steps of the procedure are as follows:

1. At time $i \geq 1$ with $i = 1, 2...N$, a new node i appears at polar coordinates (r_i, θ_i), with $r_i = 2ln(i)$ relating to node birth and θ_i sampled randomly and uniformly from $[0, 2\pi)$. Every existing node $j < i$ increases its radial coordinate by $r_j(i) = \beta r_j + (1 - \beta)r_i$, to simulate popularity fading.

2. New node i selects a subset of previously appeared nodes to connect to. If the amount of pre-existing nodes is not larger than m, node i connects to all of them, alternatively node i connects to m nodes, with distance-dependent connection probability $p_{ij} = 1/\left(1 + e^{(x_{ij} - R_i)/2T}\right)$, where x_{ij} the hyperbolic distance between nodes i and j, $x_{ij} = \text{arcosh}(\cosh r_i \cosh r_j - \sinh r_i \sinh r_j \cos\theta_{ij})$ $\approx r_i + r_j + 2ln(\theta_{ij}/2)$ (such approximations often become exact in the large graph size limit), θ_{ij} the angular distance between nodes i and j, $\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$ and R_i the radius of the hyperbolic disk that encloses the network at the current time is set to:

$$R_i = \begin{cases} r_i - 2ln(\frac{2T}{\sin(\pi T)}\frac{(1-e^{-(1-\beta)r_i/2})}{m(1-\beta)}) & \text{if } \beta < 1 \text{ and } T > 0 \\ r_i - 2ln(\frac{2(1-e^{-(1-\beta)r_i/2})}{\pi m(1-\beta)}) & \text{if } T \to 0 \\ r_i \text{ \& existing nodes do not move,} & \text{if } \beta = 1 \end{cases}$$

3. Previous steps are repeated until total of N nodes joined the network.

4.2 HyperMap Embedding Algorithm

In contrast to networks generated by the PSO model where link formation takes place between new and old nodes, in many real networks, new links could also appear at a certain rate between old nodes (internal links) as well. In order to account for this deficiency, Papadopoulos *et al.* [61] proposed the E-PSO model, a generalized version of PSO model that takes into consideration and effectively distinguishes external and internal links. It reproduces the scale-free degree distribution and clustering of real networks, but also several other important properties including the average neighbor degree, the distribution of hop length

of shortest paths and the average node betweenness [60]. This model is the basis of another hyperbolic Poincaré embedding, the HyperMap algorithm [61]. HyperMap is a Maximum Likelihood Estimation (MLE) framework, where a search space of PS models is explored in order to find the best fit for the topology of the network of interest, ensuring a better embedding quality. This exploration is greatly precise, though computationally heavy.

In this method, the whole network is initially disassembled to disconnected nodes, and the next steps are as follows:

- Nodes are sorted in decreasing order of their degree, $k_1 > k_2 > ... > k_t$, and indexed accordingly, $i = 1, 2, ..., t$.
- When node $i = 1$ is born (the one with the largest degree), is initially mapped at the origin of the hyperbolic disk, with radial coordinate $r_1 = 0$ and angular coordinate θ_1 randomly sampled from $[0, 2\pi]$.
- For the rest of the nodes $i \in [2, t]$, that are introduced one by one, node i is assigned an initial radial coordinate $r_i = 2ln(i)$, while every pre-existing node $j < i$ updates its radial coordinate according to the concept of popularity fading by $r_j(i) = \beta r_j + (1 - \beta)r_i$. The angular coordinate is chosen by maximizing node's local likelihood that the network is generated by the E-PSO model, given by equation:

$$\mathcal{L}_i = \prod_{1 \leq j < i} p(x_{ij})^{\alpha_{ij}} [(1 - p(x_{ij})]^{1-\alpha_{ij}}$$

We note that a_{ij} is the corresponding value in the adjacency matrix, where $\alpha_{ij} = 1$ if $\{u, v\} \in E$ and $\alpha_{ij} = 0$ if $\{u, v\} \notin E$. Only the pre-existing nodes contribute to the product and the likelihood is eventually a function of θ_i, since $p(x_{ij})$ depends on x_{ij}, which in turn depends on θ_i. In an improved version of this approach further periodic correction steps are also applied for better adjustment of the angular coordinates.

4.3 Angular Optimization

Driven by the work of Alanis-Lobato *et al.* [62] and their approach of combining LaBNE and HyperMap strategies, in order to balance between computational speed and accuracy, we examined a different version of refinement, pursuing a more efficient and accurate network embedding. An assumption that the network to be embedded was generated according to the PSO model, was made. The angular optimization process includes the following steps (pseudocode is described further on and the code is available upon request):

- Calculate the global logarithmic loss, at network level:

$$LL = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} A_{ij} ln[p(x_{ij})] - \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} (1 - A_{ij}) ln[1 - p(x_{ij})]$$

  where A_{ij} corresponds to adjacency matrix values (1 if nodes i and j are connected, 0 otherwise), while $p(x_{ij})$ is the distance-dependent connection probability, equal to $p(x_{ij}) = 1/\left(1 + e^{(x_{ij} - R_N)/2T}\right)$. Also, R_N is the radius of the hyperbolic disk enclosing the network at the end of the network generation process and T is the network temperature. Lastly, x_{ij} is the hyperbolic distance between nodes i and j, equal to $x_{ij} = \text{arcosh}(\cosh r_i \cosh r_j - \sinh r_i \sinh r_j \cos\theta_{ij})$, where $\theta_{ij} = \pi - |\pi - |\theta_i - \theta_j||$ is the angular distance between them.

- Iterate over all nodes.

- Examine new potential angular coordinates inside a window defined by the angular positions of the second neighbors of each specific node, with a step properly chosen to distribute the tested positions in equal distances.

- Update the current angular position only if a lower logarithmic loss contribution value was found (minimization problem). In this case, update the logarithmic loss contribution of every other node as well the angular positions of the adjacent nodes.

- The whole optimization process can be repeated as many times as defined by the user, who also sets the number of each node's potential angular positions.

Algorithm 1 Angular Optimization

Input: network, inferred polar coordinates, PSO parameters (γ, T), the number of correction rounds, the number of each node's tested angular positions in each round

Output: optimized polar coordinates

Compute the average node degree of the network ($2m$), the popularity fading parameter (β) and the adjacency matrix (A)

if $T = 0$ **then return** inferred polar coordinates

else

if $\beta = 1$ **then**

$radius \leftarrow 2ln(N)$

else

$radius \leftarrow 2ln(N) - 2ln(\frac{2T}{\sin(\pi T)} \frac{(1-e^{-(1-\beta)2ln(N)/2})}{m(1-\beta)})$

end if

Calculate the initial global logarithmic loss and each node's contribution

$$LL = -\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} A_{ij} ln[p(x_{ij})] - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (1 - A_{ij}) ln[1 - p(x_{ij})]$$

Calculate current angular node order

Calculate current 1st and 2nd neighbors of each node

for i in optimization rounds **do**

for n in range(N) **do**

Determine the angular arc between node n's 2nd neighbors

Determine the candidate equidistantly distributed angular positions

for a in candidate angular positions **do**

Calculate node's n log-loss when placed in the angular position a

end for

if tested log-loss $<$ current log-loss **then**

Update the angular position of node n with the one minimizing log-loss and also update the angular position of node n's neighbors

Based on its new position, update the log-loss' contribution of the node n as well the one of the other nodes

else

The current angular position is the best and nothing changes

end if
end for
end for
end if
return updated polar coordinates

In machine learning, logarithmic loss (or log-loss) is indicative of how close the prediction probability is to the corresponding ground truth (0 or 1 in case of binary classification). In the case above, log-loss represents the divergence between the connection probability of two nodes $p(x_{ij})$ and their corresponding actual/true value in the graph's adjacency matrix A_{ij}. The more those two values differ, the higher the log-loss value. Log-loss can alternatively be defined as the negative log-likelihood function.

4.4 Rigel Embedding Algorithm

Against the previous approaches which focus on fitting the graph of interest to a network model, Rigel algorithm [63] aims to an embedding where distances between nodes in the Hyperbolic Space imitate the geodesic paths in the original graph. More specifically, node distance measurements are approximated through a Graph Coordinate System (GCS), which embeds nodes of a high dimensional graph (extremely high number of dimensions that make calculations excessively time-consuming) into positions in a fixed-dimension coordinate space. After accomplishing the embedding of network G into the coordinate space, a GCS can approximate node geodesic distance queries in a small amount of time and independently of the graph size, i.e. fixed $O(1)$ time. The initial step is computationally expensive and scales with graph size, demanding $O(N)$ time for a graph of size N. Also, in opposition to the Poincaré disk model that was used before, Rigel is based on the Hyperboloid model. Calculating the distance between two points in this model is computationally simpler than in other hyperbolic models, and its complexity is independent of the curvature of the space. So, the distance between two n-dimensional points x and y, with curvature parameter c, is given by:

$$\delta(x,y) = \textbf{arccosh}\left(\sqrt{\left(1+\sum_{i=1}^{n} x_i^2\right)\left(1+\sum_{i=1}^{n} y_i^2\right)} - \sum_{i=1}^{n} x_i y_i\right)|c|$$

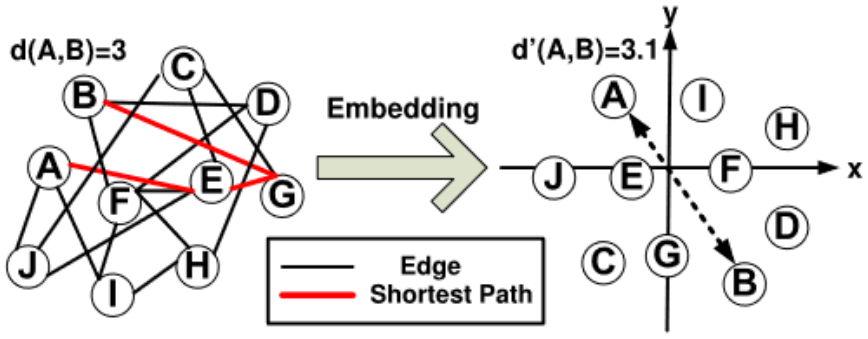

Figure 4-2: [63] Graph Embedding to an Euclidean Space via Rigel algorithm, aiming to preserve the shortest path distance in the coordinate space. For example, the geodesic distance between nodes A and B is 3 (left) and after the embedding its corresponding Euclidean distance between their points is 3.1 (right).

The embedding process is achieved as follows:

- A subset of l nodes is selected as landmarks, where $l << N$, and N the total number of nodes in the network. Those reference points were chosen as high-degree nodes, since the computational complexity of degree centrality is low, while its results remain efficient in relation to other approaches [64].
- For each of the landmarks, shortest path trees to all other nodes are calculated, using Breadth-First-Search (BFS) algorithm.
- "Bootstrapping" step: a general optimization algorithm is used in order to determine the coordinates of these nodes, while aiming for distances in the coordinate space that adequately approximate the corresponding geodesic paths in the graph (Fig. 4-2).
- A random subset of $k < l$ nodes from the landmarks set is chosen for the mapping of each remaining node. These nodes are assigned to coordinates that minimize the deviations of the distances between the node and the corresponding k nodes between the coordinate space and their actual hop distance in the graph. Simplex method from linear programming, is used for the optimization.

Both curvature of the coordinate space and the number of dimensions play an important role in embedding accuracy. In particular, curvature affects the level of distortion between the original node distances and their mappings on the Hyperbolic Space. As shown in [63], curvature of -1 provides the most accurate results, so this value will be used during the execution of Rigel algorithm in the context of this thesis. At the same time, an increase of the dimension of Hyperbolic Space leads to an increase in accuracy. Noting that the number of landmarks should be greater or equal to the chosen dimensions, there should be a balance between accuracy and complexity.

Chapter 5

Materials and Methods

5.1 Network Topologies

For the best possible documentation of the results, a number of different graph datasets were used, some synthetic and others real. The real networks were selected from the Stanford Network Analysis Project (SNAP) [65] as well the Biological General Repository for Interaction Datasets (BioGRID) [66]. Only the largest connected component was considered, while self-loops and multiple edges were discarded and edge directionality or weight were ignored. Information along with the key properties of these networks are presented below and in Table 5.1.

The autonomous systems (AS) correspond to sub-graphs of routers comprising the Internet. AS are constructed based on the BGP (Border Gateway Protocol) logs, that reveal a who-talks-to-whom communication network. The dataset [67] is the daily instance of January 2 2000 and the data were collected from University of Oregon Route Views Project (www.routeviews.org). This AS topology is part of a collection of 733 AS graphs, and is available for download at https://snap.stanford.edu/data/as-733.html.

The LastFM network corresponds to a social network of LastFM Asia which was retrieved in March 2020 from the public API. Nodes are LastFM users from Asian countries and edges are mutual follower relationships among them. The used dataset [68] is available for download at https://snap.stanford.edu/data/feather-lastfm-social.html.

Lastly, the human interactome is a BioGRID multi-validated dataset (MV), excluding genetic interactions or physical interactions that failed to pass a

specific set of criteria. These criteria concern the participants of each interaction and are summarized to the following: different Experimental System as well as Publication Source, different Experimental System when the Publication Source is the same and different Publication Source when the Experimental System is the same. To acquire the dataset one could download the BIOGRID-MV-Physical zip file at https://downloads.thebiogrid.org/BioGRID/ (4.3.196 version, Tab 3.0 format), and keep only the records concerning human interactions, by filtering the Organism ID Interactor field to be equivalent to '9606'. BioGRID is an open access repository that archives protein-protein, genetic and chemical interactions and post-translational modifications. Data relate to all major model organisms and human, and are comprehensively curated from the biomedical literature.

