Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Στοχαστική Δυναμική Απόψεων για Πρόβλεψη Ενδιαφερόντων σε Κοινωνικά Δίκτυα

Διπλωματικη Εργασια

του

ΜΑΡΙΟΥ Α. ΠΑΠΑΧΡΗΣΤΟΥ

Επιβλέπων: Δημήτρης Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εργαστηριο Επιστημης Υπολογιστων, Λογικης και Αλγοριθμων (CoReLab)

Αθήνα, 25 Ιουνίου 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επιστήμης Υπολογιστών, Λογικής και Αλγορίθμων (CoRe-Lab)

Στοχαστική Δυναμική Απόψεων για Πρόβλεψη Ενδιαφερόντων σε Κοινωνικά Δίκτυα

Διπλωματικη Εργασια

του

ΜΑΡΙΟΥ Α. ΠΑΠΑΧΡΗΣΤΟΥ

Επιβλέπων: Δημήτρης Φωτάκης
Αν. Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Ιουνίου 2020.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτρης Φωτάκης	Άρης Παγουρτζής	Νίκος Παπασπύρου
Αν. Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, 25 Ιουνίου 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Επιστήμης Υπολογιστών, Λογικής και Αλγορίθμων (CoRe-Lab)

(Υπογραφή)

.....

Μαριος Α. Παπαχρηστου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σε αυτήν τη διπλωματική εργασία, εκμεταλλευόμαστε τη δομή πυρήνα-περιφέρειας και τις ισχυρές ομοφιλικές ιδιότητες των διαδικτυακών κοινωνικών δικτύων για την ανάπτυξη ταχύτερων και ακριβέστερων αλγορίθμων για την πρόβλεψη ενδιαφέροντος των χρηστών. Ο πυρήνας των σύγχρονων κοινωνικών δικτύων αποτελείται από σχετικά λίγους επιδραστικούς χρήστες, των οποίων τα προφίλ ενδιαφέροντος είναι διαθέσιμα στο κοινό, ενώ η πλειονότητα των περιφερειακών χρηστών ακολουθεί αρκετά από αυτά με βάση κοινά ενδιαφέροντα. Η προσέγγισή μας είναι να απορρίψουμε ένα μεγάλο μέρος του δικτύου, που αποτελείται από συνδέσεις μεταξύ περιφερειακών κόμβων και να προβλέψουμε τα συμφέροντα των περιφερειακών κόμβων ξεκινώντας από τα συμφέροντα των επιδραστικών τους συνδέσεων. Για το σκοπό αυτό, χρειαζόμαστε ένα επίσημο μοντέλο που να εξηγεί πώς τα κοινά ενδιαφέροντα οδηγούν σε συνδέσεις δικτύου. Με γνώμονα τις ισχυρές ομοφιλικές ιδιότητες που παρουσιάζουν τα σύγχρονα διαδικτυακά κοινωνικά δίκτυα (ΔΚΔ), προτείνουμε ένα νέο γενετικό μοντέλο βασισμένο στη δυναμική της γνώμης για απόψεις στα ΔΚΔ με βάση την ανταλλαγή απόψεων των πρακτόρων με τους k πλησιέστερους γείτονές τους. Πιο συγκεκριμένα, προτείνουμε ένα στοχαστικό μοντέλο σχηματισμού ενδιαφέροντος, το Nearest Neighbor Influence Model (NNIM), το οποίο είναι εμπνευσμένο από το μοντέλο Hegselmann-Krause και στοχεύει να εξηγήσει πώς η ομοφιλία διαμορφώνει το δίκτυο. Με βάση το NNIM, αναπτύσσουμε μια αποτελεσματική προσέγγιση για την πρόβλεψη των συμφερόντων των περιφερειακών χρηστών. Στη συνέχεια αναπτύσσουμε έναν αλγόριθμο για συμπερασμό παραμέτρων μέσω Μεταβολικού Expectation Maximization (EM) και αποδεικνύουμε ότι, κάτω από εύλογες υποθέσεις, οι εξισώσεις συμπερασμού μέσου πεδίου μοιάζουν με τις συμβατικές επαναληπτικές εξισώσεις των κλασικών ΔΔΑ και μπορούν να κλιμακωθούν αποτελεσματικά σε δίκτυα με εκατομμύρια χρήστες. Αποδεικνύουμε θεωρητικά ότι οι εξισώσεις μέσου πεδίου συγκλίνουν εντός πεπερασμένου χρόνου και ότι η Απόσταση Ολικής Μεταβολής (Total Variation Distance) σε κάθε χρονικό βήμα φράσσεται αυστηρά από πάνω από μια εκθετική συνάρτηση με βάση $1/\sqrt{k}$. Τέλος, αξιοποιούμε την ικανότητα του μοντέλου μας να προβλέπει ενδιαφέροντα μέσω κόμβων με μεγάλη επιρροή στα διαδικτυακά κοινωνικά δίκτυα, ξεπερνώντας τα σχετικά μοντέλα δυναμικής γνώμης και μεθόδους ενσωμάτωσης κόμβων (node embeddings) μέσω τυποποιημένων σημείων αναφοράς που υπάρχουν στην πρόσφατη βιβλιογραφία από άποψη των μετρικών Μέσου Τετραγωνικού Σφάλματος (ΜΤΣ) και Περιοχής κάτω από τη Χαρακτηριστική Καμπύλη Λειτουργίας του Δέκτη (ΠΧΚΛΔ) και υπολογιστικού χρόνου σε δίκτυα διαφόρων μεγεθών.

Λέξεις Κλειδιά. δυναμική διάδοσης απόψεων, στοχαστική δυναμική διάδοσης απόψεων, k πλησιέστεροι γείτονες, διάσημοι, μέγιστη κάλυψη, μεγιστοποίηση αναμενόμενης πιθανοφάνειας, συναρτήσεις Lyapunov, διαδικτυακά κοινωνικά δίκτυα, πρόβλεψη ενδιαφερόντων, ενσωμάτωση κόμβων

Abstract

In this Diploma Thesis, we take advantage of the core-periphery structure and the strong homophilic properties of online social networks to develop faster and more accurate algorithms for user interest prediction.

The core of modern social networks consists of relatively few influential users, whose interest profiles are publicly available, while the majority of peripheral users follow enough of them based on common interests. Our approach is to discard a large part of the network, consisting of connections between peripheral nodes, and to predict the interests of the peripheral nodes starting from the interests of their influential connections. To this end, we need a formal model that explains how common interests lead to network connections.

Driven by strong homophilic properties that modern Online Social Networks (OSN) exhibit, we propose a novel opinion-dynamics-based generative model for opinions in OSN based on the agents' opinions' exchange with their k nearest neighbors. More specifically, we propose a stochastic interest formation model, the Nearest Neighbor Influence Model (NNIM), which is inspired by the Hegselmann-Krause model and aims to explain how homophily shapes the network. Based on NNIM, we develop an efficient approach for predicting the interests of the peripheral users.

Furthermore, we develop an algorithm for parameter inference via Variational Expectation Maximization (EM) and prove that under reasonable assumptions, the mean-field inference equations resemble the conventional Opinion Dynamics iterative equations and can scale efficiently to networks with millions of agents. We theoretically prove that the mean-field equations converge within finite time and the total variation distance at each timestep is upper bounded by an exponential function with base $1/\sqrt{k}$.

Finally, we leverage our model's ability to predict interests via highly-influential nodes in Online Social Networks overcoming relevant Opinion Dynamics models and node embedding methods via standardized benchmarks that exist in recent literature in terms of AUC-ROC, RMSE and computational time in networks of various sizes. Concluding, we set up a theoretical basis for the stochastic formulation of Opinion Dynamics and establish a framework for label inference in very large graphs.

Keywords. opinion dynamics, stochastic opinion dynnamics, k nearest neighbors, influencers, maximum coverage, variational expectation-maximization, Lyapunov potential, locality sensitive hashing, online social networks, interest prediction, multilabel classification, node embeddings, auc-roc, rmse

Ευχαριστίες

Με την ολοκλήρωση της παρούσας Διπλωματικής Εργασίας και του προπτυχιακού μου Τίτλου Σπουδών αισθάνομαι την υποχρέωση να ευχαριστήσω εγκαρδίως μια σειρά από ανθρώπους που μου παρείχαν βοήθεια και συμπαράσταση καθόλη τη διάρκεια της πορείας μου.

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα και Καθηγητή μου κ. Δημήτρη Φωτάκη, ο οποίος με εισήγαγε στα μονοπάτια της Πληροφορικής από τα πρώτα έτη των σπουδών μου. Υπήρξε δίπλα μου σε όλη τη διάρκεια εκπόνησης της παρούσας εργασίας, η οποία δεν θα ήταν εφικτή χωρίς τη συνεισφορά και την καθοδήγησή του. Επιπλέον, υπήρξε ένας εξαιρετικός μέντορας και μου προσέφερε ανεκτίμητη βοήθεια όσον αφορά τα μελλοντικά μου βήματα ως νέος ερευνητής.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Καθηγητή Διομίδη Σπινέλλη, τον οποίο γνώρισα ως επιβλέποντα στην πρακτική μου στο Google Summer of Code το καλοκαίρι του 2018, και ο οποίος παραμένει μέντοράς μου μέχρι και σήμερα. Η εμπειρία του, τόσο ως μηχανικού λογισμικού όσο και ως επιστήμονα πληροφορικής υπήρξε αρωγός για την ανάπτυξη των ικανοτήτων μου, υποδεικνύοντάς μου πως να είμαι επαγγελματίας στη δουλειά μου. Η έκκλησή του για συμμετοχή μου στο Business Analytics Lab του ΟΠΑ ως προπτυχιακός ερευνητής μου άνοιξε νέους ορίζοντες όσον αφορά την ερευνητική μου δραστηριότητα, καταφέρνοντας, μέσω αυτού, να δημοσιεύσω εργασία στο ESEC/FSE.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους φίλους μου και τους συμφοιτητές μου για την υποστήριξη που μου παρείχαν όλα αυτά τα χρόνια, οδηγώντας με στο να αναπτύξω την προσωπικότητα και το χαρακτήρα μου. Θα τους θυμάμαι για πάντα, όπου και να βρίσκομαι. Τους παραθέτω με τυχαία σειρά: Σταύρος Κ., Γιάννης Δ., Χάρης Φ., Χρήστος Π., Ειρώ Μ., Δημήτρης Χ., Κώστας Σ., Κωσταντίνος Α., Παναγιώτης Κ., Χάρης Π., Φώτης Α., Δημήτρης Κ., Ειρήνη Τ., Κωσταντίνος Κ., Μαίρη Π., Δημήτρης Γ., Μαριλένα Οικ., Ελένη Π., Ελισσαίος Σ., Άγγελος Α., Μιχαήλ Φ., Μιχάλης Κ., Ηλίας Π., Λυκούργος Μ., Φωτεινή Δ., Γιάννης Μ., Πέτρος Ν., Ιωάννα Δ., Πάνος Μ., Γιώτα Κ., Γιώργος Β., Οδυσσέας Μ., Γιάννης Β., Μάνος Π., Μαίρη Κ., Σοφία Γ., Στέφανος Σ., Παναγιώτης Τ., Λυδία Ν., Δημήτρης Μ., Μιλτιάδης Σ., Γιώργος Ρ., Νεοκλής Β., Ραφαήλ Κ., Ιάσων Ν., Χαρίτων Χ.,

Ευχαριστώ, επίσης, την Τριμελή Επιτροπή εξέτασης της παρούσας Εργασίας, τους κκ. Αριστείδη Παγουρτζή και Νίκο Παπασπύρου, για τις πολύτιμες υποδείξεις και παρατηρήσεις τους πάνω στην Εργασία.