Six synthetic networks were generated based on the Popularity-Similarity Optimization (PSO) model as described in section 4.1.1. These networks were grown for a broad range of parameters (Table 5.2), including: number of nodes in-between $N = 500 - 5000$, quite low γ and temperature values - around $2.1 - 2.3$ and $0.1 - 0.3$ respectively, and average degree analogous to the number of nodes. With these parameters' combinations, the cold-temperature regime ($\gamma = 2$ & $T = 0$) was avoided, while a community structure with quite strong clustering, without "loose-ends" (nodes with only one connection) or high density was preserved, in order to study network dynamics. The code is available upon request.

Table 5.1: The number of nodes N, edges L, average node degree $2m$, scaling exponent γ, temperature T and average clustering coefficient $\bar{c}$ are reported for each real network.

Network	N	L	$2m$	γ	T	$\bar{c}$
Autonomous System	6474	12572	3.88	2.06	0.69	0.25
LastFM Asia	7624	27806	7.29	2.58	0.73	0.22
H. sapiens PPIs	10179	54518	10.71	2.27	0.8	0.17

Table 5.2: The number of nodes N, edges L, average node degree $2m$, scaling exponent γ, temperature T and average clustering coefficient $\bar{c}$ are reported for each network generated based on the PSO model. The values marked in brackets are the ones observed in the final topology and not the target ones.

Network	N	L	$2m$	γ	T	$\bar{c}$
G1	500	(1092)	4 (4.37)	2.1	0.10	0.59
G2	500	(1026)	4 (4.10)	2.3	0.30	0.48
G3	1000	(3369)	6 (6.74)	2.1	0.05	0.59
G4	1000	(3062)	6 (6.12)	2.3	0.35	0.42
G5	2500	(10506)	8 (8.40)	2.2	0.20	0.58
G6	5000	(25858)	10 (10.34)	2.2	0.20	0.60

5.2 Real Network Parameters Estimation

Network's scaling exponent (γ) and temperature (T) are associated with different statistical and topological features of the graphs generated by the PSO model. Thus, their estimation could be accomplished by observing the properties of the network to be embedded or by reproducing similar topologies to detect those features. Therefore, the estimation of the scaling exponent γ was accomplished by fitting a power-law to the degree distribution of the network. Aaron Clauset *et al.* addressed the issue of fitting power-laws to distributions [69]. An implementation of their algorithm with Python language was used in this thesis (the code is available at https://pypi.org/project/plfit/). The following function implements maximum likelihood estimators for fitting the power-law distribution to data. It accepts as input the list of networks degrees, and the output of interest is the $\_alpha$ attribute corresponding to the power-law exponent (γ).

```
myplfit = plfit.plfit(network_degrees)
myplfit._alpha
```

To continue, temperature appears as a natural parameter controlling clustering in the network. Its estimation is based on the statement that clustering is maximized at $T = 0$, almost linearly decreases to zero at $T = 1$ and remains asymptotically zero for any $T \geq 1$ [9, 60]. For each real network, a linear least-squares regression was calculated to model the relationship between temperature

and clustering coefficient. The point ($T = 1, \bar{c} = 0$) was used as x-intercept, while y-intercept corresponded to the average clustering coefficient of 10 artificial networks generated by the PSO model using $T = 0$ and the same structural properties as the network of interest. After determining the regression line, its slope and intercept values are used along with the real clustering coefficient to compute the real network's temperature.

5.3 Hypermap Algorithm

HyperMap is a Maximum Likelihood Estimation approach. Its main purpose is the exploration of the space of PSO models with structural parameters similar to the network of interest, in order to find the best topological match. This process is very accurate, albeit computationally demanding. HyperMap is implemented in C++ by the DK Lab, while a R wrapper of it is available at https://github.com/galanisl/NetHypGeom.

Algorithm 2 HyperMap embedding

```
hm <- labne_hm(net= network, gma= 2.06, Temp= 0.69, w= 2*pi)
```

To use the HyperMap algorithm, the R packages *NetHypGeom* and *igraph* were installed and loaded into R working space. To embed the network, the function *labne_hm()* was called by setting LaBNE+HM's window to $2 * \pi$. The function also expects as input the network's scaling exponent (gma) and the network's temperature ($Temp$). These parameters were determined prior to the embedding step, as described in section 5.2. Speed-up heuristic was set to its default value 10 (also referred to as fast hybrid version of HyperMap), while correction steps were not applied through this thesis, since their effect has been reported as not significant [70].

5.4 LaBNE Algorithm

LaBNE is a Laplacian-based embedding of a complex network to the Poincaré disk. It is based on a non-linear dimension reduction of the Laplacian matrix, followed by nodes' organization, with the radial coordinates resembling the nodes' degree ranking and the angular coordinates obtained from a few algorithmic

steps (including solving an eigenvalue problem). LaBNE is extremely fast, with low computational complexity. However, it can be inaccurate by solely focusing to keep connected nodes as close as possible, disregarding that disconnected nodes should be far from each other as well. It highly depends on topological information, so performs better when the average node degree and clustering coefficient of a network are high [59].

LaBNE is implemented in R, exploiting the R package *RSpectra* which contains high-performance solvers for large-scale eigenvalue problems. The spatial regression step is the most complex and time-consuming. As the algorithm of *RSpectra* is generally better at finding large eigenvalues, the appropriate way is to utilize the spectral transformation and calculate the largest eigenvalues of A^{-1}, whose reciprocals are exactly the smallest eigenvalues of A.

Algorithm 3 LaBNE embedding

```
labne <- labne_hm(net = network, gma = 2.06, w = 0)
```

The code of LaBNE is available at https://github.com/galanisl/NetHypGeom. As seen above for the process of network embedding, the function *labne_hm()* was called by setting LaBNE+HM's window to 0. The function expects only one additional input, the network's scaling exponent (gma). Its value is automatically computed by fitting a power-law distribution to the network's degrees dataset, unless it is specified by the user.

5.5 Angular Optimization

LaBNE produces a draft geometric configuration of the network of interest to the Hyperbolic Space. Then, this configuration is passed on a refinement step in order to improve the inferred angular coordinates and produce the final mapping. The reason behind this approach is to benefit from LaBNE's fast embedding and reduce the search space of possible angular coordinates that the next algorithmic step will have to explore in order to minimize a logarithmic loss function.

Algorithm 4 Angular Optimization

```
angOpt <- angular_optimization(network, coordinates, gamma,
temperature, candidates=5, reps=3)
```

The inputs and parameters needed to execute the angular optimization, are the network, the previously inferred coordinates (by LaBNE or any other algorithm complying with the same PSO model), the network's scaling exponent $gamma$, the temperature T, and user's choice on the number of iterations and potential new angular positions. The optimization process with the iteration over all nodes can be repeated a few times until the angular coordinates settle in an optimum position. The definition of the range of possible new positions in-between the second neighbours is based on the assumption that only minor adjustments will be needed to improve an already quite good embedding.

5.6 Rigel Algorithm

Rigel is designed as a command line tool, including two phases -the embedding and the querying. Rigel's embedding phase is written in C++ Standard Template Library, and is entirely self-contained with no additional dependencies. Only CMake is required for its compiling. Rigel expects 0-based indexing of the nodes and its execution is split in two steps: embedding of landmark nodes (bootstrapping), followed by embedding of the rest of the nodes against those landmarks. Example commands of this implementation are the following:

Algorithm 5 RIGEL embedding

Landmark embedding

```
$ ./rigel.exe -b 16 -e -1 -i 16 -L 30 -l distanceMatrix.txt
-o 30L10D -r landmarks.txt -t serially -u 10179 -x 10 -y -1
```

Non-landmark embedding

```
$ ./rigel.exe -b 16 -e -1 -i 16 -L 30 -l distanceMatrix.txt
-o 30L10D -r landmarks.txt -t serially -u 10179 -x 10 -y 0
```

Specifically, the first command embeds 30 landmarks ($-L$) of a graph with 10179 nodes ($-u$) into a 10 ($-x$) dimensional Hyperbolic Space with curvature -1. The top 16 out of the 30 landmarks are embedded first ($-i$) and the rest of

them are aligned against them. The second command describes the embedding of the non-landmarks nodes. Every one of these is aligned against a random selection of 16 landmarks ($-b$). The whole process runs serially.

In order to find a set of optimal parameters, concerning the number of dimensions and landmarks, dimensionality values between 2 and 14 along with 30 and 50 landmarks were examined, for every dataset. For the landmarks' selection, two strategies were tested: high-degree and high-closeness centrality. For the evaluation of Rigel's accuracy of the estimated distances between nodes, the Average Relative Error (ARE) metric was used [64]. Small values indicate realistic predicted distances. Relative Error is calculated by:

$$RE = \frac{|d_{x,y}^{measured} - d_{x,y}^{predicted}|}{d_{x,y}^{measured}}$$

, where $d_{x,y}^{predicted}$ is the estimated distance in the embedding space, computed using x and y's coordinates based on the Hyperboloid model, while $d_{x,y}^{measured}$ is the actual shortest path distance between x and y on the real graph. Intuitively, a larger vector of coordinates would lead to more precise distance estimations and smaller relative errors. Apparently, using 30 landmarks and 10-dimensional coordinates was the best combination for the total of the examined datasets. Regarding the selection of landmarks, both approaches had very similar ARE values. The high-degree strategy was adopted for the presented results, because it was computationally faster.

5.7 Evaluation Criteria of Hyperbolic Embedding

Network embedding techniques aim to simplify network's interpretation, visualize graphs despite the ever-growing amount of data, and highlight the structural features of complex networks in the geometric (hyperbolic) space. Those representations facilitate downstream applications including greedy routing and link prediction. These tasks are also considered to evaluate the quality and efficacy of each embedding method.

5.7.1 Greedy Routing

An important structural property of any network is its navigability, which is associated with a navigation technique called greedy routing (GR). This task examines the possibility of shipping information from a source node to a target node, without global knowledge of the network topology, using exclusively local topological information. At every stage of the process, every node knows its own 'address' as well the ones of its adjacent nodes, while every routed packet includes the 'address' of its target. Given these conditions the employed protocol in this thesis proceeds as follows: the source node computes the hyperbolic distance using nodes' coordinates, from itself and from every adjacent node to the target, and ships the packet (also referred to as information or commodities) to its neighbor that is closer in terms of hyperbolic distance to the target node than the node itself. This step is repeated until the packet reaches the target. In this case the delivery is considered successful and the whole process is known as greedy routing or greedy forwarding. However, if the packet is sent to a previously visited node (and so creates a loop), then the packet is dropped and the delivery is considered unsuccessful [9, 71, 72].

The fact that greedy routing uses only local information results in reduction of the complexity imposed by shortest path computations, thus making it suitable for large-scale systems [63]. A disadvantage of greedy routing emerges when a packet gets stuck at local minima of distance, where a node closer to the destination than itself does not exists [73]; but this can be avoided by the appropriate choice of hyperbolic network embedding and nodes' coordinates. For the greedy routing task, a good embedding would be the one that achieves a high rate of successful deliveries which at the same time, are similar in size to the true shortest paths. A graph embedding where greedy routing is successful for every pair of source and target nodes is called greedy embedding. Conversely, in a greedy embedding, for every pair of nodes a distance-decreasing path exists. In the context of this thesis, in order to evaluate the performance of each embedding algorithm, the percentage of successful paths (namely successfully delivered packets) was measured, while the average hop length of the successful paths was compared to the mean shortest path length.

5.7.2 Link Prediction

Another way to evaluate the performance of graph embedding algorithms is link prediction. It could be viewed as a way to predict lost, impending and spurious links of graphs or to convey the degree that an evolutionary process could be modelled based on the topological properties of a network [74]. The goal is to estimate the likelihood of the existence or the non-existence of a non-observed link between a pair of nodes in the network, based on the observed links as well the attributes of nodes. Since the connection probability is a descending function of the hyperbolic distance between two nodes (as previously described), then the predicted links are more likely to exist between nodes closer in terms of hyperbolic distance.

In order to evaluate the accuracy of each link prediction technique, a common framework was employed. Specifically, a subset of links from a given network was removed at random, and the ability of each examined algorithm to predict these missing links using the incomplete data was tested [8]. In greater detail:

- L links, equal to 10% of the total edges, were removed uniformly at random from the observable network topology
- The link-prediction technique assigned a confidence score to each non-observed link of the pruned network to quantify its existence likelihood and then sorted them decreasingly. The better the score, the better the candidate interaction and the higher in the list of the predicted links.
- L putative edges from the top of the sorted list were selected and their proportion included in the initially removed set of edges (step 1) was computed, as indicative of algorithm's precision. For the next chapters, this will be referred to as link-prediction score.

Apart from the above, another framework to assess algorithm's performance and quantify its accuracy of link-prediction over a variety of thresholds, is the precision-recall (PR) curve. In order to plot the PR curve, the sorted list of putative edges was scanned with a moving score threshold to compute the fraction of predicted links that actually belong among the removed edges (precision), along with the fraction of predicted links out of the total of removed edges (recall-sensitivity). Precision is represented as $\frac{TP}{TP+FP}$, and Recall as $\frac{TP}{TP+FN}$, where TP is the number of true positives, FP the number of false positives and FN the number of false negatives.

5.7.3 Semantic Similarity

Apart from the aforementioned criteria, in the case of the PPIs network we applied another methodology in order to detect the potential association between the hyperbolic distances and the functional divergence of proteins on the PPIs network. Namely ad hoc workflows were constructed to examine if the proposed hyperbolic embedding algorithm is able to pose small distances between functionally relevant proteins, comparing them with randomly selected protein sets. For this purpose, biomedical databases, which contain the functional annotation of proteins and measures that perform calculations on these semantic schemes where used.