Θα ήθελα να απευθύνω ένα πολύ μεγάλο και εγκάρδιο 'ευχαριστώ' στους Ιατρούς Β.Κ. και Μ.Σ. η συνεισφορά των οποίων υπήρξε καιριότατη για τη ζωή και την αρτιμέλειά μου

στην σοβαρότατη περιπέτεια υγείας που είχα το χειμώνα του 2014-15. Η επιστημονική τους εξειδίκευση και η ηθική εργασίας υπήρξαν παραδειγματικά για μένα, πρωτίστως για να μην χάσω το κουράγιο μου στο διάβασμα μέσα στο νοσοκομείο — δεδομένου ότι εκείνο τον καιρό έδινα Πανελλήνιες — αλλά και στο πώς πρέπει να συνυπάρχει η επιστήμη με την αρετή.

Κυρίως, όμως, θα ήθελα να ευχαριστήσω την οικογενειά μου, τους γονείς μου Απόστολο και Παναγιώτα και τις αδερφές μου Αναστασία και Ελένη-Αγγελική, που βρίσκονταν πάντα δίπλα μου σε οποιαδήποτε επιλογή μου και με ωθούσαν πάντα να ακολουθώ τα όνειρά μου. Η υπέρμετρη αγάπη τους υπήρξε, και συνεχίζει να υπάρχει, το μεγαλύτερο στήριγμα στην πορεία μου. Η παρούσα Διπλωματική Εργασία είναι αφιερωμένη σε αυτούς.

Μάριος Παπαχρήστου
Αθήνα, 25 Ιουνίου 2020

Σαν βγεις στον πηγαιμό για την Ιθάκη,
να εύχεσαι να ΄ναι μακρύς ο δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις.
— Ιθάκη, Κ.Π. Καβάφης

Contents

Περίληψη 1

Abstract 2

Ευχαριστίες 3

Contents 7

Εκτενής Περίληψη στα Ελληνικά 9
Εισαγωγή . . . 9
Κίνητρο . . . 10
Συνεξελικτικά Παίγνια Διαμόρφωσης Άποψης . . . 12
Γενετικό Μοντέλο . . . 13
Ομοφιλία . . . 13
Γενετικό Μοντέλο Επιρροής από τους k Κοντινότερους Γείτονες . . . 15
Θεωρητικές Ιδιότητες . . . 20
Πρόβλεψη Ενδιαφερόντων σε ΔΚΔ από 'Δίασημους' Χρήστες σε Δίκτυα με Δομή Πυρήνα-Περιφέρειας . . . 23
Δεδομένα . . . 24
Πειραματική Διάταξη . . . 24
Συζήτηση . . . 25
Συμπεράσματα . . . 26
Αντίκτυπος . . . 26
Μελλοντικές Προεκτάσεις . . . 27

1 Introduction 29
1.1 Motivation . . . 29
1.2 Approach and Contribution . . . 30
1.3 Thesis Structure . . . 31

2 Characteristics of Social Networks 35
2.1 Motivation . . . 35
2.2 Homophily . . . 35

2.3 Scale-free Degree Distributions . . . 36
2.3.1 Examples of Scale-free Distributions . . . 37
2.4 Small-world . . . 38
2.5 Densification Laws and Shrinking Diameters . . . 39
2.6 Core-periphery structure . . . 40

3 Generative Models 45
3.1 Motivation . . . 45
3.2 Maximum Likelihood Estimation . . . 45
3.3 Inference of Latent Variable Models through Expectation-Maximization . . . 46
3.3.1 Latent Variable Models . . . 46

4 Opinion Dynamics 53
4.1 Motivation . . . 53
4.2 Models of Opinion Dynamics . . . 53
4.3 Coevolutionary Opinion Formation Games . . . 58

5 Dynamical Systems on the Steady State 61
5.1 Lyapunov Stability and Lyapunov Functions . . . 61
5.2 Markov Chains . . . 63

6 Unsupervised Feature Learning 67
6.1 Dimensionality Reduction . . . 67
6.1.1 Principal Components Analysis . . . 67
6.1.2 Random Projections . . . 68
6.1.3 Similarity for Sets: MinHash . . . 69
6.2 Nearest Neighbor Search . . . 70
6.2.1 Brute-force Approach . . . 70
6.2.2 Intrinsic Dimensionality . . . 71
6.2.3 KD Trees . . . 72
6.2.4 Ball Tree . . . 72
6.2.5 Locality Sensitive Hashing . . . 73
6.3 Dynamic Continuous Indexing . . . 73

7 Stochastic Opinion Dynamics for Interest Prediction in Social Networks 77
7.1 The Nearest Neighbor Influence Model . . . 77
7.1.1 Model Inference through Variational Expectation-Maximization . . . 79
7.1.2 Model Convergence and Convergence Rate . . . 86
7.1.3 Complexity and Implementation . . . 93
7.2 Generalization Example: Multivariate Gaussian Opinions . . . 93
7.3 Experiments . . . 94
7.3.1 Datasets . . . 94

7.3.2 Influencer Identification . . . 96
7.3.3 Further Processing . . . 97
7.3.4 Experimental Setting . . . 97
7.3.5 Discussion . . . 97
7.3.6 More Experiments . . . 98
7.4 Further Related Work . . . 100
7.5 Conclusion . . . 102
7.6 Future Work . . . 103

A Concentration Bounds **105**
A.1 Motivation . . . 105
A.2 Talagrand's Inequality . . . 108
A.2.1 Statement of Talagrand's Inequality . . . 109
A.3 McDiarmid's Inequality . . . 110
A.4 Nomenclature for the Asymtotic Behaviour of Randomized Algorithms . . . 111

List of Figures **113**

List of Tables **115**

Bibliography **117**

Bibliography **117**

Εκτενής Περίληψη στα Ελληνικά

Εισαγωγή

Ο πλούτος των σημερινών δικτύων είναι τεράστιος. Κάποιος μπορεί να παρατηρήσει δίκτυα σχεδόν παντού: κοινωνικά δίκτυα, δίκτυα κυκλοφορίας, βιολογικά δίκτυα, δίκτυα παραγωγής και δίκτυα αλληλεπίδρασης σωματιδίων είναι μερικά πολύ ζωντανά παραδείγματα. Η τάση των διαφόρων μορφών ζωής, κάτω από τη φύση ή την κοινωνία, να συνδέονται και να συνεργάζονται δημιουργούν πλούσια πρότυπα που διέπουν την καθημερινή μας ζωή.

Είναι καλά κατανοητό ότι τα περισσότερα διαδικτυακά κοινωνικά δίκτυα μεγάλης κλίμακας (ΔΚΔ) εμφανίζουν τη λεγόμενη δομή πυρήνα-περιφέρειας core-periphery structure (βλ. [48, 114, 88, 115, 99] και τις αναφορές εντός). Οι κόμβοι αυτών των δικτύων χωρίζονται τους χωρίζονται φυσικά σε έναν πυρήνα C κόμβων που είναι στενά συνδεδεμένοι μεταξύ τους και σε ένα περιφερειακό σύνολο U, όπου οι κόμβοι είναι αραιά συνδεδεμένοι, αλλά είναι σχετικά καλά συνδεδεμένο με τον πυρήνα. Στις περισσότερες περιπτώσεις, οι κεντρικοί κόμβοι κυριαρχούν σχεδόν στο υπόλοιπο δίκτυο, με την έννοια ότι ένα μικρό κλάσμα δn κόμβων υψηλού βαθμού κυριαρχεί σε ένα κλάσμα $(1-a)n$ των 'δεσμευμένων' (engaged) κόμβων του δικτύου (όπου "δεσμευμένο' αναφέρεται σε κόμβους με βαθμό πάνω από ένα συγκεκριμένο όριο). Εάν περιοριζόμαστε μόνο στους δεσμευμένους κόμβους, ακόμη και ένα υπογραμμικό κλάσμα κόμβων κυριαρχεί σχεδόν σε όλα (δείτε επίσης [16, 17, 18]). Αυτοί οι σημαντικοί κόμβοι πυρήνα, οι οποίοι διαθέτουν μεγάλο αριθμό εισερχόμενων συνδέσεων, ή ακολούθων, είναι επίσης γνωστοί (και εξυπηρετούν) ως διάσημοι/επηρεαστές (celebrities/influencers) του δικτύου. Οι επηρεαστές τείνουν να εκθέτουν δημόσια — κυρίως για εμπορικούς λόγους [52, 37, 49, 100, 19] — τις πληροφορίες προφίλ τους (φίλοι και ενδιαφέροντα), επομένως οι πληροφορίες μπορούν να συγκεντρωθούν εύκολα, για παράδειγμα μέσω κλήσεων σε REST APIs.

Μια άλλη σημαντική κινητήρια δύναμη που διαμορφώνει τη δομή του κοινωνικού δικτύου είναι η 'ομοφιλία', δηλαδή η ιδιότητα υπό την οποία τα συνδεδεμένα άτομα σε ένα κοινωνικό δίκτυο έχουν παρόμοια ενδιαφέροντα [73, 72]. Τα σύγχρονα ΔΚΔ μεγάλης κλίμακας φαίνεται να εμφανίζει ισχυρές ομοφιλικές τάσεις, οι οποίες αποτελούσαν σημαντικό μέρος των κινήτρων μας (βλ. Επίσης Κεφάλαιο 2).

Προσέγγιση και Συμβολή. Σε αυτήν τη Διπλωματική Εργασία, αξιοποιούμε τις ομοφιλικές τάσεις και τη δομή πυρήνα-περιφέρειας του σύγχρονου ΟΣΝ για να αποκτήσουμε επεκ-

τάσιμες και ακριβείς μεθόδους μάθησης για την πρόβλεψη των ενδιαφερόντων των περιφερειακών χρηστών ενός δικτύου. Η προσέγγισή μας είναι να προσδιορίσουμε και να χρησιμοποιήσουμε τους παράγοντες επιρροής του δικτύου ως "steady-state trend-setters" και να αφήσουμε το δίκτυο γύρω τους να εξελιχθεί σύμφωνα με μια επαναληπτική διαδικασία που ξεκίνησε από μια συγκέντρωση των χαρακτηριστικών των επηρεαστών. Ο υπογραμμικός αριθμός των επηρεαστών επιτρέπει μια αρκετά γρήγορη προετοιμασία (στη χειρότερη περίπτωση έντονα υπο-τετραγωνικό χρόνο) των ενδιαφερόντων των χρηστών. Εμπνευσμένο από το τα συνεξελικτικά παίγνια διαμόρφωσης άποψης [46, 11], στη συνέχεια αντιμετωπίζουμε το δίκτυο ως αποτέλεσμα μιας φυσικής δυναμικής διαδικασίας, όπου κάθε περιφερειακός χρήστης ενημερώνει τις δυνατότητές της σύμφωνα με τα ενδιαφέροντά της k -εγγιέστεροι γείτονες στην περιφέρεια, έως ότου επιτευχθεί συναίνεση (βλ. επίσης Κεφάλαιο 4).

Χρησιμοποιούμε τον χώρο ενδιαφέροντος του δικτύου που δημιουργείται από αυτήν τη διαδικασία για να συμπεράνουμε την πιθανότητα ότι ένας περιφερειακός χρήστης υιοθετεί συγκεκριμένα ενδιαφέροντα (μια εργασία ισοδύναμη με την ταξινόμηση πολλαπλών ετικετών). Το κλειδί για την επεκτασιμότητα του αλγορίθμου είναι ότι καθ΄όλη τη διάρκεια της διαδικασίας, κάθε περιφερειακός χρήστης αλληλεπιδρά μόνο με τους k πλησιέστερους γείτονές του.

Πιο συγκεκριμένα, ένα βασικό μέρος της προσέγγισής μας είναι το Γενετικό Μοντέλο Επιρροής από k Κοντινότερους Γείτονες (Nearest Neighbor Influence Model/NNIM), μια στοχαστική επαναληπτική διαδικασία σύμφωνα με την οποία οι χρήστες εξελίσσουν τα ενδιαφέροντά τους. Σε κάθε χρονικό βήμα, κάθε περιφερειακός χρήστης δειγματίζει έναν νέο φορέα δυαδικού ενδιαφέροντος με βάση τα ενδιαφέροντα των πλησιέστερων γειτόνων το στην περιφέρεια. Η γενική δομή του NNIM εμπνέεται από το μοντέλο Hegselmann-Krausse [46]. Ωστόσο, το NNIM είναι στοχαστικό και χρησιμοποιείται ως γενετικό μοντέλο, με στόχο να εξηγήσει, μέσω της ομοφιλίας, τη συνεκτίμηση της δομής του δικτύου και τα ενδιαφερόντα των περιφερειακών χρηστών (βλ. Κεφάλαιο 7).

Περιληπτικά, η μέθοδος πρόβλεψής μας στοχεύει στην ανάκτηση των λανθανόντων ενδιαφερόντων των περιφερειακών χρηστών που μεγιστοποιούν την πιθανοφάνεια του μοντέλου. Αν και η ιδέα είναι απλή, η αποτελεσματική εφαρμογή της απαιτεί σημαντική προσπάθεια και φροντίδα (βλ. Ενότητα 7.1.1 και Κεφάλαιο 3). Χρησιμοποιούμε Variational Expectation-Maximization, λόγω της λανθάνουσας φύσης του NNIM, δεδομένου ότι η άμεση μεγιστοποίηση της πιθανότητας λογ είναι δυσδιάκριτη. Ως αποτέλεσμα, λαμβάνουμε μια απλοποιημένη προσέγγιση μέσου πεδίου του NNIM , η οποία είναι παρόμοια με τις κλασικές εξισώσεις της κλασικής δυναμικής γνώμης, καθστώντας επιπλέον δυνατή την εγκαθίδρυση σύνδεσης μεταξύ της στοχαστικής και της ντετερμινιστικής δυναμικής της γνώμης.

Κίνητρο

Σήμερα, οι άνθρωποι τείνουν να επηρεάζονται από άλλους που έχουν κοινά ενδιαφέροντα. Αυτή η ιδιότητα των κοινωνικών δικτύων — που υφίσταται για ένα λογικό χρονικό διάστημα στις κοινωνικές επιστήμες — ονομάζεται ομοφιλία. Οι άνθρωποι μοιράζονται συνεχώς τις απόψεις τους και κάθε άτομο επηρεάζει έναν άλλο με δυναμικό τρόπο. Η ανταλλαγή πληρο-

φοριών συμβαίνει συνήθως τοπικά μεταξύ των ανθρώπων έως ότου επιτευχθεί συναίνεση, δηλαδή οι απόψεις συγκλίνουν σε ένα μόνο σημείο. Οι απόψεις μπορεί να αναφέρονται σε πολιτικές σχέσεις, υιοθεσίες νέων τάσεων, χόμπι και ενδιαφερόντων γενικά. Τέλος, αυτό που καθιστά την ανάγκη για μελέτη δυναμικής γνώμης πανταχού παρούσα είναι η εμφάνιση σύγχρονων διαδικτυακών κοινωνικών δικτύων (ΔΚΔ).

Η γραμμή έρευνας στην ΔΔΑ εστιάζει κυρίως σε μοντέλα που ορίζουν κάποια έννοια της «γειτονιάς» γύρω από κάθε χρήστη και μια διαδικασία ενημέρωσης που λαμβάνει χώρα ως αλληλεπίδραση μεταξύ του χρήστη και των γειτόνων του. Η διαδικασία είναι επαναληπτική και επαναλαμβάνεται έως ότου οι χρήστες φτάσουν σε ένα σημείο στο οποίο οι απόψεις τους δεν αλλάζουν — το οποίο αναφέρεται στη βιβλιογραφία ως *συναίνεση*. Μεταξύ των γνωστών μοντέλων ΔΔΑ είναι το μοντέλο Friedkin-Johnsen (FJ) [**?**], το μοντέλο DeGroot [24] και το μοντέλο Hegselmann-Krause (HK) [46]. Παρόλο που αυτά τα μοντέλα παρουσιάζουν ισχυρές μαθηματικές ιδιότητες όπως η σταθερότητα και η σύγκλιση πεπερασμένου χρόνου σε μια κατάσταση συναίνεσης, δεν διαθέτουν θεμελιώδεις ιδιότητες όπως η *στοχαστικότητα* που μπορούν να αξιοποιήσουν τη δύναμη του μοντέλου με πολλούς τρόπους. Μια φυσική επέκταση για τη συμπερίληψη στοχαστικών ιδιοτήτων είναι ότι οι απόψεις μια χρονική στιγμή δεν αντιπροσωπεύονται από ντετερμινιστικές μεταβλητές αλλά αντλούνται από μια κατανομή $\mathcal{D}$ με ένα σύνολο παραμέτρων $\boldsymbol{\theta}^{(t)}$ που καθορίζονται μέσω αλληλεπιδράσεων των παραγόντων στο προηγούμενο βήμα. Το μοντέλο που προτείνουμε μέσω αυτής της εργασίας, έχει αυτές τις ιδιότητες. Πιο συγκεκριμένα, σε κάθε βήμα t κάθε χρήστης έχει μια δυαδική άποψη που δίδεται από μια κατανομή Bernoulli. Στη συνέχεια, ο χρήστης «κοιτάζει» τους k-Κοντινότερους Γείτονες (ΚΚΓ) θέτει την πιθανότητα να ενημερώσει τη γνώμη της ίση με τον μέσο όρο των παρατηρούμενων απόψεων. Μολονότι, αυτής της φύσεως τα μοντέλα κατέχουν προφανή προτερήματα έναντι των ντετερμινιστικών αναλόγων τους, τα συγκεκριμένα μοντέλα υποφέρουν από θεμελιώδη υπολογιστικά προβλήματα, με το σημαντικότερο από αυτά την εξαγωγή παραμέτρων. Η απευθείας εκτίμηση μέγιστης πιθανοφάνειας (ΕΜΠ) στα μοντέλα αυτά χρειαεται εκθετικό χρόνο για να υπολογιστεί και εν προκειμένω το πρόβλημα απευθείας συμπερασματολογίας είναι πρακτικώς αδύνατο λόγω της λανθάνουσας φύσης του μοντέλου. Αυτό το κενό έρχεται να καλυφθεί από τον αλγόριθμο Expectation-Maximization (EM) που επιτρέπει την προσεγγιστική και υπολογιστικά αποδοτική μεγιστοποίηση της αναμενόμενης τιμής της πιθανοφάνειας υποθέτοντας μια μεταβολική κατανομή πάνω στις κρυφές μεταβλητές. Έτσι, με γνώση μόνο της αρχικής κατάστασης των απόψεων καταφέρνουμε να εξάγουμε τις παραμέτρους για όλο το γενετικό μοντέλο (ΓΜ). Χρησιμοποιώντας την προσέγγιστη του μέσου πεδίου (mean-field) (ΜΠ) καταφέρνουμε να εξάγουμε επαναληπτικές εξισώσεις για την συμπερασματολογία παραμέτρων για τις οποίες *αποδεικνύουμε θεωρητικά* ότι συγκλίνουν σε πεπερασμένο χρόνο καθώς και ότι για μεγάλο αριθμό χρηστών η ταχύτητα σύγκλισης — ως προς την Απόσταση Ολικής Μεταβολής (ΑΟΜ) - φράσσεται από μια εκθετική συνάρτηση με βάση $1/\sqrt{k}$. Επιπλέον, παρατηρούμε ότι η μέθοδος μας γενικεύεται σε μια οικογένεια στοχαστικών μοντέλων δυναμικής απόψεων, στα οποία μπορεί να γίνει συναγωγή συμπερασμάτων με παρόμοιο τρόπο.