Generally, the identity and the role of a protein in the cellular system are defined by a plethora of features. Some of them are estimated through experiments and encoded in digitized formats, such as the amino acid sequence which is represented as a string of letters. On the other hand, the functionality or the contribution of a protein variation in a disease can be represented only using semantics. Therefore, various biomedical ontologies have been constructed to describe the existing knowledge under a specific domain of biology and provide an apparatus of controlled vocabularies to define the functional and phenotypic characteristics of proteins. The association of proteins with semantic terms is usually called genomic annotation and it could be assigned using either experimental results or reference and phylogenetic-based approaches. Gene Ontology (GO) is the foremost biomedical ontology for the functional annotation of the proteomic universe, as it describes in various layers the role of proteins in cells, for thousands of species [75]. GO is separated into three main subdomains: molecular function (MF), biological process (BP) and cellular component (CC). Namely, according to GO, the functionality of a protein is drawn by three sets of semantic terms: the molecular functions where the protein participates, the biological processes (broader mechanisms and pathways) where these interactions are encompassed and the components of cellular topology, where these interactions are performed.

In order to estimate the functional similarity of two proteins, semantic similarity measures are used to quantify the relatedness of their semantic annotation in the same ontological domain [76]. Although the semantic similarity measures are separated in node-based, edge-based and hybrid measures, all of them take advantage of the graphical representation of the ontology to calculate

the similarity of two terms. In general, node-based approaches assume that the similarity of two terms is proportional to the shared semantic component between them, which is estimated using their ancestral branches on the graph [77]. Then, aggregated measures could be used to calculate the semantic similarity of two proteins, averaging the pairwise similarities of their annotations. In this study, two different workflows were constructed exploiting the GO annotation of proteins to evaluate the predictive potential of the used embedding algorithm. The former one was used to assess the question if functional related proteins have the same hyperbolic distance compared to random sets of proteins, while the latter one calculated the correlation of hyperbolic and semantic distances in the PPIs network.

Comparison of Functionally Relevant Protein Sets with Random Sets

This part of the analysis is targeted to evaluate how the hyperbolic distances of proteins, annotated with the same semantic term, differ from the distances of randomly selected proteins. The developed workflow (Fig. 5-1) was implemented for both GO-BP and GO-MF annotations. The size of GO annotation has its maximum value (the superset of all annotated proteins) in the most generic term (the root of the graph) and then gradually decreases, as the ontological graph is traversed to the leaves, where very specific terms are located with modest or even empty annotation. In order to avoid size values with only a few terms and to reduce the amount of the examined groups, size values were grouped using an increment step, until a specific amount of proteins. Specifically, 20 and 5 were used as increment step values and 600 and 200 were used as maximum protein set sizes for GO-BP and GO-MF respectively. Also, terms with less than 5 annotated proteins were filtered out. Thus, 30 groups of terms were defined for GO-BP and 20 for GO-MF.

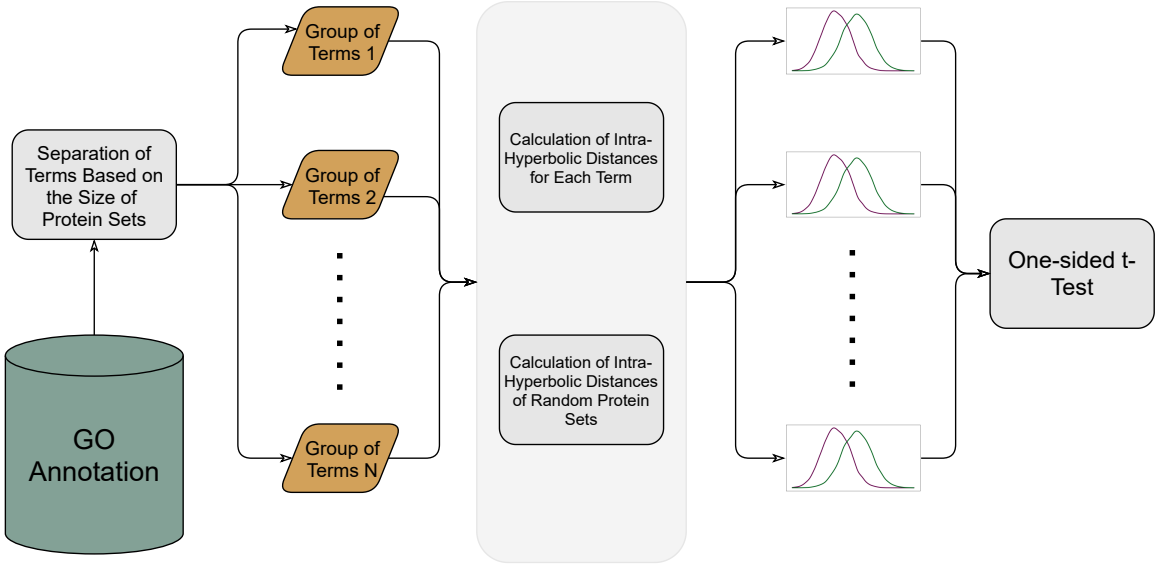

Figure 5-1: Workflow for the comparison of functionally relevant protein sets with random sets. GO terms (GO-BP and GO-MF) were separated into distinct groups according to the size of their protein sets. Then, two distributions of average hyperbolic distance were created, one for the protein sets of group terms and another one for randomly selected protein sets of equal size. The two distributions were compared using one-sided t-Test.

Afterwards, for each term two mean hyperbolic distances were calculated, one taking into account its annotated protein set, and another one using an equally-sized random set of proteins. Following this procedure for each group of terms, two distributions of mean hyperbolic distances were constructed, one which corresponded to the respective protein sets (DistP) and another one derived from randomly defined sets (DistR). A t-test for means of two independent samples was executed for each pair of distributions, assuming as null hypothesis that the expected average values of these distributions were identical and as the alternative one that the expected average of DistP is less than that of DistR (one-sided test). Alternatively stated, the mean hyperbolic distances in-between the annotated protein sets are expected to be smaller (so their corresponding embedded nodes closer) than the ones of randomly selected sets of proteins.

Correlation of Hyperbolic and Semantic Distances

The following procedure was performed to estimate if there is any correlation between hyperbolic and semantic distances, meaning that the hyperbolic topology could indicate functional relatedness to a specific degree. The hyperbolic distances (the *H.sapiens* PPIs network coordinates used to calculate these distances were created using the Rigel embedding algorithm) follow normal distribution as it is

depicted in Fig. 5-2. To calculate their correlation with the semantic distances, three distinct bands of hyperbolic distances were defined, using specific percentiles. Pairs of proteins with distances lower than the 5^{th}, between the 25^{th} and the 75^{th} and greater than the 95^{th} percentile were selected to create three adequately separated groups of pairs on the basis of hyperbolic distance. This discretization was performed instead of a uniform selection, as this approach would end in the accumulation of many pairs with hyperbolic distance around the distribution mean and without many pairs from distribution tails.

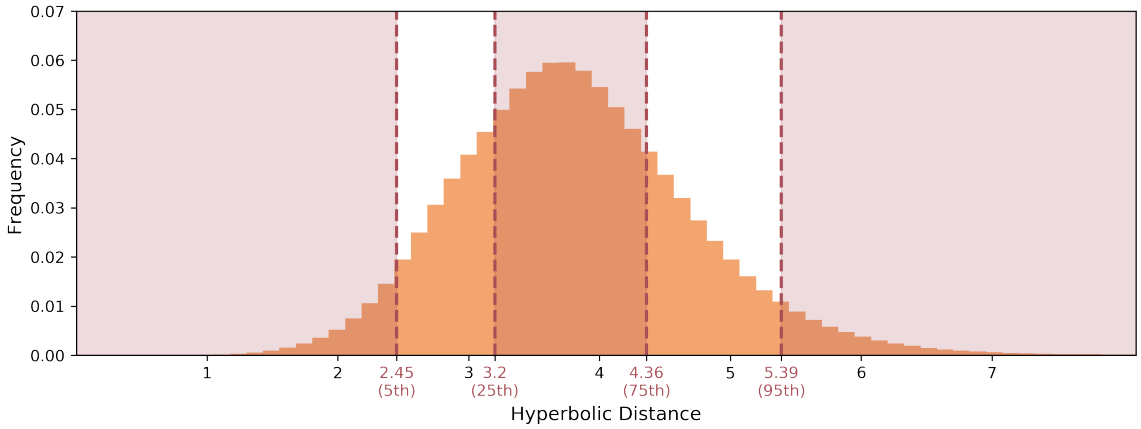

Figure 5-2: Normal distribution of hyperbolic distances.

An iterative process was executed to calculate the aforesaid correlation, using GO-BP and GO-MF annotations (Fig. 5-3). One thousand pairs of proteins were randomly selected from each band of distances. The respective semantic distances were calculated using the average value of Resnik [78], AggregateIC [79] and XGraSM [80] measures to quantify the similarity of semantic terms and subsequently the formula of Best Match Average (BMA, [81]) to aggregate them for the similarity of protein pairs. Then, Pearson's correlation of hyperbolic and semantic distances was calculated. This process was executed one hundred times for each ontology, ending up to a final average correlation score for each one.

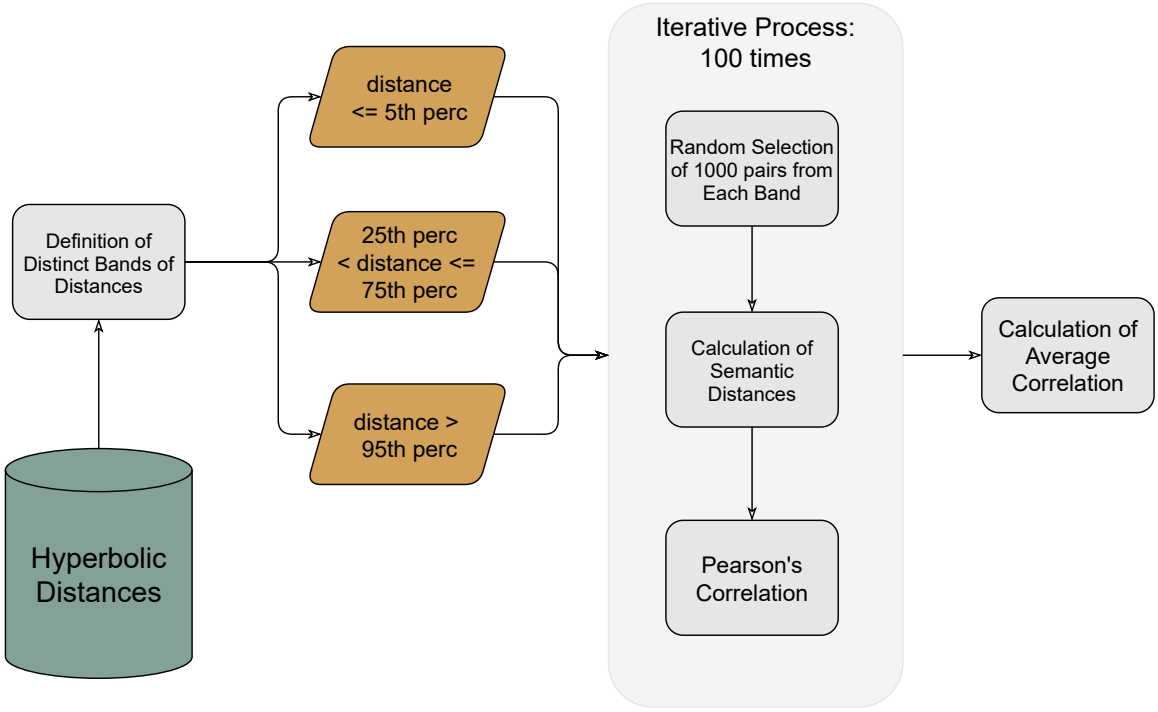

Figure 5-3: Workflow to estimate the correlation of hyperbolic and semantic distances. Three distinct bands of hyperbolic distances were selected in order to contain protein pairs with extreme high, extreme low and moderate distances. An iterative process was performed in order to calculate the Pearson's correlation.

Chapter 6

Results and Discussion

In this chapter, comparative results for the runtime and the evaluation of the accuracy and the performance of the different embedding algorithms in each network case are presented. Finally, an alternative semantic similarity criterion was applied for the biological network, in order to interpret meaningfully and evaluate the embedding based on the predictive capacity of links compared to the functional association of the corresponding proteins. The presented experiments were performed on a personal computer with the following specifications: Intel Core i5-7200U CPU @ 2.50GHz 2.70 GHz processor, 6.00 GB installed RAM and Windows 10 (64-bit) operating system.

6.1 Embedding

6.1.1 Rigel - Impact of Dimensions and Landmarks

In order to evaluate Rigel's accuracy of the estimated distances between nodes, the Average Relative Error (ARE) metric was used, ending up with a set of ideal values for each parameter concerning the magnitude of space dimensionality and landmarks. Indicatively in Fig. 6-1, for each PSO-generated network, the ARE values are plotted as the number of dimensions varies between 2 and 14, and the number of landmarks opts between 30 and 50. As expected the dimensionality of the geometric space, and secondarily the cardinality of the landmark set, plays a crucial role in the precise estimation of distances between nodes. For the most part, regardless the parameters used during the network generation, accuracy does improve significantly as the number of dimensions

increases. Additionally, in most cases, the accuracy slightly decreases when dimensions are greater than 10. The 10-dimensional Hyperbolic Space stands out as the most appropriate choice balancing between precision and the inevitable time- and computational-complexity. Lastly, for the presented networks, it was shown that a relatively smaller set of landmarks gave the same or even smaller ARE values, although the presence of more landmarks throughout the graph embedding space was supposed to allow for higher accuracy.