Τέλος, επικυρώνουμε τη μέθοδο μας σε δίκτυα πραγματικού κόσμου με τη χρήση των

συμπερασματικών παραμέτρων ως πιθανότητες να προτείνουμε συγκεκριμένα ενδιαφέροντα σε συγκεκριμένους χρήστες.

Συνεξελικτικά Παίγνια Διαμόρφωσης Άποψης

Η οικογένεια θεωρητικών μοντέλων σχηματισμού γνώμης που σχετίζονται περισσότερο με τη δική μας είναι τα Συνεξελικτικά Παίγνια Διαμόρφωσης Άποψης (Coevolutionary Opinion Formation Games / ΣΠΔΑ) που εισάγονται στο [11]. Σύμφωνα με τα ΣΠΔΑ, οι χρήστες του δικτύου εξελίσσουν τις απόψεις τους μαζί με τις γειτονιές τους (όπως στο μοντέλο HK). Αυτό επεκτείνει το έργο των Bindel et al. [12] που επιδιώκει την ελαχιστοποίηση του κόστους διαφωνίας των χρηστών, όπως θα δούμε παρακάτω, αλλά με το δίκτυο σταθερό, φτάνοντας τελικά σε μια θεωρητική παιγνιοθεωρητική κατανόηση του μοντέλου FJ. Οι συγγραφείς του [11] γενικεύουν τη λειτουργία κοινωνικού κόστους που επιβάλλουν οι Bindel et al. και φράσσουν στενά το Τίμημα της Αναρχίας (ΤΑ) [1] και το ερμηνεύουν ως έναν τρόπο να αποδωθεί αξία στο πόσο οι κόμβοι εκτιμούν την εγγενή τους άποψη και τις απόψεις των φίλων τους.

Σε ένα ΣΠΔΑ, υπάρχουν n παίκτες καθένας από τους οποίους έχει εγγενή γνώμη s_i και εκφράζει γνώμη z_i (όπου γενικά $s_i \neq z_i$). Στόχος κάθε παίκτη είναι να ελαχιστοποιήσει το κόστος $C_i(\boldsymbol{z})$ που είναι συνάρτηση των εγγενών απόψεων s_i και των εκφρασμένων απόψεων $\boldsymbol{z} = (z_1, \ldots, z_n)$ όλων των παικτών. Το αθροιστικό κοινωνικό κόστος ορίζεται ως $C(\boldsymbol{z}) = \sum_{i=1}^{n} C_i(\boldsymbol{z})$. Οι συγγραφείς θεωρούν δύο παιχνίδια. Το πρώτο είναι το Συμμετρικό ΣΠΔΑ (ΣΣΠΔΑ) όπου δίνεται η συνάρτηση κόστους κάθε παίκτη ως

$$C_i(z_i, \boldsymbol{z}_{-i}) = \sum_{j \neq i} f_{ij}(z_i - z_j) + w_i g_i(z_i - s_i) \tag{1}$$

όπου f_{ij} και g_i είναι γνωστές πραγματικές συναρτήσεις που είναι κυρτές, συνεχώς διαφορίσιμες και συμμετρικές, δηλαδή $f_{ij}(-x) = f_{ij}(x)$ και $g_i(x) = g_i(-x)$ και $g(0) = 0$. Στη συμμετρική ρύθμιση $f_{ij} = f_{ji}$ που κάνει το παιχνίδι συμμετρικό ωρτ σε ζευγάρια παικτών. Το έργο των Bindel et al. ορίζει $g(x) = x^2$ και $f_{ij}(x) = w_{ij}x^2$ όπου $w_{ij} = w_{ji}$ αντιπροσωπεύει το βάρος της ακμής $\{i, j\}$ μεταξύ παικτών i και j. Όταν οι συναρτήσεις είναι κυρτές, με χρήση του δυναμικού

$$\phi(\boldsymbol{z}) = \sum_{i} w_i g_i(z_i - s_i) + \sum_{i<j} f_{ij}(z_i - z_j) \tag{2}$$

μπορεί κανείς να αποδείξει την ύπαρξη Ισορροπίας Nash. Επιπλέον, για την εξαγωγή του φράγματος για το ΤΑ οι συγγραφείς εξετάζουν το σύνολο

[1] Το ΤΑ ενός παιχνιδιού μετρά τον τρόπο με τον οποίο η αποτελεσματικότητα ενός συστήματος υποβαθμίζεται λόγω της εγωιστικής συμπεριφοράς των παικτών του. Είναι μια γενική έννοια που μπορεί να επεκταθεί σε διαφορετικά συστήματα και έννοιες της αποτελεσματικότητας. Δίνεται ένα παιχνίδι $G = (\mathcal{N}, \mathcal{S}, u)$ με ένα σετ $\mathcal{N}$ των παικτών στρατηγική $\mathcal{S}$ για κάθε παίκτη $i \; in \mathcal{N}$, συναρτήσεις ωφέλειας $u_i : \mathcal{S} \; to \mathbb{R}$, μια συνάρτηση πρόνοιας $W : \mathcal{S} \; to \mathbb{R}$, όπως το utilitarian objective $W(s) = \sum_{i \; in \mathcal{N}} u_i(s)$ ή ο ισότιμος στόχος (egalitarian objective) $W(s) = \min_{i \; in \mathcal{N}} u_i(s)$ και ένα σύνολο $\mathcal{E} \subseteq \mathcal{S}$ ισορροπιών, το ΤΑ ορίζεται ως $\text{TA} = \frac{\max_{s \; in \mathcal{S}} W(s)}{\min_{e \in \mathcal{E}} W(e)}$.

$$\mathcal{H}_{x,y,f} = \left\{ (\lambda, \mu) \middle| f(x) + \frac{y-x}{2} f'(x) \leq \lambda f(y) + \mu f(x) \forall x, y \geq 0, \ f \text{είναι μια συνάρτηση βάρους} \right\} \tag{3}$$

και το σύνολο

$$\mathcal{H}_{u,v,g} = \left\{ (\lambda, \mu) \middle| g(u) + (v-u) g'(u) \leq \lambda g(v) + \mu g(u) \forall x, y \geq 0, \ g \text{είναι μια συνάρτηση βάρους} \right\} \tag{4}$$

και τα σύνολα $\mathcal{A}_1, \mathcal{A}_2$ που δίνονται ως $\mathcal{A}_1 = \bigcup_f \mathcal{H}_{x,y,f}$ και $\mathcal{A}_2 = \bigcup_g \mathcal{H}_{u,v,g}$. Οι συγγραφείς χρησιμοποιούν την τεχνική του Local Smoothness του [92]. δείχνουν ότι για κάθε $(\lambda, \mu) \in \mathcal{A}_1 \cap \mathcal{A}_2$ η τιμή $\lambda/(1-\mu)$ είναι το ανώτερο όριο του ΤΑ και το $\zeta = \min_{(\lambda,\mu)\in\mathcal{A}_1\cap\mathcal{A}_2} \frac{\lambda}{1-\mu}$ είναι το καλύτερο ανώτερο όριο. Όταν οι συναρτήσεις είναι κυρτές και διαφοροποιημένες, το ΤΑ είναι το πολύ 2. Τέλος, οι συγγραφείς παρέχουν μια γενική κατασκευή κάτω ορίου για το ΣΣΠΔΑ.

Στο ΣΠΔΑ Κοντινότερων Γειτόνων (ΣΠΔΑΚΚΓ), κάθε παίκτης κοιτάζει τους k πλησιέστερους γείτονές της (με συνεπές σπάσιμο ισοπαλιών) σε σχέση με τα s_i και σχηματίζει το σετ $K(\boldsymbol{z}, i)$ και υποφέρει κόστος

$$C_i(z_i, \boldsymbol{z}_{-i}) = \sum_{j \in K(\boldsymbol{z}, i)} (z_j - z_i)^2 + \alpha k (z_i - s_i)^2 \tag{5}$$

Οι συγγραφείς δείχνουν ότι το παιχνίδι K-NN έχει ΤΑ με τιμή το πολύ σταθερά για $\alpha > 1$, όπου η σταθερά βελτιώνεται μαζί με α. Τα κοινωνικά αποτελέσματα γίνονται καλύτερα όταν οι κόμβοι είναι «στενόμυαλοι», δηλ. δίνουν μεγαλύτερο βάρος στις απόψεις τους ($\alpha \to \infty$). Σε αντίθεση με τους Bindel et al., οι συγγραφείς δείχνουν ότι εάν οι κόμβοι μπορούν να επιλέξουν τους γείτονές τους με βάση τους k πλησιέστερους γείτονές τους, το ΤΑ μπορεί να φραχθεί. Τέλος, οι συγγραφείς δείχνουν ότι για μικρά α το ΤΑ είναι τουλάχιστον $1/\alpha^2$, γεγονός που εξηγεί γιατί το ΤΑ επιδεινώνεται όταν οι πράκτορες είναι πιο «ευφυείς». Η σύνδεση με το έργο μας είναι η διαμόρφωση κόστους ως αρνητικός λογάριθμος της συνάρτησης πιθανοφάνειας (βλ. την ακόλουθη Ενότητα και το Κεφ. 7).

Γενετικό Μοντέλο

Ομοφιλία

Οι ομοφιλικές ιδιότητες στα ΚΔ έχουν μεγάλη ιστορία: οι πρώτες αναφορές για την ομοφιλία συναντάται στο Συμπόσιο του Πλάτωνα με την φράση *'όμοιος ομοίω αεί πελάζει'*, δηλαδή ότι οι όμοιοι ταιριάζουν μεταξύ τους. Εξάλλου, η λέξη *ομοφιλία* προέρχεται από τις λέξεις *ομού* και *φιλία*. Η μοντελοποίηση των ομοφιλικών διαδικασιών στα ΚΔ γίνεται συνήθως μέσα των χώρων Blau [73, 72] οι οποίοι είναι πολυδιάστατα συστήματα συντεταγμένων όπου η κάθε συντεταγμένη αναφέρεται σε μια κοινωνικο-δημογραφική μεταβλητή, όπως το φύλο, η ηλικιακή ομάδα, το μορφωτικό επίπεδο, το εισόδημα κ.ά.. Η οργανωτική δύναμη σε έναν χώρο

Blau είναι η ομοφιλία σύμφωνα με την οποία η ροή της πληροφορίας από άτομο σε άτομο είναι μια φθίνουσα συνάρτηση της απόστασης μεταξύ των θέσεων των ατόμων στον κοινωνικο-δημογραφικό χώρο. Άτομα με μεγάλη απόσταση είναι σχεδόν απίθανο να αλληλεπιδρούν ενώ ταυτόχρονα οι ομόφιλοι συμμετέχοντες σχηματίζουν μεταξύ τους κοινότητες. Μαθηματικά, αν οι χρήστες u, v έχουν αναπαραστάσεις $\boldsymbol{x}_u,\ \boldsymbol{x}_v$ αντίστοιχα τότε η αρχή της ομοφιλίας υποδυκνύει ότι όταν η απόσταση $\|\boldsymbol{x}_u - \boldsymbol{x}_v\|$ είναι μικρή τότε η πιθανότητα να είναι φίλοι οι u, v είναι μεγάλη.

Η άμεση σύνδεση που μπορεί να κάνει κάποιος για την ομοφιλία είναι η περίπτωση των πλησιέστερων γειτόνων. Πιο συγκεκριμένα, για να ποσοτικοποιήσουμε την ομοφιλία σε ένα ΚΔ εισάγουμε την έννοια του ομοφιλικού δείκτη (ΟΔ). Για ένα ΚΔ $G(V, E)$ θεωρούμε ότι ο κάθε κόμβος $u \in V$ έχει ένα διάνυσμα (δυαδικών στην προκειμένη περίπτωση) χαρακτηριστικών $\boldsymbol{x}_u \in [0, 1]^d$[2]. Θεωρούμε δύο γειτονιές. Η πρώτη γειτονιά αφορά την παρατηρήσιμη γειτονιά του κόμβου $N(u) \cup \{u\}$ εντός του ΚΔ που περιλαμβάνει και τον κόμβο u. Στην περίπτωση του κατευθυνόμενου δικτύου αναφερόμαστε στην έξω-γειτονιά του κόμβου u. Η δεύτερη γειτονιά αφορά τους k_u πλησιέστερους γείτονες του κόμβου u, συμπεριλαμβανομένου και του u. Σκοπός μας είναι να μετρήσουμε ποσοτικά τη διαφορά των συναθροισμένων ενδιαφερόντων $\boldsymbol{\alpha}_u$ και $\boldsymbol{\beta}_u$ αντίστοιχα υπό μια συνάρτηση συνάθροισης $f : [0, 1]^d \to [0, 1]^d$. Δεδομένων των συναθροισμένων (με τον ίδιο τρόπο) χαρακτηριστικών $\boldsymbol{\alpha}_u$ και $\boldsymbol{\beta}_u$ μετράμε το Τετραγωνικό Σφάλμα (ΜΤΣ) μεταξύ των $\boldsymbol{\alpha}_u, \boldsymbol{\beta}_u$

$$\mathrm{RMSE}(\boldsymbol{\alpha}_u, \boldsymbol{\beta}_u) = \sqrt{\frac{\sum_{i=1}^{d}(\alpha_{ui} - \beta_{ui})^2}{d}} \tag{6}$$

Το οποίο έχει μια τιμή εντός του διαστήματος $[0, 1]$. Εν συνεχεία λαμβάνουμε τον σταθμισμένο μέσο όρο για όλους τους κόβους και μετράμε την απόστασή του από το 100%

$$\mathrm{HI} = \left(1 - \frac{\sum_{u \in V} w_u \mathrm{RMSE}(\boldsymbol{\alpha}_u, \boldsymbol{\beta}_u)}{\sum_{u \in V} w_u}\right) \times 100\% \tag{7}$$

Αν θέλουμε να εισάγουμε την έννοια της ισχύος μέσα στην μετρική μας μπορούμε να λάβουμε τον σταθμισμένο μέσο με βάρη τους έξω βαθμούς $w_u = 1/(1 + |\mathrm{out}(u)|)$, ενώ αν θέλουμε να έχοτυμε μια ομοιόμορφη θεώρηση μπορούμε να θέσουμε όλα τα w_u ίσα μεταξύ τους. Διαισθητικά η παραπάνω μετρική μας δείχνει ποσοτικά το πόσο μοιάζει μια παρατηρήσιμη γειτονιά του γράφου μέσω των ακμών του με την γειτονιά των k_u πλησιέστερων γειτόνων του. Μια μεγάλη ομοιότητα μεταξύ αυτών των χαρακτηριστικών θα υποδήλωνε ότι μπορούμε ενδεχομένως να βασιστούμε πάνω στην δομή των πλησιέστερων γειτόνων, αγνοώντας την παρατηρούμενη δομή του γράφου, επιταχύνοντας έτσι αρκετά τους αλγορίθμους που επιδρούν πάνω στο ΚΔ. Πειραματιζόμαστε με κοινωνικά δίκτυα διαφόρων μεγεθών (από $\sim 10^3$ κόμβους ώς $\sim 10^6$ κόμβους) με δυαδικά χαρακτηριστικά και μετράμε τον ΟΔ (αγγλ. HI). Για τον γρήγορο υπολογισμό χρησμοποιούμε μείωση διάστασης με Ανάλυση Κύριων Συνιστωσών (ΑΚΣ)

[2]Στα πειράματά μας τα διανύσματα είναι απλώς δυαδικά, αλλά η θεώρηση δουλεύει και για πραγματικά διανύσματα, όπου το κάθε στοιχείο υποδηλώνει την πιθανότητα ο χρήστης u να έχει την εκάστοτε ιδιότητα.

[13, 98] εξηγώντας το 95% της διασποράς των δεδομένων[3]. Τα αποτελέσματα παρατίθενται στον Πίνακα 7.1.

Παρατηρούμε πολύ υψηλά ποσοστά ομοφιλίας στα διάφορα δίκτυα που εξετάζουμε τα οποία μας παρακινούν να εισάγουμε ένα μοντέλο πάνω στο οποίο οι απόψεις των χρηστών δίνονται με βάση τους κοντινότερούς τους γείτονες.

Γενετικό Μοντέλο Επιρροής από τους k Κοντινότερους Γείτονες

Σε αυτήν την ενότητα ορίζουμε το ΓΜ κάτω από το οποίο δημιουργούνται οι απόψεις. Στο πλαίσιο μας θα υποθέσουμε ότι οι χρήστες διαμορφώνονται από διανύσματα με στοιχεία είτε μηδέν είτε ένα, καθένα από τα οποία εκφράζει εάν ο χρήστης έχει υιοθετήσει την αντίστοιχη γνώμη ή όχι. Για παράδειγμα, ένας χρήστης μπορεί να έχει την τάση να υποστηρίζει 'μπάσκετ', αλλά δεν υποστηρίζει 'ποδόσφαιρο'. Ένας τέτοιος χρήστης έχει ένα διάνυσμα απόψεων ίσο με $(1, 0)$. Υποθέτουμε ότι κάθε χρήστης έχει d απόψεις που διαμορφώνονται με *δοκιμές Bernoulli* που είναι ανεξάρτητες σε κάθε διάσταση. Σε κάθε επανάληψη, η πιθανότητα κάποιος να υοθετήσει μια συγκεκριμένη αλληλουχία ενδιαφερόντων/τάσεων είναι ο μέσος όρος των απόψεων των ΚΚΓ. Η διαδικασία επαναλαμβάνεται έως ότου φτάσουμε σε κατάσταση συναίνεσης. Φορμαλιστικά, έστω U να είναι το σύνολο των πρακτόρων και έστω $\boldsymbol{X}_v^{(t)}$ να είναι το διάνυσμα γνώμης του χρήστη v τη στιγμή t. Έστω επίσης $\boldsymbol{X}_S(t)$ να υποδηλώσει το συνολικό πίνακα απόψεων για ένα υποσύνολο $S \subseteq U$ όπου υποτίθεται ότι οι χρήστες είναι διατεταγμένοι σε μια (αυθαίρετη) σειρά. Υποδηλώνουμε το διάνυσμα πιθανότητας του $\boldsymbol{X}_v^{(t)}$ με το διάνυσμα $\boldsymbol{\xi}_u^{(t)}$. Η διαδικασία δημιουργίας δηλώνει ότι κάθε φορά $t \geq 0$ ο χρήστης κατασκευάζει ένα σύνολο $\mathcal{K}^{(t)}(u)$ που περιέχει τους k πλησιέστερους γείτονες u σε σχέση με τις απόψεις του. Στη συνέχεια, στο γύρο $t+1$ κάθε χρήστης περνάει μια διαδικασία εξομάλυνσης (μέσος όρος) και ενημερώνει τα ενδιαφέροντά του σύμφωνα με το $\boldsymbol{Be}\left(\boldsymbol{\xi}_u^{(t+1)}\right)$ τέτοια ώστε

$$\boldsymbol{\xi}_u^{(t+1)} = \frac{1}{k} \sum_{v \in \mathcal{K}^{(t)}(u)} \boldsymbol{X}_u^{(t)} \tag{8}$$

Η επιλογή των γειτόνων γινεται σύμφωνα με την απόσταση Hamming που ορίζεται ως το άθροισμα των συντεταγμένων στις οποίες διαφωνούν οι χρήστες, δηλαδή για δύο χρήστες $u, v \in U$ ορίζεται ως

$$\sum_{i=1}^{d} \mathbf{1}\left\{X_{ui}^{(t)} \neq X_{vi}^{(t)}\right\} = \sum_{i=1}^{d} \left(X_{ui}^{(t)} - X_{vi}^{(t)}\right)^2 \tag{9}$$

Μακροσκοπικά, η πιθανότητα μια άποψη $1 \leq i \leq d$ να είναι 1 μεταξύ των χρηστών του υποσυνόλου $S \subseteq U$ τη χρονική στιγμή $t \geq 0$ μοντελοποιείται από την ποσότητα $\mu_{iS}^{(t)}$, όπου δοσμένων των $\boldsymbol{\mu}_S^{(t)}$

$$\Pr\left[\boldsymbol{X}_S^{(t)} \middle| \boldsymbol{\mu}_S^{(t)}\right] = \prod_{i=1}^{d} \prod_{u \in S} \left(\mu_{iS}^{(t)}\right)^{X_{iu}^{(t)}} \left(1 - \mu_{iS}^{(t)}\right)^{1 - X_{iu}^{(t)}} \tag{10}$$

[3] αγγλ. explained variance

Οι παράμετροι αυτοί μπορούν να υπολογιστούν εκ των υστέρων μετά τον υπολογισμό των παραμέτρων των απόψεων. Για μια παρόμοια αντιμετώπιση παραπέμπουμε στην αναφορά [55, 54].