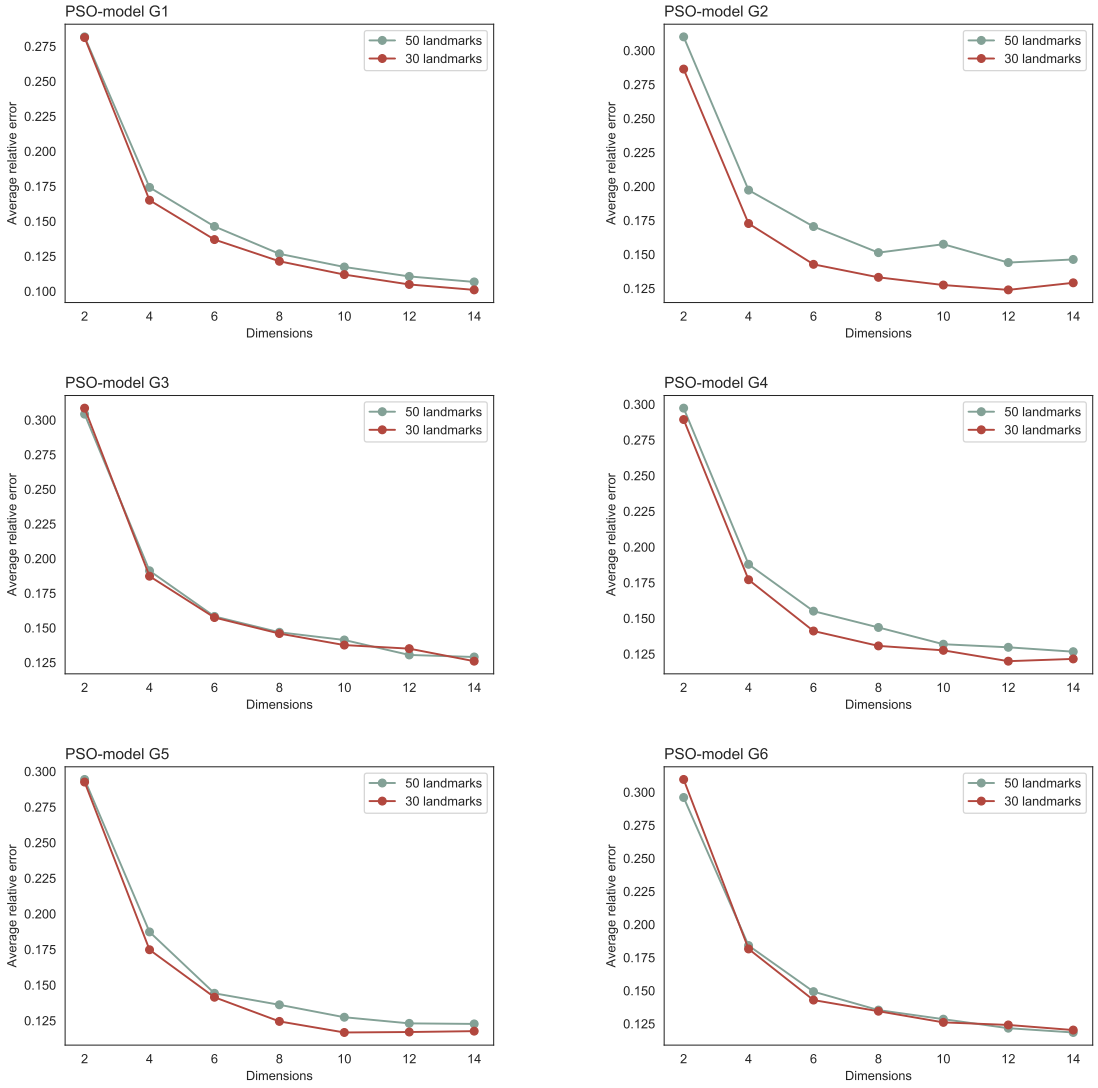

Figure 6-1: Average relative errors of different coordinate dimensions and landmark sets for each PSO-generated network.

The same observations and conclusions from the PSO-generated networks apply to the real complex networks as well (Fig. 6-2), with the combination of 30 landmarks and 10-dimensional coordinates being the choice standing out as the best trade-off between complexity and accuracy.

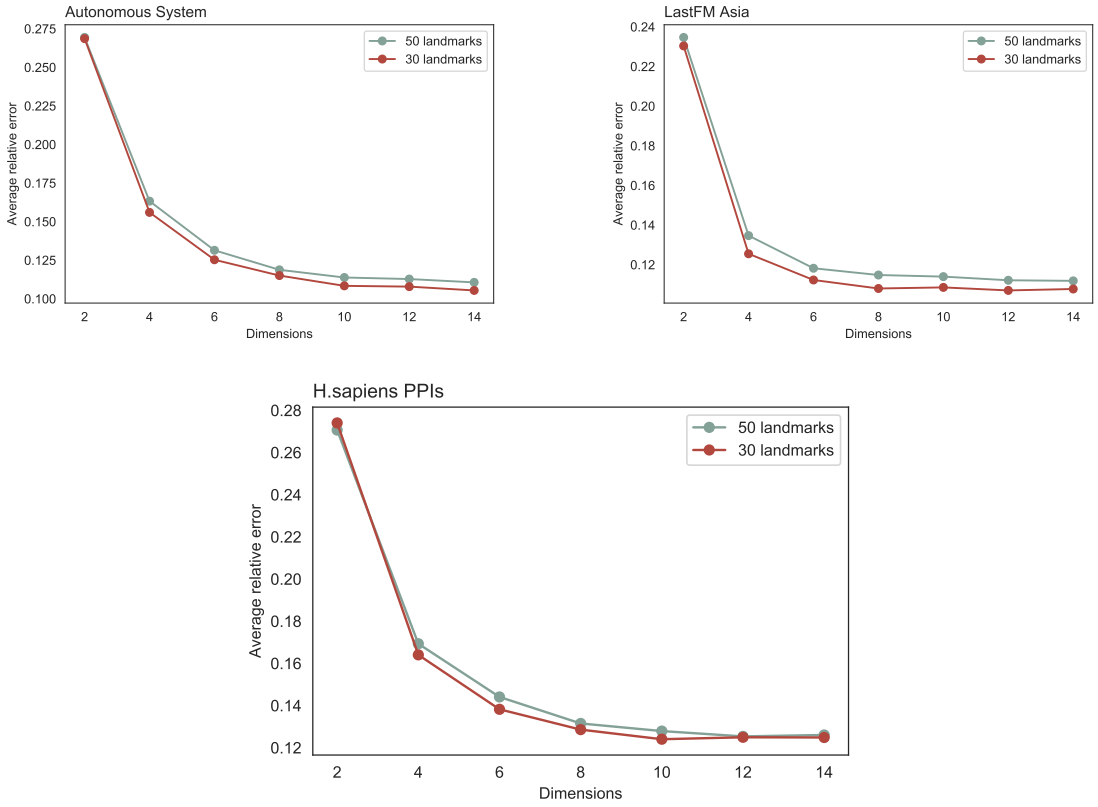

Figure 6-2: Average relative errors of different coordinate dimensions and landmark sets for each real complex network.

6.1.2 Execution Time

Apart from the performance evaluation that will be described below, another examined aspect of the embedding process was the runtime required for each algorithm to map a network into the Hyperbolic Space. Firstly, the great superiority of LaBNE in terms of its time performance was pointed out, based on the high-performance subroutines used to solve large scale eigenvalue problems. In every tested graph the mapping was completed in the order of seconds or even milliseconds.

In case of the MLE approximation algorithm (HyperMap) the time needed to perform the embedding task was 2100 up to 167700 times slower than LaBNE, for the smallest and the biggest networks respectively, although using the speed-up heuristic, which only applies in a subset of nodes with degrees smaller than a fixed value (the default value 10 was used). So despite HyperMap's extreme embedding accuracy, its application is mostly feasible in datasets with a few hundreds of nodes, making impractical its application to big biological networks. Specifically, HyperMap didn't manage to complete the embedding in 48 hours, and for that reason it was excluded from the table 6.1, setting the need to examine different concepts to find an approach that avoids its prohibitive computational time.

In the suggested angular optimization technique, the computational cost is proportional to the size of the graph, the number of candidate new angular positions and the number of iterative rounds, quadratically to the former and linearly to the latter ones. Likewise, the recorded running time ranged between 30 seconds and 2 and a half hours, depending on the number of nodes. Keeping the number of candidates and rounds at low values, the running time decreases and the logarithmic loss improves. In such a condition, the computational cost could be assumed as dependent only on the graph size.

Lastly, in case of Rigel embedding, the most time consuming part is the initial computation of the breadth-first-search (BFS) trees, used to fix the positions of the landmarks' subset. Once these are set, the remaining node positions are computed using Simplex method, in time almost linear to the network size. This approach allows the embedding of even massive networks in a few hours, which means that it needs only a few seconds for the networks used in the current study.

Table 6.1: Time (in seconds) needed by each algorithm to embed the networks (both the PSO-generated and the real complex ones) to Hyperbolic Space. In the case of the *H.sapiens* PPIs network, HyperMap algorithm failed to complete in a reasonable amount of time and so excluded.

Dataset	LaBNE	LaBNE+ang.opt.	HyperMap	Rigel
G1	0.01	32.1	40.28	23.77
G2	0.01	29.6	21.74	21.67
G3	0.02	114	241.48	24.12
G4	0.03	133	95.94	18.63
G5	0.04	776	733.87	46.82
G6	0.07	2942	2932.46	75.61
Autonomous System	0.07	4407	11740.68	107.68
LastFM Asia	0.21	6177	15121.19	147.33
H.sapiens PPIs	1.17	8890	-	132.66

6.2 Greedy Routing

As mentioned in section 5.7.1, one of the adopted methods for the performance evaluation of each embedding algorithm, was the Greedy Routing, and more specifically the percentage of successfully delivered packets along with the average hop stretch of the successful paths. A set of randomly selected source-target pairs was selected to approximate the GR measurements, as for the large-scale networks, where total possible number of links surpasses the $\frac{N(N-1)}{2} \geq 10^4$ threshold, it is impossible to include every pair of source-target nodes in the computations. For each PSO-generated network, and every tested algorithm, 500 source-target pairs were selected at random. The same process was repeated 10 times and the average values are the ones presented below. Noting that since network embedding was completed using two different representation models of the Hyperbolic Space, Poincaré disk and the Hyperboloid, the distance computations were also performed with two different metrics during greedy routing evaluation.

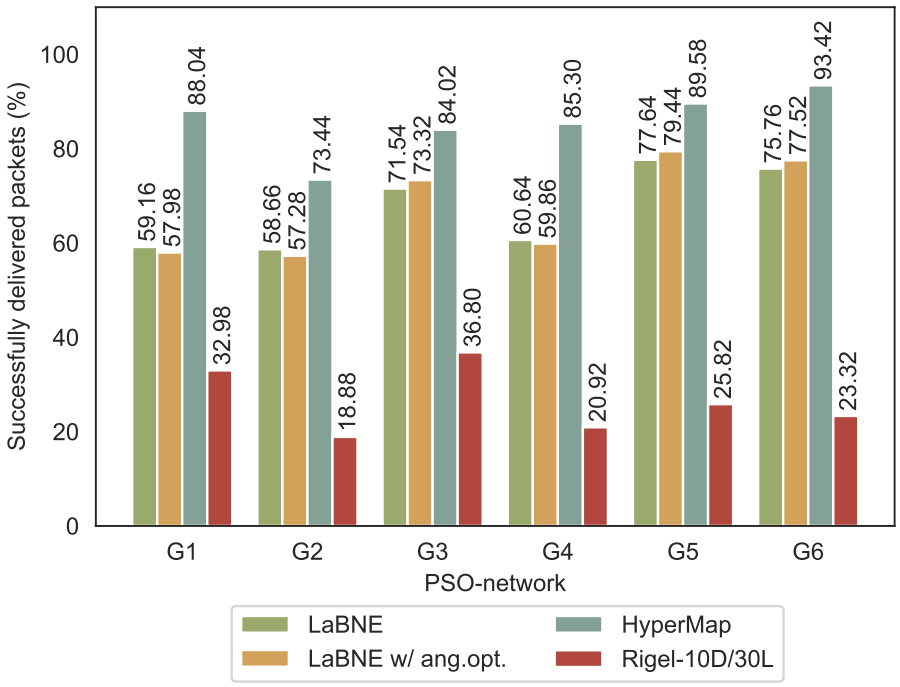

Figure 6-3: Greedy routing efficiency when the inferred hyperbolic coordinates are used for navigation purposes. Indicated by bars is the percentage of successfully delivered packets for LaBNE (olive), LaBNE with Angular Optimization (nectar), HyperMap (cyan) and Rigel with 10dimensions and 30landmarks (carmine) for the studied PSO-generated networks.

As illustrated in Fig. 6-3, HyperMap inferred hyperbolic coordinates allow for efficient navigability, above 80%, in almost all the tested networks. Using coordinates inferred by LaBNE as well as the ones derived from the angular optimization process, the proportion of paths that do not get looped and reach their destinations drops around $60 - 75\%$. Since the optimization step is based on the minimization of logarithmic loss, it does not seem to offer any significant improvement in terms of navigability. Lastly, in case of Rigel's coordinate system, the percentages of successfully delivered packets didn't manage to reach values greater than 37%, making it the most unsuitable option when the navigation in a greedy manner throughout the network space is the main objective of the embedding process. Since Rigel is not a greedy embedding algorithm guaranteeing the greedy routing success in the Hyperbolic Space, the lack of greedy paths in the topology is a quite expected behavior.