Προκειμένου να προβούμε σε συμπερασμό παραμέτρων για το μοντέλο αυτό πρέπει να ορίσουμε μια συνάρτηση κόστους που έχει στατιστικό νόημα. Η πιο συχνά χρησιμοποιούμενη συνάρτηση είναι αυτή της πιθανοφάνειας. Επομένως, ορίζουμε το κόστος ως την στιγμιαία πιθανοφάνεια του μοντέλου τη χρονική στιγμή t, ήτοι

$$\mathcal{L}_{\xi}^{(t+1)}\left(\boldsymbol{\xi}_U^{(t+1)}\right) = \log \sum_{\boldsymbol{X}_U^{(t)}} \Pr\left[\boldsymbol{X}_U^{(t)} | \boldsymbol{\xi}_U^{(t+1)}\right] \tag{11}$$

Η συνάρτηση αυτή έχει καταβολές και από τη μοντελοποίηση τέτοιων διαδικασιών υπό την παιγνιοθεωρητική σκοπιά όπου σκοπός του καθενός είναι να ελαχιστοποιήσει το κόστος ασυμφωνίας του. Περισσότερα μπορούν να βρεθούν στην αναφορά [12], η οποία θα μπορούσε, υπό τη δική μας μοντελοποίηση, να είναι ο αρνητικός λογάριθμος πιθανοφάνειας μοντέλου με Γκαουσιανές απόψεις. Ένας άλλος τρόπος να εξηγήσουμε τη μορφή της παραπάνω πιθανοφάνειας είναι η Μαρκοβιανή ιδιότητα, δηλαδή οι απόψεις ενός χρήστη δεν επηρεάζονται παρά μόνον από την προηγούμενη κατάσταση.

Εν γένει, ο ακριβής συμπερασμός παραμέτρων με ΕΜΠ γνωρίζοντας μόνο την αρχική κατάσταση του δικτύου $\boldsymbol{X}_U^{(0)}$ χρειάζεται εν γένει *εκθετικό χρόνο* για να λυθεί. Ο λόγος για τον οποίο συμβαίνει αυτό είναι διότι κανείς πρέπει να αθροίσει πάνω σε όλα τα ενδεχόμενα για τις λανθάνουσες μεταβλητές $\boldsymbol{X}_U^{(t)}$ για $t \geq 1$. Υποθέτοντας μια μεταβολική κατανομή $Q^{(t)}$ που ακολουθούν οι μεταβλητές των απόψεων, μπορούμε να χρησιμοποιήσουμε την ανισότητα Jensen για να εξάγουμε ένα κάτω φράγμα

$$\mathcal{L}_{\xi}^{(t+1)} \geq \mathbb{E}_{Q^{(t)}}\left[\log \Pr\left[\boldsymbol{X}_U^{(t)} \middle| \boldsymbol{\xi}_U^{(t+1)}\right]\right] + \mathbb{E}_{Q^{(t)}}\left[-\log Q(\boldsymbol{X}_U^{(t)})\right] \tag{12}$$

πάνω στην πιθανοφάνεια (ως προς τις μεταβλητές των απόψεων). Παρόμοιο αποτέλεσμα μπορούμε να λάβουμε και ως προς τις μακροσκοπικές παραμέτρους $\boldsymbol{\mu}_U^{(t)}$. Η ποσότητα

$$\mathcal{L}_{Q,\xi}^{(t+1)} = \mathbb{E}_{Q^{(t)}}\left[\log \Pr\left[\boldsymbol{X}_U^{(t)} \middle| \boldsymbol{\xi}_U^{(t+1)}\right]\right] \tag{13}$$

ονομάζεται Evidence Lower Bound (ELBO) και αποτελεί ένα κάτω φράγμα στην πιθανοφάνεια. Η χρήση αυτής της συνάρτησης ως αντικειμενικής είναι ευρεία στα προβλήματα στατιστικής μοντελοποίησης [54, 55, 50, 101], λόγω της ευκολίας στην βελτιστοποίησή της. Η υπολοιπόμενη ποσότητα

$$\mathcal{H}_{Q,\xi}^{(t+1)} = \mathbb{E}_{Q^{(t)}}\left[-\log Q(\boldsymbol{X}_U^{(t)})\right] \geq 0 \tag{14}$$

είναι η εντροπία της μεταβολικής κατανομής. Πολλές επιλογές έχουν προταθεί για τη μεταβολική κατανομή. Η πιο συγγενής στο πρόβλημά μας αποτελεί η *προσέγγιση μέσου πεδίου* [50, 101], ορμώμενη από τη στατιστική φυσική.[4] Στην προσέγγιση αυτή, θεωρούμε ανεξάρτητες και ισόνομες μεταβλητές, ήτοι

[4] Η παραπάνω πρακτική είναι ευρέως γνωστή και ως Expectation Maximization (EM) [26]

$$Q^{(t)} = \prod_{u \in U} \prod_{i=1}^{d} \left(\phi_{iu}^{(t)}\right)^{X_{iu}^{(t)}} \left(1 - \phi_{iu}^{(t)}\right)^{1 - X_{iu}^{(t)}} \tag{15}$$

Το φράγμα ELBO της Εξ. 13 μπορεί να γραφτεί ως

$$\mathcal{L}_{Q,\xi}^{(t+1)} = \mathbb{E}_{Q^{(t)}} \left[\sum_{i=1}^{d} \sum_{u \in U} \sum_{v \in S} \mathbf{1}\left\{ v \in \mathcal{K}^{(t)}(u) \right\} \left(X_{iv}^{(t)} \log \xi_{iu}^{(t+1)} + \left(1 - X_{iv}^{(t)}\right) \log\left(1 - \xi_{iu}^{(t+1)}\right) \right) \right] \tag{16}$$

Σε αυτό το σημείο καλούμαστε να προσεγγίσουμε την ποσότητα

$$\mathbf{1}\left\{ v \in \mathcal{K}^{(t)}(u) \right\} \left(X_{iv}^{(t)} \log \xi_{iu}^{(t+1)} + \left(1 - X_{iv}^{(t)}\right) \log\left(1 - \xi_{iu}^{(t+1)}\right) \right) \tag{17}$$

Την οποία θα προσεγγίσουμε με χρήση φραγμάτων Chernoff. Πιο συγκεκριμένα, θα δείξουμε ότι για αρκούντως μεγάλο $|U| = n$ και κατάλληλα φραγμένο k το σύνολο $\mathcal{K}^{(t)}(u)$ προσεγγίζεται από το σύνολο $K^{(t)}(u)$ που περιέχει τους ΚΚΓ ως προς το χώρο των παραμέτρων (κάθε τυχαίο διάνυσμα αναπαρίσταται με το διάνυσμα των παραμέτρων του) για κάθε χρήστη u. Πιο συγκεκριμένα αποδεικνύουμε το παρακάτω θεώρημα για την απόσταση Hamming, το οποίο γενικεύεται και για άλλες νόρμες με τις κατάλληλες αλλαγές. Αρχικά αποδεικνύουμε το εξής ενδιάμεσσο Λήμμα

Λήμμα 1. *Έστω $\boldsymbol{X}, \boldsymbol{Y} \in \{0,1\}^d$ δύο διανύσματα με ανεξάρτητες συντεταγμένες Bernoulli με διανύσματα παραμέτρων $\boldsymbol{p} = \mathbb{E}[\boldsymbol{X}]$ και $\boldsymbol{q} = \mathbb{E}[\boldsymbol{Y}]$ αντίστοιχα, δηλαδή $X_i \perp\!\!\!\perp X_j$, $Y_i \perp\!\!\!\perp Y_j$ για κάθε $i \neq j$ και $X_i \perp\!\!\!\perp Y_j$ για κάθε $1 \leq i, j \leq d$ και $\|\boldsymbol{X} - \boldsymbol{Y}\|$ η απόσταση Hamming μεταξύ τους*

1. *Για κάθε $\eta > \eta_0$ όπου $\eta_0 = \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{Y}\|] - \|\boldsymbol{p} - \boldsymbol{q}\| \geq 0$, ισχύει η ανισότητα συγκέντρωσης*

$$\Pr[|\|\boldsymbol{X} - \boldsymbol{Y}\| - \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{Y}\|]| \geq \eta] \leq 2\exp\left(-\frac{2(\eta - \eta_0)^2}{d}\right) \tag{18}$$

2. *Για κάθε $\epsilon > 0$ ισχύει η ανισότητα συγκέντρωσης (χειρότερης περίπτωσης)*

$$\Pr\left[|\|\boldsymbol{X} - \boldsymbol{Y}\| - \|\boldsymbol{p} - \boldsymbol{q}\|| > \frac{(1+\epsilon)d}{2}\right] \leq 2\exp\left(-\frac{\epsilon^2 d}{2}\right) \tag{19}$$

3. *Δοθέντος $\delta \in [0,1]$ και για κάθε $\epsilon > 0$ πρέπει να διαλέξουμε διάσταση*

$$d = \Omega\left(\frac{\log(2/\delta)}{\epsilon^2}\right) \tag{20}$$

έτσι ώστε

$$\Pr\left[|\|\boldsymbol{X} - \boldsymbol{Y}\| - \|\boldsymbol{p} - \boldsymbol{q}\|| \leq \frac{(1+\epsilon)d}{2}\right] \geq 1 - \delta \tag{21}$$

Η ανάλυση μας βασίζεται στην ανισότητα φραγμένων διαφορών [30] και στη γενικώτερη περίπτωση στην ανισότητα του Talagrand και τη χρήση της τριγωνικής ανισότητας. Το Λήμμα μας εγγυάται ότι στη χειρότερη περίπτωση δεν θα δούμε σχεδόν ποτέ το σφάλμα να ξεπερνά το $\frac{(1+\epsilon)d}{2}$. Η χειρότερη περίπτωση είναι αυτή στην οποία όλα είναι πυκνά και ισοπίθανα, δηλαδή $p_i = q_i = 1/2$ για κάθε $1 \leq i \leq d$. Προφανώς, σε αιραιότερες διατάξεις, όπως αυτές που ισχύουν στην πραγματικότητα το φράγμα είναι αισθητά καλύτερο και εξαρτάται από την ενεργό διάσταση (συντεταγμένες με αρκούντως μεγάλη μάζα). Η συμπεριφορά της εκλέπτυνσης εξαρτάται από την ποσότητα η_0 που ισούται με

$$\eta_0 = \sum_{i=1}^{d}(p_i(1-q_i) + q_i(1-p_i)) - \sum_{i=1}^{d}(p_i - q_i)^2 \tag{22}$$

Στην περίπτωση που έχουμε ένα σύνολο διανυσμάτων $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ και πραγματοποιήσουμε $\binom{m}{2}$ ανεξάρτητες δοκιμές όπως παραπάνω, η απαιτούμενη διάσταση μειώνεται σε

$$d = \Omega\left(\frac{\log(2/\delta)}{\epsilon^2 m^2}\right) \tag{23}$$

Στη συνέχεια, οπλισμένοι με το προηγούμενο Λήμμα, διερευνουμε την συμπεριφορά της συμμετρικής διαφοράς $\mathcal{K} \ominus K = (\mathcal{K} \setminus K) \cup (K \setminus \mathcal{K})$, όπου αγνοούμε τους δείκτες t, u χάριν ευκολίας. Συγκεκριμένα αποδεικνύουμε το εξής θεώρημα

Θεώρημα 1. *Έστω $\epsilon > 0$ κάποιος πραγματικός αριθμός και έστω $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \{0,1\}^d$ διανύσματα από ανεξάρτητες δοκιμές Bernoulli. Διαλέγοντας*

$$k \leq C\left(4n\exp\left(-\epsilon^2 d\right) + \log n\right)$$

γείτονες, για κάποιο $C > 1$ έχουμε ότι $\lim_{n\to\infty} \Pr_k[|\mathcal{K} \ominus K| \geq 1] = 0$.

Η ανάλυση για την απόδειξη αυτού του Θεωρήματος χωρίζεται σε δύο μέρη. Το πρώτο μέρος αφορά στο φράξιμο της πιθανότητας κάποιου adversary στο να επηρεάσει τους ΚΚΓ και το δεύτερο σκέλος ομοιάζει με ένα πρόβλημα μπαλλών και κάδων, κλασικής μεθόδου στην ανάλυση πιθανοτικών αλγορίθμων. Με βάση τα παραπάνω και με το γεγονός ότι 'σχεδόν βεβαίως' το σύνολο των τυχαίων ΚΚΓ προσεγγίζει το σύνολο των ΚΚΓ στον χώρο των παραμέτρων (κατά αναμενόμενη τιμή). Με βάση τα παραπάνω μπορούμε να απλοποιήσουμε την Εξ. 13 στην μορφή

$$\mathcal{L}_{Q,\xi}^{(t+1)} \approx \sum_{i=1}^{d}\sum_{u\in U}\sum_{v\in K^{(t)}(u)}\left[\phi_{iv}^{(t)}\log\phi_{iu}^{(t+1)} + \left(1-\phi_{iv}^{(t)}\right)\log\left(1-\phi_{iu}^{(t+1)}\right)\right] \tag{24}$$

Λαμβάνοντας τις μερικές παραγώγους ως προς τις μεταβολικές παραμέτρους

$$\frac{\partial\mathcal{L}_{Q,\xi}^{(t+1)}}{\partial\phi_{iu}^{(t+1)}} = 0 \tag{25}$$

καταλήγουμε στο σύνολο των επαναληπτικών εξισώσεων

$$\phi_{iu}^{(t+1)} = \frac{1}{k} \sum_{v \in K^{(t)}(u)} \phi_{iv}^{(t)} \tag{26}$$

για κάθε $1 \leq i \leq d$ και για κάθε $u \in U$. Οι εξισώσεις αυτές προσομοιάζουν με τα κλασικά μοντέλα ΔΔΑ που έχουν ντετερμινισμό. Με άλλα λόγια, οι εξισώσεις συμπερασμού στα στοχαστικά μοντέλα συμπίπτουν με τις εξισώσεις των ντετερμινιστικών. Όσον αφορά τις μακροσκοπικές παραμέτρους εύκολα διαπιστώνει κανείς ότι για κάθε υποσύνολο $S \subseteq U$ έχουμε ότι

$$\begin{aligned} \mathcal{L}_{Q,\mu}^{(t)} = \\ \mathbb{E}_{Q^{(t)}} \left[\sum_{u \in S} \sum_{i=1}^{d} \left(X_{iu}^{(t)} \log \mu_{iS}^{(t)} + \left(1 - X_{iu}^{(t)}\right) \log \left(1 - \mu_{iS}^{(t)}\right) \right) \right] = \\ \sum_{u \in S} \sum_{i=1}^{d} \left(\phi_{iu}^{(t)} \log \mu_{iS}^{(t)} + \left(1 - \phi_{iu}^{(t)}\right) \log \left(1 - \mu_{iS}^{(t)}\right) \right) \end{aligned} \tag{27}$$

Λαμβάνοντας πάλι παραγώγους έχουμε ότι

$$\frac{\partial \mathcal{L}_{Q,\mu}^{(t+1)}}{\partial \mu_{iS}^{(t+1)}} = 0 \tag{28}$$

Ισοδύναμα

$$\boldsymbol{\mu}_S^{(t)} = \frac{1}{|S|} \sum_{v \in S} \boldsymbol{\phi}_v^{(t)} \tag{29}$$

Γενίκευση. Η προσέγγιση αυτή είναι πολύ ενδιαφέρουσα και μπορεί να επεκταθεί και σε άλλου είδους κατανομές. Για παράδειγμα αν οι απόψεις ακολουθούν Γκαουσιανή κατανομή με άγνωστη μέση τιμή $\boldsymbol{\xi}_u^{(t)}$ και γνωστό πίνακα διασποράς Σ τότε κατάντιστοιχία

$$\mathcal{L}_Q^{(t+1)} = \mathbb{E}_{Q^{(t)}} \left[\sum_{u \in U} \sum_{v \in \mathcal{K}^{(t)}(u)} \mathbf{1}\{v \in \mathcal{K}^{(t)}(u)\} \left[-\frac{1}{2} \log\left((2\pi)^d |\Sigma|\right) - \frac{1}{2} \left(\boldsymbol{X}_v^{(t)} - \boldsymbol{\xi}_u^{(t+1)}\right)^T \Sigma^{-1} \left(\boldsymbol{X}_v^{(t)} - \boldsymbol{\xi}_u^{(t+1)}\right) \right] \right] \tag{30}$$

Ενώ λαμβάνοντας τις μέσες τιμές κάποιος μπορεί να καταλήξει με τον ίδιο τρόπο στην Εξ. 26 και στην Εξ. 29. Αν δε κάποιος θέλει να συμπεράνει τη μακροσκοπική $\Sigma_S^{(t)}$ ενός υποσυνόλου $S \subseteq U$, μπορεί να καταλήξει με τον ίδιο ακριβώς τρόπο στην εξίσωση

$$\Sigma_S^{(t)} = \frac{1}{|S|} \sum_{u \in S} \left(\boldsymbol{\phi}_u^{(t)} - \boldsymbol{\mu}_S^{(t)}\right) \left(\boldsymbol{\phi}_u^{(t)} - \boldsymbol{\mu}_S^{(t)}\right)^T \tag{31}$$

Είναι ενδιαφέρον — και ανοικτό — το πως γενικεύεται η παραπάνω μεθοδολογία σε εκθετικές οικογένειες και με ποιες εγγυήσεις.

Κανονικοποίηση. Για να προσθέσουμε κανονικοποίηση στο μοντέλο ώστε να μην κάνει overfit μπορούμε να εισάγουμε πλασματικές απόψεις μέσω των συναρτήσεων κανονικοποίησης $\omega^{(t)}$ όπου, π.χ. αν προσθέσουμε τις αρχικές απόψεις έχουμε ότι

$$\omega^{(t)} = \alpha \sum_{u \in U} \sum_{i=1}^{d} \left[\phi_{iu}^{(0)} \log \phi_{iu}^{(t)} + \left(1 - \phi_{iu}^{(0)}\right) \log \left(1 - \phi_{iu}^{(t)}\right) \right] \tag{32}$$

Θέτοντας τις παραγώγους ίσες με το 0 έχουμε ότι για τη συνολική πιθανοφάνεια ειπεισέρχεται ο όρος της κανονικοποίησης

$$\frac{\partial \mathcal{L}_{Q,\xi}^{(t+1)}}{\partial \phi_{iu}^{(t+1)}} + \alpha \frac{\partial \omega^{(t+1)}}{\partial \phi_{iu}^{(t+1)}} = 0 \tag{33}$$

ο οποίος δίνει τις επαυξημένες εξισώσεις

$$\boldsymbol{\phi}_u^{(t+1)} = \frac{1}{k+\alpha} \sum_{v \in K^{(t)}(u)} \boldsymbol{\phi}_v^{(t)} + \frac{\alpha}{k+\alpha} \boldsymbol{\phi}_u^{(0)} \tag{34}$$

Ως sanity check μπορούμε να δούμε ότι για $\alpha = 1$ προσθέτουμε μια άποψη και πάμε σε $k+1$ γείτονες από k. Τέλος σε μια γενικώτερη περίπτωση που προσθέτουμε R πλασματικές απόψεις $\boldsymbol{\psi}_{ru}$ με βάρη α_r όπου $1 \leq r \leq R$ έχουμε ότι

$$\boldsymbol{\phi}_u^{(t+1)} = \frac{1}{k+A} \left(\sum_{v \in K^{(t)}(u)} \boldsymbol{\phi}_v^{(t)} + \sum_{r=1}^{R} \alpha_r \boldsymbol{\psi}_{ru} \right) \tag{35}$$

$$A = \sum_{r=1}^{R} \alpha_r \tag{36}$$

Παρατηρήστε ότι η παραπάνω εξίσωση αντιστοιχεί σε σταθμισμένο μέσο όρο με $k+R$ βάρη, όπου τα πρώτα k βάρη είναι ίσα με 1 και τα υπόλοιπα είναι ίσα με α_r. Η μορφή της κανονικοποίησης αυτής είναι λογική για μεταβλητές Bernoulli. Αντίστοιχα, σε Γκαουσιανές μεταβλητές, τα ίδια αποτελέσματα θα λάβουμε με L2 Regularization. Ενώ εν γένει το μοντέλο της Εξ. 26 συγκλίνει — όπως θα αποδείξουμε αναλυτικά στην εργασία μας — η σύγκλιση του κανονικοποιημένου μοντέλου αποτελεί ανοικτό πρόβλημα.