Table 6.2: Average geodesic path length l_G and average hop length $\bar{h}$ of the successful paths using the inferred hyperbolic coordinates of each examined algorithm, for the studied PSO-generated networks.

Dataset	l_G	$\bar{h}_{\textbf{LaBNE}}$	$\bar{h}_{\textbf{LaBNE+ang.opt.}}$	$\bar{h}_{\textbf{HyperMap}}$	$\bar{h}_{\textbf{Rigel}}$
G1	3.14	2.99	2.98	3.30	2.64
G2	3.54	3.47	3.44	3.84	2.80
G3	2.74	2.81	2.81	2.93	2.47
G4	3.25	3.37	3.33	3.64	2.85
G5	2.89	2.97	2.97	3.20	2.62
G6	2.90	3.04	3.04	3.23	2.61

Table 6.3: Greedy routing score, or average hop stretch of successfully delivered packets for the considered source-target pairs of each PSO-generated networks, for each examined algorithm.

Dataset	$GR_{\textbf{LaBNE}}$	$GR_{\textbf{LaBNE+ang.opt.}}$	$GR_{\textbf{HyperMap}}$	$GR_{\textbf{Rigel}}$
G1	1.035	1.036	1.067	1.066
G2	1.056	1.056	1.122	1.074
G3	1.072	1.071	1.087	1.084
G4	1.085	1.079	1.142	1.107
G5	1.058	1.056	1.114	1.109
G6	1.074	1.070	1.116	1.114

The results for the length of the utilized paths that lead to successfully delivered packets (average hop length $\bar{h}$) are presented in Table 6.2. Concerning to Rigel, the low hop length values pinpoint the difficulty of its coordinates to rightly navigate a package using paths with length greater than 2 hops. In contrast, HyperMap's coordinates successfully generate greedy paths with lengths close to the geodesic ones, or a bit longer than them. Finally, LaBNE with or without the Angular Optimization, shows moderate ability to generate greedy paths. When dividing the length of the greedy paths by the length of the corresponding shortest paths between source and target nodes in the graph (hop

stretch), the resulting average values are close to 1 (Table 6.3). This result indicates that the vast majority of greedy paths are also shortest paths and so optimal for all algorithms.

In terms of PSO-model parameters, routing efficiency tends to be higher when the networks appear more heterogeneous and strongly clustered, and is reduced when they lose their connectedness or local link density. More specifically the network G6 with the highest clustering coefficient (0.60) is also the one with the highest routing efficiency (93.42%), while the G2 network with a clustering coefficient among the lowest (0.48), has the lowest efficiency (73.44%). The reported values refer to HyperMap, but the same trend is observed to LaBNE with or without optimization. Network density is related mainly to temperature T, while degree heterogeneity is controlled by the scaling parameter γ, with the low T and γ values producing more navigable networks. In a network with weak clustering, path redundancy reduces, the geometrical/hierarchical structure decays and the dependence between edge probability and hyperbolic distance decreases gradually.

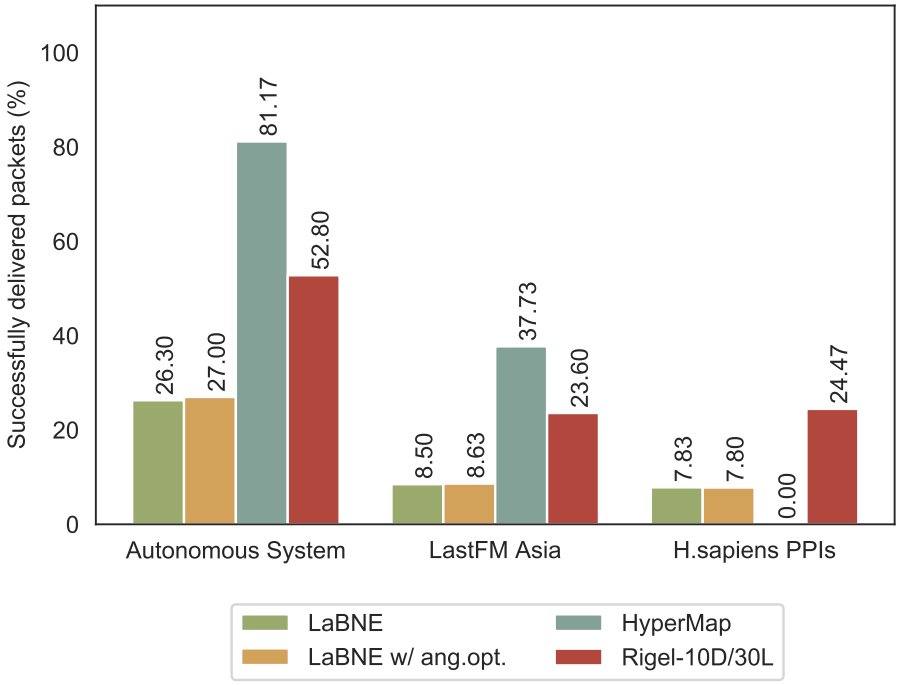

Figure 6-4: Greedy routing efficiency when the inferred hyperbolic coordinates are used for navigation purposes. Indicated by bars is the percentage of successfully delivered packets for LaBNE (olive), LaBNE with Angular Optimization (nectar), HyperMap (cyan) and Rigel with 10 dimensions and 30 landmarks (carmine) for the studied real complex networks. In the case of the *H.sapiens* PPIs network, HyperMap results was not included.

Table 6.4: Average geodesic path length l_G and average hop length $\bar{h}$ of the successful paths using the inferred hyperbolic coordinates of each examined algorithm, for the studied real complex networks.

Dataset	l_G	$\bar{h}_{\textbf{LaBNE}}$	$\bar{h}_{\textbf{LaBNE+ang.opt.}}$	$\bar{h}_{\textbf{HyperMap}}$	$\bar{h}_{\textbf{Rigel}}$
Autonomous System	3.71	3.73	3.74	4.07	3.65
LastFM Asia	5.23	4.74	4.79	5.96	5.20
H.sapiens PPIs	3.96	3.81	3.76	-	4.05

Table 6.5: Greedy routing score, or average hop stretch of successfully delivered packets for the considered source-target pairs of each real complex networks, for each examined algorithm.

Dataset	$GR_{\textbf{LaBNE}}$	$GR_{\textbf{LaBNE+ang.opt.}}$	$GR_{\textbf{HyperMap}}$	$GR_{\textbf{Rigel}}$
Autonomous System	1.097	1.099	1.125	1.112
LastFM Asia	1.151	1.155	1.272	1.140
H.sapiens PPIs	1.208	1.204	-	1.181

When it comes to quality evaluation of the hyperbolic embedding of a complex network, greedy routing score is probably the most widely used metric and therefore a fair measure of comparison between embedding algorithms based on PSO model and not only (since it makes no assumption about the model that generated the examined network). Apart from the comparison of the absolute values of greedy routing efficiency between different algorithms, another way to evaluate the result is to compare it with the one calculated using the polar coordinates originally assigned during the network generation. Indeed, the intrinsic routing efficiency has a success rate between $92.26 - 98.36\%$. However, in the case of real networks' embedding, that intrinsic GR-score cannot be determined computationally. Because of this, one cannot decide if an obtained GR-score is close or far from that network's "ground truth", though he could still compare embedding algorithms, by ranking their achieved GR-scores in a decreasing manner.

Given that, HyperMap's greedy routing was found the one with the highest efficiency, however its percentage is reduced significantly in the case of LastFM Asia network (Fig. 6-4). Although HyperMap was not used for the embedding

of the *H.sapiens* PPIs network, it would not be groundless to assume that its performance would be the same as in the LastFM Asia network. This assumption is based on the fact that the estimated parameters of the network (quite high temperature (0.74) and γ (2.27) values) lead to a very low clustering coefficient equal to 0.17. Theoretically, this clustering coefficient is not ideal to embed using HyperMap or even LaBNE algorithm. Despite those difficulties, hop stretch values appear quite close to 1 (Table 6.5), regardless the network or the algorithm, indicating that even that smaller percentage of successful greedy paths are also the shortest ones in the network.

6.2.1 Greedy Routing in Pathway Databases

In order to examine the effectiveness of greedy navigation in the case of the *H.sapiens* PPIs network, a specific part of that immense network was isolated. Particularly, the signal transduction pathways, listed in KEGG [82] and WikiPathways [83] were used. Thanks to that component of the *H.sapiens* PPIs network, the cell controls its own function, its intercellular communication and its reaction to environmental stimuli [84]. KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY database is a knowledge base providing high-level molecular and functional information generated from genomic data, containing graphical representations of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. As well, WikiPathways is a collaborative open-science molecular pathway database capturing mechanistic knowledge in pathway diagrams, facilitating data visualization and analysis. Pathway information from both these databases were retrieved from gmt files (Gene Matrix Transposed), including lists of datanodes per pathway, unified to Entrez Gene identifiers. Those files are correspondingly available for download at https://www.pathwaycommons.org/archives/PC2/v12/ and https://wikipathways-data.wmcloud.org/. Also, the Entrez Gene IDs were mapped to HGNC Symbols as the ones used in the studied *H.sapiens* PPIs network. Then, the starting- and end- points of each functional pathway were isolated, to use as source-target pairs in the greedy routing evaluation.

For this step of the analysis, the coordinates inferred by Rigel (since are the ones with the more promising results) and LaBNE (as a measure of comparison) were used. Routing efficiency was found 57.7% (in contrary to 32.4% with LaBNE) while the mean hop stretch (length of the computed greedy paths between two

nodes divided by the length of their shortest paths in the graph) was 1.18, a value close to 1. This result indicates that the navigated greedy paths, guided by network's geometry while using local information only, were the shortest paths as well.

6.3 Logarithmic Loss

Logarithmic loss (LL) is the negative log-likelihood, so for algorithms that operate in order to maximize the likelihood, the logarithmic loss is a common quality metric. The smaller the logarithmic loss the better the embedding results. In order to quantify this, the network's LL was computed using the inferred coordinates for each embedding algorithm. These results were compared against the LL obtained using the original / "ground truth" polar coordinates that came up during the PSO-network generation process (henceforth referred as intrinsic LL value of the network or $LL_{original}$). The closer the $LL_{inferred}$ compared to the $LL_{original}$ the better the embedding quality. In the following Table 6.6, these differences between the LL values are presented while in Fig. 6-5 the log-scaled values of the computed LL values are plotted.

As shown below, HyperMap succeeded to keep the network's LL as close as possible to the intrinsic value, after the mapping process. This result verifies why HyperMap is probably the most widely used algorithm, that exploits the logarithmic loss minimization problem. On the other hand, it is no wonder that the LL associated with LaBNE inferred coordinates is the one with the greatest divergence from the intrinsic LL. When LaBNE was followed by the angular optimisation step, it managed to approximate the intrinsic network's LL, providing an about 18% lower LL compared to LaBNE without optimisation, but about a 20% higher value compared to HyperMap. That reduction of LL is anticipated, since the optimization is actually performed with regard to that. Noting that, since Rigel is based on the Hyperboloid model, the logarithmic loss with respect to the PSO model cannot be considered as a proper quality metric for this embedding method; therefore, Rigel was excluded from this part.

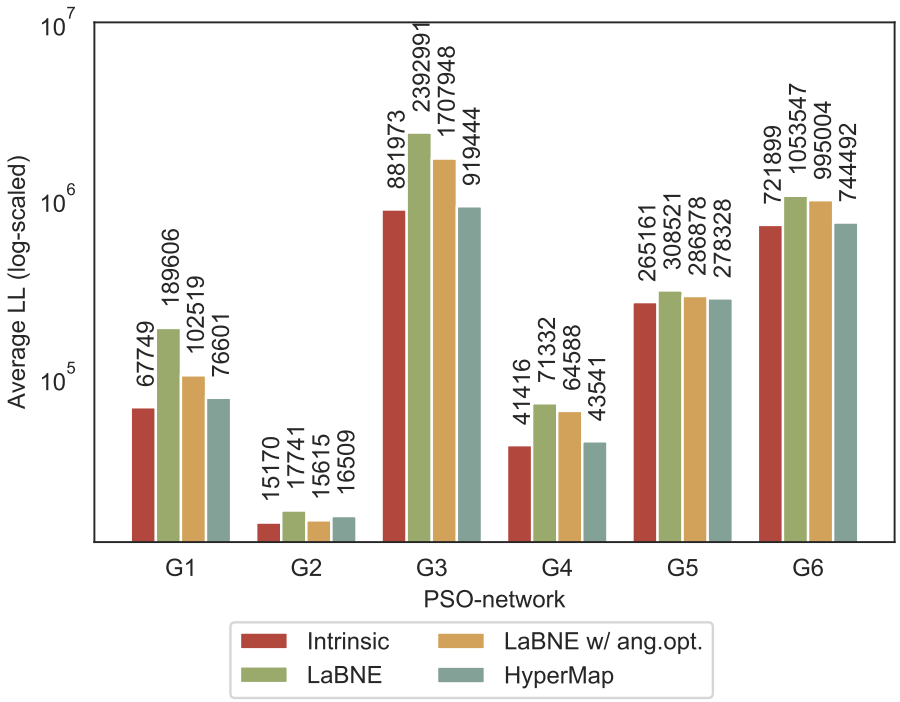

Figure 6-5: Logarithmic Loss when the original and the inferred hyperbolic coordinates are used for embedding purposes. Indicated by bars is the log-scaled LL score for each embedded PSO-generated network using LaBNE (olive), LaBNE with Angular Optimization (nectar) or HyperMap (cyan) coordinates. The intrinsic LL value (carmine) is also included. The recorded LL values (without the log-transformation) are annotated on the top of each bar.

Table 6.6: Differences between LL values based on the inferred coordinates of each algorithm and the original ones of each PSO-generated networks. The smaller the divergence of the $LL_{inferred}$ from the $LL_{original}$, the smaller the value in the table, and the better the embedding quality.

Dataset	$LL_{\mathbf{intr}} - LL_{\mathbf{LaBNE}}$	$LL_{\mathbf{intr}} - LL_{\mathbf{LaBNE+ang.opt.}}$	$LL_{\mathbf{intr}} - LL_{\mathbf{HyperMap}}$
G1	121856	34770	8852
G2	2571	445	1339
G3	1511019	825975	37471
G4	29916	23172	2125
G5	43361	21717	13167
G6	331648	273105	22593

As seen below, the difference in network's LL between the algorithms' efficiency was amplified when computing this metric in the real complex networks. The HyperMap's score was 3 to 5 times lower than the LaBNE's score. In this case, the angular optimization did contribute with a 7% reduction on average, but definitely not enough to reach HyperMap's values.

An important assumption made by the tested optimization framework is the fact that the previously inferred coordinates given as an input, are already quite satisfying; thus, only minor adjustments are needed. So the magnitude of the improvement is controlled and restricted by the accuracy and efficiency of the previous algorithm. A brief way to confirm that property is to check the impact of angular optimization onto the results produced by HyperMap algorithm. Indeed, when that framework (with same configuration) was applied to the Autonomous System coordinates inferred by HyperMap, it managed to reduce network's LL by 30%, namely from 189809 to 135013. This actually highlights the important role of the quality of the input draft embedding.

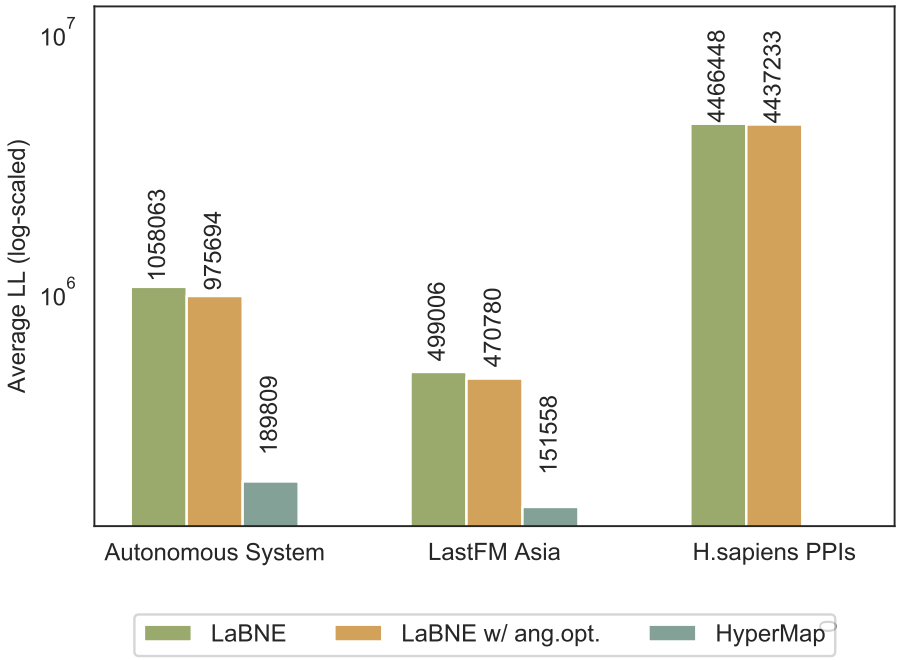

Figure 6-6: Logarithmic Loss using the inferred hyperbolic coordinates to embed. Indicated by bars is the log-scaled LL score for each embedded real complex network using LaBNE (olive), LaBNE with Angular Optimization (nectar) or HyperMap (cyan) coordinates. The recorded LL values (without the log-transformation) are annotated on the top of each bar. In the case of the *H.sapiens* PPIs network, HyperMap results were excluded.

6.4 Link Prediction

As mentioned in section 5.7, apart from the Greedy Routing, an alternative approach of the embedding evaluation is the predictive potential of an algorithm by quantifying the existence likelihood of an edge between two nodes based on their hyperbolic distance. The smaller the distance, the more likely a link to exist. For this reason, firstly, a random percentage of links was removed from the network. Then, the remaining nodes were used as an input to the described embedding process (including the parameters' determination step). The choice of links to delete is restrained by two factors: 1. the fact that the deletion of a large number of edges can cause distortion, loss of connectedness and node isolation, and 2. the existence of bridges and/or links with high edge betweenness centrality that are important for the integrity of the network structure, so they shouldn't be removed. Those two restrictions were taken into consideration during the link deletion. And so, the amount of destroyed links was set to 10%. However in the case of Autonomous System and LastFM Asia networks, the structure broke into more than one components. For that reason and since the used network is that of *H.sapiens* PPIs, their corresponding results are not presented.

As seen in Fig. 6-7, the performance of both LaBNE (with or without angular optimization), HyperMap and Rigel improves as clustering increases. LaBNE's results in most cases were similar or slightly inferior to HyperMap's, a quite promising finding especially for the case of massive graphs, if we take into consideration LaBNE's extremely fast embedding process. Angular optimization offered mostly no added value to LaBNE's results.

An actual aftermath of these pretty low clustering values of the networks was the massive number of potential links. That, in combination with the fact that the differences between the computed hyperbolic distances were often located after the third decimal, created many ties whose actual order may cause non-negligible variance in the precision values, that appeared very low in every examined case.

In the *H.sapiens* PPIs network, Rigel performs better than LaBNE, allowing for higher Recall without sacrificing that much of Precision. Further on, the predictive potential of the Rigel algorithm over this protein network will be interpreted in the light of semantic similarity, aiming to detect an association between the inferred distances in the Hyperbolic Space and the functional divergence of proteins on the PPIs network.

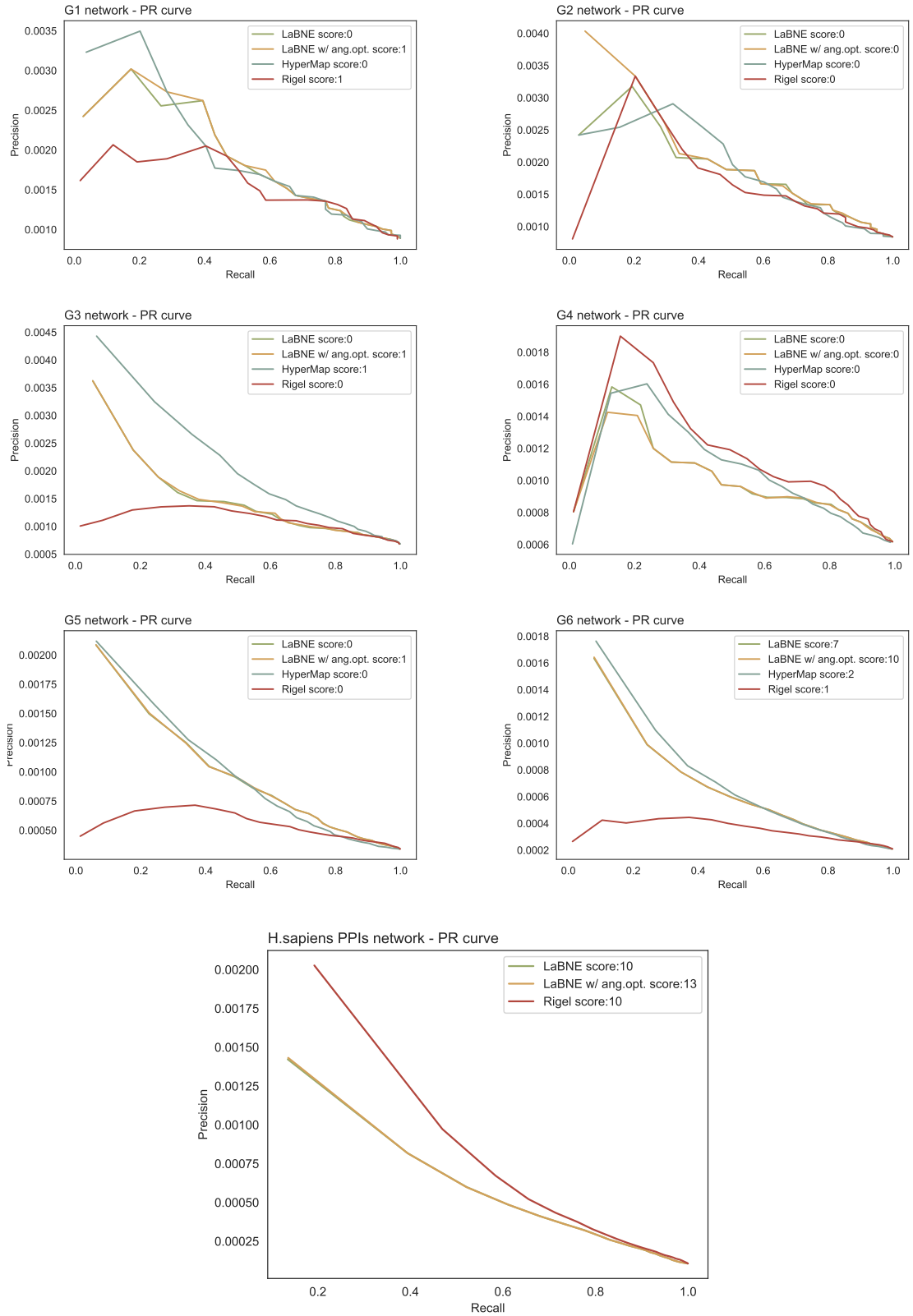

Figure 6-7: Precision-Recall curves to assess link-prediction performance of each embedding algorithm for each PSO-generated network as well the *H.sapiens* PPIs network. The score mentioned in the legends refers to the link-prediction score described in section 5.7.2.

6.5 Semantic Similarity

6.5.1 Comparison of Functionally Relevant Protein Sets with Random Sets

As mentioned in section 5.7.3, 30 and 20 groups of ontological terms were generated for GO-BP and GO-MF respectively. The boxplots of mean hyperbolic intra-distances for the term groups (distP) as well the corresponding boxplots for the random sets (distR) are depicted in Fig. 6-8A, Fig. 6-9A and Fig. 6-8B, Fig. 6-9B, respectively.

In both cases, the distP distributions have a greater deviation from the distR ones, whose variance also decreases as the size of random protein sets increases. This is an expected behavior due to law of Large Numbers, because as the set of randomly selected proteins increases in size, their average intra-distance converges to the average value of the whole population and its variance decreases. Apart from that, the crucial question was to find out if proteins participating in a common biological mechanism have lower hyperbolic distance than randomly selected proteins. Thus, the subsequent statistical tests for each group of terms, assumed as alternative hypothesis that the mean value of distP distribution is lower than the corresponding distR one.

The hypothesis testing was performed adopting p-value=0.01 as the threshold for the null hypothesis rejection. The analysis revealed that for both GO domains, the null hypothesis is rejected for a few groups of terms, which have relatively small size of annotated proteins (Fig. 6-8C and Fig. 6-9C). Specifically, for GO-MF only the first two group of terms (with size below 20) have significantly smaller hyperbolic intra-distance from the equally sized random sets. The null hypothesis is accepted for greater in size protein sets, indicating that the distP converges to the distR. In the case of GO-BP, the size limit below which the null hypothesis is always rejected cannot be posed with certainty, however it is evident that as the size of annotated protein sets increases the probability to reject the null hypothesis by chance increases as well. Also, it is seen that the decision of null hypothesis is not affected by the number of examined proteins sets. To conclude, although the analysis cannot draw conclusions linking the hyperbolic distance to the functional association of a pair of proteins, the above results pinpoint that the overall functional relatedness of proteins in a small set, could be also captured by the average hyperbolic intra-distance in that set.

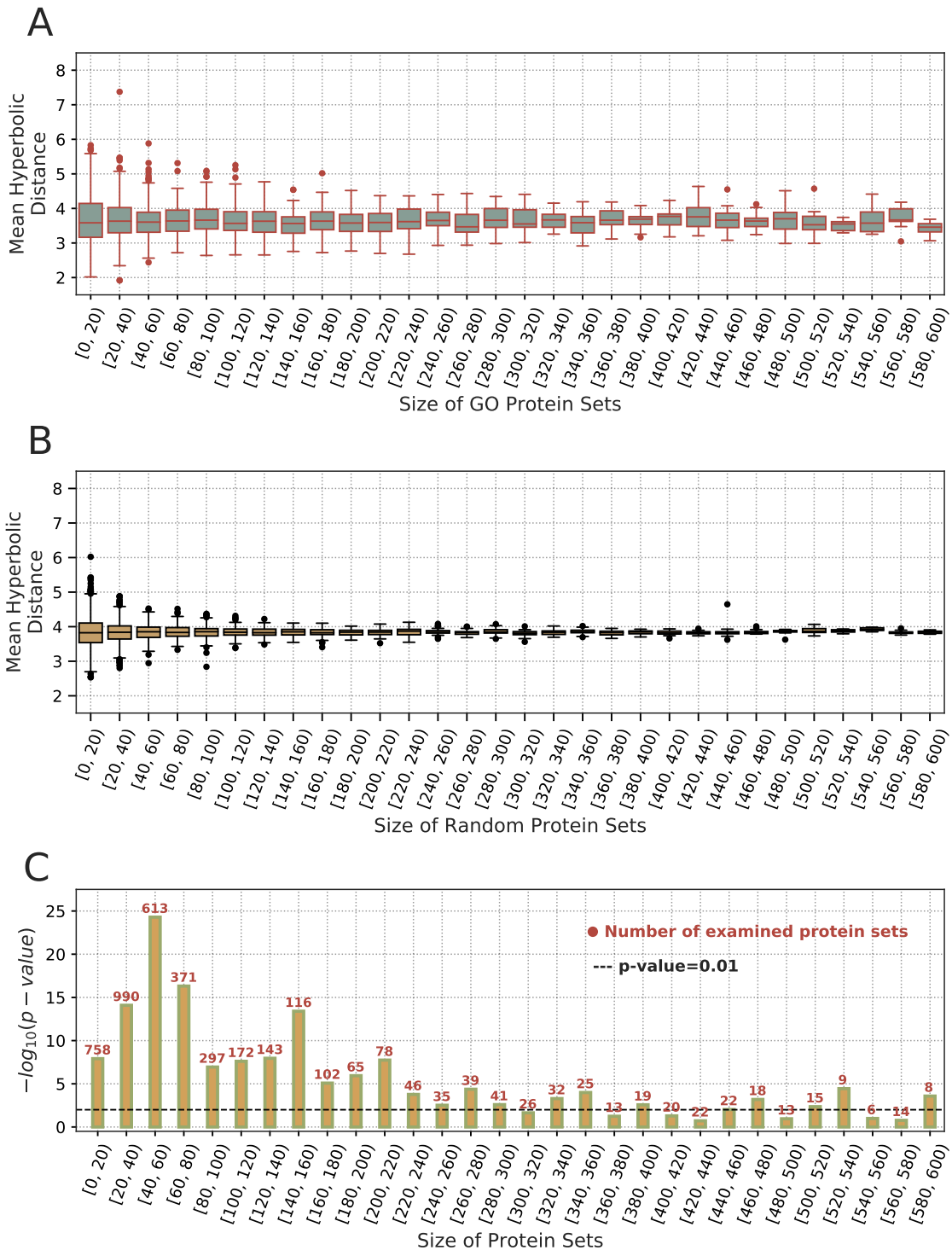

Figure 6-8: **A.** Boxplots of hyperbolic intra-distances for each group of terms, for GO Biological Process (GO-BP) domain. Each box shows the median, minimum, maximum, 1^{st} and 3^{rd} quartile of the mean hyperbolic distances, while the whiskers show the rest of the distribution, and the circles mark the outliers. **B.** Boxplots of hyperbolic intra-distances for each corresponding equally sized random set of proteins, for GO-BP domain. The consequence of the law of Large Numbers is visible, as the larger the sample the closer the observed sample average gets to the population average. **C.** One-sided T-test results for GO-BP domain, to determine whether there is a significant difference between the expected average values for each pair of distP and distR distributions.

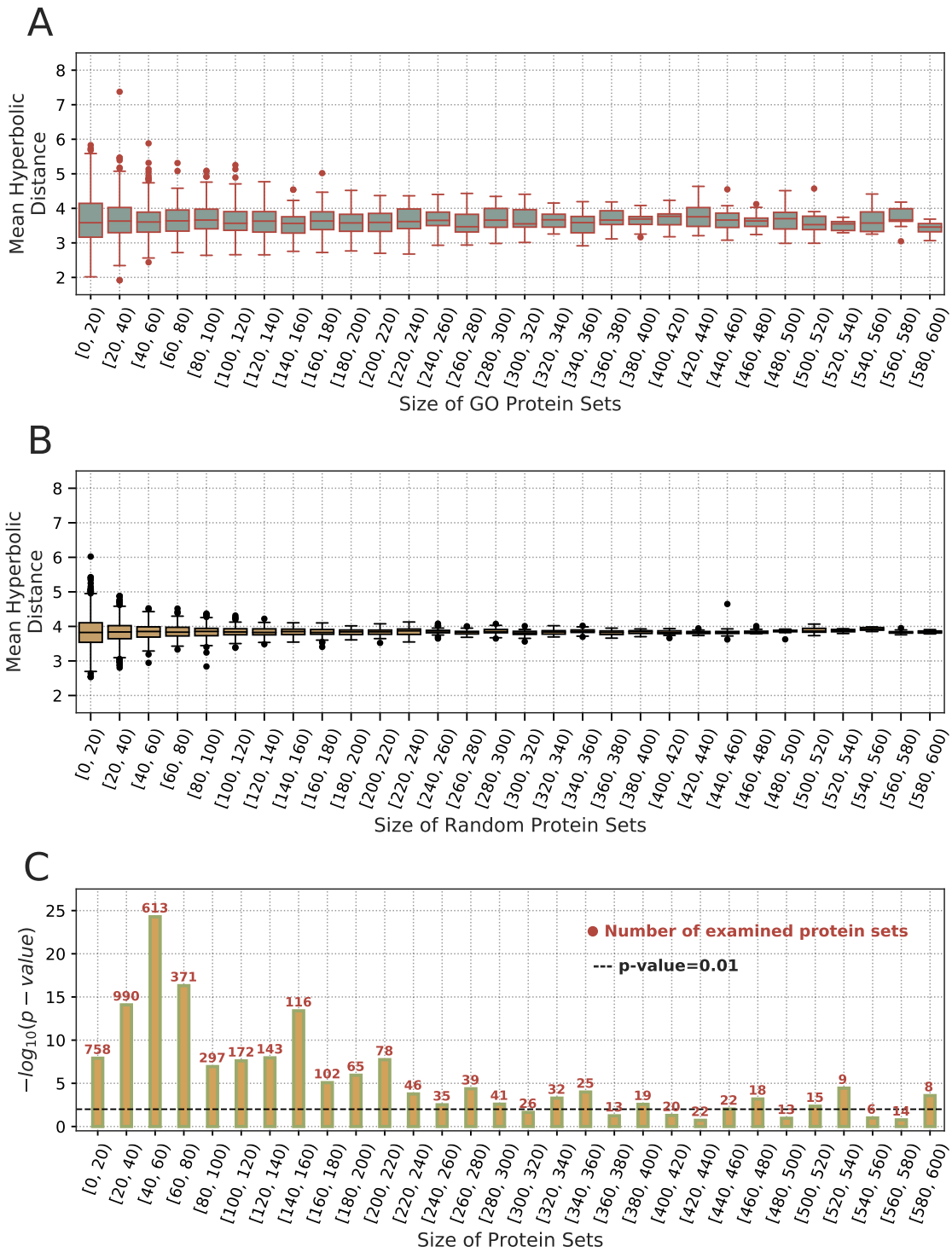

Figure 6-9: **A.** Boxplots of hyperbolic intra-distances for each group of terms, for GO Molecular Function (GO-MF) domain. Each box shows the median, minimum, maximum, 1^{st} and 3^{rd} quartile of the mean hyperbolic distances, while the whiskers show the rest of the distribution, and the circles mark the outliers. **B.** Boxplots of hyperbolic intra-distances for each corresponding equally sized random set of proteins, for GO-MF domain. The consequence of the law of Large Numbers is visible, as the larger the sample the closer the observed sample average gets to the population average. **C.** One-sided T-test results for GO-MF domain, to determine whether there is a significant difference between the expected average values for each pair of distP and distR distributions.

6.5.2 Correlation of Hyperbolic and Semantic Distances

The conducted Monte-Carlo process (100 times uniform selection of 3000 protein pairs from three different groups) revealed that the correlation of hyperbolic and semantic distances based on GO-MF and GO-BP is 0.247 and 0.295 respectively. This slightly weak positive correlation indicated that hyperbolic distance cannot unmistakably determine the functional similarity between a pair of proteins. This result enhances the previous evidences of hypothesis testing. Nonetheless, the positive direction of correlation implies that in some cases the proximity in Hyperbolic Space mentions functional relatedness. The pattern that emerges is that topological co-localization does not indicate strong functional association and as well the functional relevance is not a necessary and sufficient condition for vicinity in the Hyperbolic Space of the embedded network.

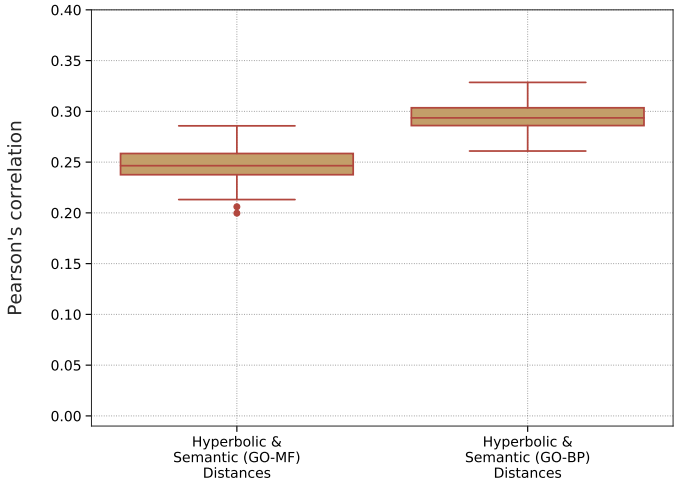

Figure 6-10: Boxplot of Pearson correlation coefficients between hyperbolic and semantic distances both for GO-MF and GO-BP.

Chapter 7

Conclusion

7.1 Summary of Results

In the context of this thesis, the embedding of complex networks into the Hyperbolic Space using different algorithms was studied. A comparative analysis regarding the execution time and the evaluation of the accuracy and performance in each network case was conducted. For the special case of human interactome, in order to give a more biologically significant interpretation to the results, the inferred hyperbolic distances of unlinked proteins were compared to their corresponding functional divergence by using biomedical ontologies. Concisely, some of the more important remarks are the following:

- LaBNE was extremely faster than any other tested algorithm, needing only a few milliseconds to embed thousands of nodes. Rigel algorithm is following close, with runtime of seconds for all the tested networks. Concerning the proposed Angular Optimization method, it scaled quadratically to the number of nodes, resulting to execution time between seconds and a few hours. Lastly, HyperMap is without doubt the most time-consuming algorithm (lasts from seconds to couple of hours), to the extent that did not complete the embedding of human interactome in less than 2 days.

- Networks that are more heterogeneous and with stronger clustering, are more navigable. In other words, greedy routing is expected to be more efficient in networks with small scaling exponent γ as well low temperature T. In cases of weak clustering, the path redundancy reduces, the hierarchical structure decays and the dependence between edge probability and hyperbolic distance decreases gradually. Independently of the examined

network, HyperMap inferred coordinates had proven to offer the more efficient navigability, although its efficiency reduces when the clustering weakens. Concerning LaBNE and Rigel, the former one performs better in the case of networks generated by the Popularity-Similarity model while the latter one becomes more greedy-routing-efficient in the case of real complex networks. Overall, for all the algorithms, the vast majority of greedy paths also appear to be the shortest ones and so the optimal ones.

- Regarding the human interactome, signal transduction was additionally examined. Proteins can only interact with their adjacent neighbors, so without any knowledge of the global network structure, signals sent from a source manage to cascade through the network and reach their target. Given that, if a signal can effectively traverse the network from source to target, using the greedy path available, was tested. That seemed to be confirmed for almost 60% of examined pathways, with the hop stretch of these paths being close to 1 indicating their compatibility with the shortest paths of the network.

- The aforementioned remarks concerning the greedy routing efficiency are still applicable in the case of link prediction, where the performance of every algorithm applied in any network, is crucially dependent on the network's clustering. Given networks with weak clustering, the performance in terms of precision is bad for all the algorithms, and probably their application should be restricted to highly clustered networks.

- The proposed Angular Optimization framework is based on the assumption that the previously inferred coordinates are satisfying enough and so the adjustment window is limited in an angular range defined by every node's neighbors. Thus, its application would make sense for refinement reasons and under certain conditions.

- Regards to the semantic similarity testing, it seems that for the case of GO terms' groups (both GO-BP and GO-MF) which include a small number of annotated proteins, the functional relevance of these proteins could be resembled by the average value of hyperbolic distances between them. Also, a weak positive correlation between vicinity in the Hyperbolic Space and functional similarity of a protein-protein pair was revealed.

7.2 Insights for Future Research

Concluding this thesis, based on the aforementioned outcomes, a few indicative future research steps along with some interpretive ideas are presented:

- Since biological networks, such as those of protein-protein interactions, display weak clustering, more algorithms appropriate for that key property, should be tested.
- The proposed Angular Optimization framework could be evaluated after being applied to coordinates, inferred by LaBNE (or any other draft geometric configuration), but in a different context where the initial results are more satisfying.
- For the case of greedy routing in the pathway databases, the protein members participating in those identified greedy paths should be cross-examined with the members in the original signal transdustion pathways. Not reported members could be interesting as novel pieces of the signaling cascade.
- Taking into consideration that proteins involved in the same disease are more likely to interact with each other, rather than with random proteins, the previously unlinked protein pairs found in close distance in the Hyperbolic Space could be used as new proteins potentially associated with the same disease. Populating unknown parts of the disease pathways, could ideally offer novel drug development targets, or biomarkers of disease classification and progression.
- A primary problem in constructing networks based on experimental data is the fact that some links might not be directly observable (due to experimental constraints), or others could be falsely detected (mostly due to the nature of interactors). So, two unlinked proteins in the interactome should not be treated as true negatives. Luckily, biological networks share topological properties with other complex networks (social and technological) and so they could be modelled and studied accordingly. That shared structure, allows the development of algorithms that would assist towards the prediction of missing or forthcoming interactions.

Bibliography

[1] R. Solé, R. Ferrer-Cancho, J. Montoya, and S. Valverde, "Selection, tinkering, and emergence in complex networks," *Complexity*, vol. 8, no. 1, 2002. DOI: 10.1002/CPLX.10055.

[2] I. Zelinka, D. Davendra, J. Lampinen, R. Senkerik, and M. Pluhacek, "Evolutionary algorithms dynamics and its hidden complex network structures," in *2014 IEEE Congress on Evolutionary Computation (CEC)*, 2014, pp. 3246–3251. DOI: 10.1109/CEC.2014.6900441.

[3] R. Pastor-Satorras and A. Vespignani, "Epidemic dynamics and endemic states in complex networks," *Phys. Rev. E*, vol. 63, 6 May 2001. DOI: 10.1103/PhysRevE.63.066117.

[4] M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, 1 Jul. 2002. DOI: 10.1103/PhysRevE.66.016128.

[5] M. Porter, J. Onnela, and P. Mucha, "Communities in networks," *Notices of the American Mathematical Society*, vol. 56, no. 9, Feb. 2009.

[6] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, Feb. 2010. DOI: 10.1016/j.physrep.2009.11.002.

[7] A. Clauset, C. Moore, and M. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98–101, Jun. 2008. DOI: 10.1038/nature06830.

[8] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011. DOI: 10.1016/j.physa.2010.11.027.

[9] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá, "Hyperbolic geometry of complex networks," *Physical Review E*, vol. 82, no. 3, Sep. 2010. DOI: 10.1103/PhysRevE.82.036106.

[10] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002. DOI: 10.1126/science.1073374.

[11] P. Holme, M. Huss, and H. Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, Mar. 2003. DOI: 10.1093/bioinformatics/btg033.

[12] D. Meunier, R. Lambiotte, A. Fornito, K. Ersche, and E. Bullmore, "Hierarchical modularity in human brain functional networks," *Frontiers in Neuroinformatics*, vol. 3, p. 37, Oct. 2009. DOI: 10.3389/neuro.11.037.2009.

[13] T. J. Akiki and C. G. Abdallah, "Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks," *Scientific Reports*, vol. 9, no. 1, p. 19 290, Dec. 2019. DOI: 10.1038/s41598-019-55738-y.

[14] T. Enver, S. Soneji, C. Joshi, *et al.*, "Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells," *Human Molecular Genetics*, vol. 14, no. 21, pp. 3129–3140, Sep. 2005. DOI: 10.1093/hmg/ddi345.

[15] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002. DOI: 10.1126/science.1075762.

[16] G. Ghoshal, "Structural and dynamical properties of complex networks.," Jan. 2009.

[17] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. DOI: 10.1126/science.286.5439.509.

[18] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, Oct. 2000. DOI: 10.1038/35036627.

[19] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, Jun. 2002. DOI: 10.1073/pnas.122653799.

[20] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, Association for Computing Machinery, 2007, pp. 29–42. DOI: 10.1145/1298306.1298311.

[21] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, Mar. 2001. DOI: 10.1038/35065725.

[22] P. Erdos and A. Renyi, "On the evolution of random graphs," *Publ. Math. Inst. Hungary. Acad. Sci.*, vol. 5, pp. 17–61, 1960.

[23] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998. DOI: 10.1038/30918.

[24] E. N. Gilbert, "Random graphs," *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1141–1144, Dec. 1959. DOI: 10.1214/aoms/1177706098. [Online]. Available: https://doi.org/10.1214/aoms/1177706098.

[25] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4, pp. 175–308, Feb. 2006. DOI: https://doi.org/10.1016/j.physrep.2005.10.009.

[26] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, May 1967. DOI: https://doi.org/10.1007/978-3-658-21742-6_94.

[27] M. Newman and D. Watts, "Renormalization group analysis of the small-world network model," *Physics Letters A*, vol. 263, no. 4-6, pp. 341–346, Dec. 1999. DOI: https://doi.org/10.1016/S0375-9601(99)00757-4.

[28] Wikimedia Commons, *Small world network example*, 2013. [Online]. Available: https://commons.wikimedia.org/wiki/File:Small-world-network-example.png.

[29] H. A. Simon, "On a class of skew distribution functions," *Biometrika*, vol. 42, no. 3/4, pp. 425–440, Dec. 1955. DOI: https://doi.org/10.2307/2333389.

[30] D. Price, "A general theory of bibliometric and other cumulative advantage processes," *J. Amer. Soc. Inform. Sci.*, vol. 27, pp. 292–306, 1976.

[31] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, Jan. 2003. DOI: 10.1137/s003614450342480.

[32] K. Rosen, *Handbook of Discrete and Combinatorial Mathematics*, ser. Discrete Mathematics and Its Applications. CRC Press, 2017.

[33] A.-L. Barabási, *Linked: The New Science of Networks*. Perseus Books Group, May 2002, vol. 71. DOI: 10.1119/1.1538577.

[34] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The web as a graph: Measurements, models, and methods," in *Computing and Combinatorics*, T. Asano, H. Imai, D. T. Lee, S.-i. Nakano, and T. Tokuyama, Eds., Springer Berlin Heidelberg, 1999, pp. 1–17.

[35] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, "Stochastic models for the web graph," in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 2000, pp. 57–65. DOI: 10.1109/SFCS.2000.892065.

[36] Euclid and T. L. Heath, *The thirteen books of Euclid's Elements*, S. T. L. Heath, Ed. New York: Dover Publications, 1956.

[37] D. Hilbert, *The Foundations of Geometry*. Open Court Publishing, 1980.

[38] Wikimedia Commons, *Lines through a given point p and asymptotic to line r*, 2009. [Online]. Available: https://commons.wikimedia.org/wiki/File:Hyperbolic.svg.

[39] Y. Zhou, B. H. Smith, and T. O. Sharpee, "Hyperbolic geometry of the olfactory space," *Science advances*, vol. 4, no. 8, eaaq1458–eaaq1458, Aug. 2018. DOI: 10.1126/sciadv.aaq1458.

[40] D. Taimina, "Crocheting adventures with hyperbolic planes: Tactile mathematics, art and craft for all to explore, second edition," CRC Press, 2018. DOI: 10.1201/9780203732731.

[41] D. Henderson and D. Taimina, "Experiencing geometry : Euclidean and non-euclidean with history," Jan. 2020. DOI: 10.3792/EUCLID/9781429799850.

[42] D. Thouless, "Defects and geometry in condensed matter physics," *Physics Today*, vol. 56, no. 5, pp. 62–64, 2003. DOI: 10.1063/1.1583539.

[43] J. Bingham and S. Sudarsanam, "Visualizing large hierarchical clusters in hyperbolic space," *Bioinformatics*, vol. 16, no. 7, pp. 660–661, 2000.

[44] M. Boguñá, F. Papadopoulos, and D. Krioukov, "Sustaining the internet with hyperbolic mapping," *Nature Communications*, vol. 1, no. 1, p. 62, Sep. 2010. DOI: 10.1038/ncomms1063.

[45] J. F. Barrett, *The hyperbolic theory of special relativity*, 2019. arXiv: 1102.0462 [physics.gen-ph].

[46] C. Charmousis, "Anti de sitter black holes," 2009.

[47] J. Ratcliffe, *Foundations of Hyperbolic Manifolds*. 1994. DOI: 10.1007/978-3-030-31597-9.

[48] B. Iversen, *Hyperbolic Geometry*. 1993. DOI: 10.2307/3618121.

[49] B. Loustau, *Hyperbolic geometry*, 2020. arXiv: 2003.11180 [math.DG].

[50] P. Dawkins, *Hyperboloid of two sheets*. [Online]. Available: https://tutorial.math.lamar.edu/Classes/CalcIII/QuadricSurfaces.aspx.

[51] Wikimedia Commons, *Hyperboloid projection*, 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:HyperboloidProjection.png.

[52] ——, *Poincare disc hyperbolic parallel lines*, 2008. [Online]. Available: https://commons.wikimedia.org/wiki/File:Poincare_disc_hyperbolic_parallel_lines.svg.

[53] D. Krioukov, F. Papadopoulos, A. Vahdat, and M. Boguñá, "Curvature and temperature of complex networks," *Physical Review E*, vol. 80, no. 3, Sep. 2009. DOI: 10.1103/physreve.80.035101.

[54] M. Gromov, M. Katz, P. Pansu, S. Semmes, J. Lafontaine, and S. M. Bates, *Metric structures for Riemannian and non-Riemannian spaces*. 1999.

[55] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/59dfa2df42d9e3d41f5b02bfc32229dd-Paper.pdf.

[56] A. Cvetkovski and M. Crovella, "Hyperbolic embedding and routing for dynamic graphs," *IEEE INFOCOM 2009*, pp. 1647–1655, 2009. DOI: 10.1109/INFCOM.2009.5062083.

[57] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*, vol. 14, 2002.

[58] I. T. Jolliffe, *Principal component analysis*. New York, NY: Springer, 2002. DOI: 10.1007/b98835.

[59] G. Alanis-Lobato, P. Mier, and M. A. Andrade-Navarro, "Efficient embedding of complex networks to hyperbolic space via their laplacian," *Scientific Reports*, vol. 6, no. 1, p. 30 108, Jul. 2016. DOI: 10.1038/srep30108.

[60] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov, "Popularity versus similarity in growing networks," *Nature*, vol. 489, no. 7417, pp. 537–540, Sep. 2012. DOI: 10.1038/nature11459.

[61] F. Papadopoulos, C. Psomas, and D. Krioukov, "Network mapping by replaying hyperbolic growth," *IEEE/ACM Transactions on Networking*, vol. 23, no. 1, pp. 198–211, Feb. 2015. DOI: 10.1109/TNET.2013.2294052.

[62] G. Alanis-Lobato, P. Mier, and M. A. Andrade-Navarro, "Manifold learning and maximum likelihood estimation for hyperbolic network embedding," *Applied network science*, vol. 1, no. 1, 2016. DOI: 10.1007/s41109-016-0013-0.

[63] X. Zhao, A. Sala, H. Zheng, and B. Zhao, "Efficient shortest paths on massive social graphs," *7th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pp. 77–86, 2011.

[64] X. Zhao and H. Zheng, "Orion: Shortest path estimation for large social graphs," in *WOSN*, 2010.

[65] J. Leskovec and A. Krevl, *SNAP Datasets: Stanford large network dataset collection*, http://snap.stanford.edu/data, Jun. 2014.

[66] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: A general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. Database issue, pp. D535–D539, Jan. 2006. DOI: 10.1093/nar/gkj109.

[67] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Association for Computing Machinery, 2005, pp. 177–187. DOI: 10.1145/1081870.1081893.

[68] B. Rozemberczki and R. Sarkar, "Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, 2020, pp. 1325–1334. DOI: 10.1145/3340531.3411866.

[69] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, Nov. 2009. DOI: 10.1137/070710111.

[70] F. Papadopoulos, R. Aldecoa, and D. Krioukov, "Network geometry inference using common neighbors," *Physical Review E*, vol. 92, no. 2, Aug. 2015. DOI: 10.1103/physreve.92.022807.

[71] M. Boguñá, D. Krioukov, and K. C. Claffy, "Navigability of complex networks," *Nature Physics*, vol. 5, no. 1, pp. 74–80, Jan. 2009. DOI: 10.1038/nphys1130.

[72] F. Papadopoulos, D. V. Krioukov, M. Boguñá, and A. Vahdat, "Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces," *2010 Proceedings IEEE INFOCOM*, pp. 1–9, 2010.

[73] Y.-J. Kim, R. Govindan, B. Karp, and S. Shenker, "On the pitfalls of geographic face routing," in *Proceedings of the 2005 Joint Workshop on Foundations of Mobile Computing*, ser. DIALM-POMC '05, Association for Computing Machinery, 2005, pp. 34–43. DOI: 10.1145/1080810.1080818.

[74] W.-Q. Wang, Q.-M. Zhang, and T. Zhou, "Evaluating network models: A likelihood analysis," *EPL (Europhysics Letters)*, vol. 98, no. 2, Apr. 2012. DOI: 10.1209/0295-5075/98/28004.

[75] T. G. O. Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, no. D1, pp. D330–D338, Nov. 2018. DOI: 10.1093/nar/gky1055.

[76] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLOS Computational Biology*, vol. 5, no. 7, pp. 1–12, Jul. 2009. DOI: 10.1371/journal.pcbi.1000443.

[77] S. Harispe, D. Sánchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," *Journal of Biomedical Informatics*, vol. 48, pp. 38–53, 2014, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.11.006.

[78] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'95, Morgan Kaufmann Publishers Inc., 1995, pp. 448–453. DOI: 10.5555/1625855.1625914.

[79] X. Song, L. Li, P. K. Srimani, P. S. Yu, and J. Z. Wang, "Measure the semantic similarity of go terms using aggregate information content," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 3, pp. 468–476, 2014. DOI: 10.1109/TCBB.2013.176.

[80] G. K. Mazandu and N. J. Mulder, "Dago-fun: Tool for gene ontology-based functional analysis using term information content measures," *BMC Bioinformatics*, vol. 14, no. 1, p. 284, Sep. 2013. DOI: 10.1186/1471-2105-14-284.

[81] ——, "Information content-based gene ontology functional similarity measures: Which one to use for a given biological data type?" *PLOS ONE*, vol. 9, no. 12, pp. 1–20, Dec. 2014. DOI: 10.1371/journal.pone.0113859.

[82] M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, Jan. 2000. DOI: 10.1093/nar/28.1.27.

[83] M. Kutmon, A. Riutta, N. Nunes, *et al.*, "Wikipathways: Capturing the full diversity of pathway knowledge," *Nucleic acids research*, vol. 44, no. D1, pp. D488–D494, Jan. 2016. DOI: 10.1093/nar/gkv1024.

[84] V. Arunachalam, S. Ulrich, F. Raphaele, *et al.*, "A directed protein interaction network for investigating intracellular signal transduction," *Science Signaling*, vol. 4, no. 189, rs8–rs8, Sep. 2011. DOI: 10.1126/scisignal.2001699. [Online]. Available: https://doi.org/10.1126/scisignal.2001699.