Υλοποίηση. Η υλοποίση μπορεί να πραγματοποιηθεί χρησιμοποιώντας τις διάφορες διαθέσιμες δομές για την εύρεση ΚΚΓ. Οι επιμέρους πολυπλοκότητες παρατίθενται στον Πίνακα 7.2. Στην υλοποίησή μας έχουμε χρησιμοποιήσει Locality Sensitive Hashing, και μπορούμε σε ένα απλό υπολογιστικό σύστημα 2 πυρήνων και 72GB μνήμης RAM να τρέξουμε το μοντέλο σε εκατομμύρια κόμβους σε χρόνους της τάξης των 10^2 δευτερολέπτων.

Θεωρητικές Ιδιότητες

Σύγκλιση σε Πεπερασμένο Χρόνο. Μπορούμε να δείξουμε με χρήση Θεωρίας Αυτομάτου Ελέγχου (βλ. Κεφ. 5) ότι οι Εξ. 26 συγκλίνουν σε πεπερασμένο χρόνο. Αρχικά,

δείχνουμε ότι το συστημά μας συγκλίνει εν γένει σε άπειρο χρόνο με χρήση μιας συνάρτησης δυναμικού (συνάρτηση Lyapunov) η οποία δείχνουμε ότι είναι αρνητικά ορισμένη και στη συνέχεια εκμεταλευόμενοι το όριό της καθώς και την ιδιότητα ότι όταν δύο σύνολα $W, Z \subseteq U$ διαχωρίζονται δεν ξανασυναντιούνται σε μεγαλύτερους χρόνους καταλήγουμε στο συμπέρασμα ότι η σύγκλιση γίνεται σε πεπερασμένο χρόνο.

Πιο συγκεκριμένα, η επιλογή του δυναμικού γίνεται με τη βοήθεια της συζυγούς ακολουθίας [81] $\Pi(t) = \Pi_U^{(t)}$ της ακολουθίας $\Phi(t) = \Phi_U^{(t)}$ όπου

$$\Phi(t+1) = A(t)\Phi(t) \tag{37}$$

όπου ο $A(t)$ είναι γραμμο-στοχαστικός πίνακας που έχει τιμές $1/k$ στις θέσεις των ΚΚΓ και 0 αλλού. Η ακολουθία $\Pi(t)$ έχει συντεταγμένες της μορφής $\pi_u^{(t)} > p$ όπου $p \in (0,1)$ για κάθε $u \in U$ την οποία έχουμε ότι έχει το ίδιο σύνολο εξισώσεων στο χώρο γραμμών, ήτοι

$$\Pi^T(t+1) = \Pi^T(t)A(t) \tag{38}$$

Η συνάρτηση Lyapunov που ορίζουμε είναι η

$$V(t) = \sum_{i=1}^{n} \pi_u(t) \|\phi_u(t) - \Pi^T(t)\Phi(t)\|_2^2 \tag{39}$$

Για την οποία δείχνουμε ότι

$$V(t) = V(t+1) + \frac{1}{2}\sum_{u,v} H_{uv}(t)(\phi_u^{(t)} - \phi_v^{(t)})^2 \tag{40}$$

$$H(t) = A^T(t)\mathrm{diag}(\pi_u(t+1))A(t) \tag{41}$$

Τα στοιχεία $H_{uv}(t)$ του πίνακα $H(t)$ είναι

$$H_{uv}(t) = \frac{1}{k^2}\sum_{w} \pi_w(t+1)\mathbf{1}\{u \in K^{(t)}w\}\mathbf{1}\{v \in K^{(t)}(w)\} \tag{42}$$

Επομένως εκτός του σημείου ισοροπίας η συνάρτηση $V(t)$ είναι αρνητικά ορισμένη, ήτοι

$$V(t+1) = V(t) - \frac{1}{2k^2}\sum_{w} \pi_w(t+1) \sum_{u,v \in K^{(t)}(w)} (\phi_u^{(t)} - \phi_v^{(t)})^2 < V(t) \tag{43}$$

Επίσης, για δύο συνεχόμενα σύνολα χρηστών $W, Z \subseteq U$ σε μια διάσταση ορίζουμε την μετρική

$$\delta_{WZ}^{(t)} = \min_{w \in W, z \in Z} \|\phi_w^{(t)} - \phi_z^{(t)}\| \tag{44}$$

Στη συνέχεια χρησιμοποιώντας το γεγονός ότι $\lim_{t\to\infty} V(t) = 0$ καταλήγουμε στο συμπέρασμα ότι άπαξ και δυο σύνολα χωριστούν τη στιγμή t_0 δηλαδή η απόσταση των κοντινότερων γειτόνων των εγγύτερων σημείων $w^* \in W$ και $z^* \in Z$ είναι μικρότερη από $\delta_{WZ}^{(t_0)}$, τότε για όλους τους ύστερους χρόνους $t \geq t_0$ παραμένουν χωριστά. Επομένως οι χρήστες δημιουργούν

συστάδες σε πεπερασμένο χρόνο. Η πλήρης απόδειξη των θεωρημάτων βρίσκεται στο Κεφ. 7. Παραθέτουμε την τελική διατύπωση

Θεώρημα 2 (Σύγκλιση)**.** *Το σύστημα των Εξ. 26 συγκλίνει σε πεπερασμένο χρόνο.*

Ρυθμός Σύγκλισης. Ο ρυθμός σύγκλισης μπορεί να προσδιοριστεί χρησιμοποιώντας θεωρία Μαρκοβιανών Αλυσίδων. Αρχικά, δίνουμε την έννοια της Απόστασης Ολικής Μεταβολής (AOM) για δύο μέτρα πιθανότητας μ, ν και μιας σ-άλγεβρας $\mathcal{F}$ πάνω σε ένα δειγματικό χώρο Ω (αριθμήσιμο) η οποία είναι ίση με

$$d_{TV}(\mu, \nu) = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)| \tag{45}$$

Για μια Μαρκοβιανή Αλυσίδα που επιδέχεται στάσιμη κατανομή π[5] με χώρο καταστάσεων Ω με n καταστάσεις, πίνακα μετάβασης P, και διάνυσμα κατάστασης $\pi(t)$, όπου με $\pi(A, t)$ συμβολίζουμε το

$$\pi(A, t) = \sum_{x \in A} \pi_x(t) \qquad A \subseteq [n] \tag{46}$$

ορίζουμε το χρόνο μείξης ως τον ελάχιστο χρόνο $t_0(D)$ τέτοιος ώστε η AOM μεταξύ της στάσιμης κατανομής π και της $\pi(t)$ είναι μικρότερη η ίση από D, δηλαδή

$$t_0(D) = \inf\{t \geq 0 \mid d_{TV}(\pi(t), \pi) \leq D\} \tag{47}$$

Γνωρίζουμε από το Θεώρημα Perron-Frobenius ότι η AOM τη χρονική στιγμή t κυριαρχείται από τη δεύτερη μεγαλύτερη ιδιοτιμή $|\lambda_2| < 1$ στην t-οστή δύναμη, δηλαδή

$$d_{TV}(\pi(t), \pi) = O(|\lambda_2|^t) \tag{48}$$

Για να εφαρμόσουμε τον κανόνα αυτό στη δική μας περίπτωση πρέπει να βρούμε τον πιο ¨αργό' πίνακα μετάβασης, ο οποίος και θα επηρεάσει τη σύγκλιση. Χρησιμοποιώντας το αποτέλεσμα της εικασίας του Alon που αποδείχθηκε πολύ καιρό μετά από τον Friedman[6] καταλήγουμε στο εξής θεώρημα

Θεώρημα 3. *Η AOM των Εξ. 26 μειώνεται σαν $o(k^{-t/2})$ για $k \geq 2$ και μεγάλο n. Πιο συγκεκριμένα, για κάποιο $\delta \in [0, 1]$ αρκετά μικρό και $n = \Omega\left(\delta^{-1/\tau}\right)$ χρήστες όπου $\tau = \lceil(\sqrt{k-1} + 1)/2\rceil - 1$ τότε με πιθανότητα τουλάχιστον $1 - \delta$ τη AOM μειώνεται σαν $o(k^{-t/2})$*

Η πλήρης απόδειξη βρίσκεται στο Κεφ. 7.

[5] Μια ειδική διανομή για μια αλυσίδα Μαρκο έτσι ώστε εάν η αλυσίδα ξεκινά με τη στάσιμη κατανομή της, η οριακή κατανομή όλων των καταστάσεων ανά πάσα στιγμή θα είναι πάντα η στάσιμη κατανομή. Υποθέτοντας ότι δεν υπάρχει δυνατότητα μείωσης, η σταθερή κατανομή είναι πάντα μοναδική αν υπάρχει και η ύπαρξή της μπορεί να υπονοείται από τη θετική επανάληψη όλων των καταστάσεων. Η στάσιμη κατανομή έχει την ερμηνεία της περιοριστικής κατανομής όταν η αλυσίδα είναι εργοδική.

[6] Περισσότερες λεπτομέρειες μπορούν να διαβαστούν στο Κεφ. 5

Πρόβλεψη Ενδιαφερόντων σε ΔΚΔ από 'Δίασημους' Χρήστες σε Δίκτυα με Δομή Πυρήνα-Περιφέρειας

Σε αυτήν την ενότητα, θα παρουσιάσουμε μια εφαρμογή του μοντέλου μας. Πιο συγκεκριμένα, στόχος μας είναι να προβλέψουμε τα ενδιαφερόντα των χρηστών σε ένα ΔΚΔ. Έχουμε ένα κοινωνικό δίκτυο $G(V, E)$ με $|V| = n$ κόμβοι και $|E| = m$ εδγες όπου κάθε κόμβος σε ένα υποσύνολο C V έχει d - διαστατικό δυαδικό διάνυσμα που δείχνει εάν ο χρήστης c $in C$ εμπλέκεται σε ένα συγκεκριμένο ενδιαφέρον $1 \leq i \leq d$ ή όχι. Για παράδειγμα, σε ένα δίκτυο συν-συγγραφέων τα χαρακτηριστικά αντιπροσωπεύουν τα συνέδρια στα οποία ένας ερευνητής έχει δημοσιεύσει και οι ακμές αντιπροσωπεύουν συν-συγγραφείς και στοχεύουμε να προτείνουμε στους ερευνητές πού να δημοσιεύσουν τις εργασίες τους.

Δομή Πυρήνα-Περιφέρειας. Η δομή πολλών κοινωνικών δικτύων μπορεί να αποσυντεθεί σε ένα σχετικά μικρό *πυρήνα* [48, 114, 88, 115, 99, 82, 109] C με κόμβους και μια *περιφέρεια* P, όπου οι περιφερειακοί κόμβοι είναι αραιά συνδεδεμένοι μεταξύ τους, αλλά σχετικά καλά συνδεδεμένοι με τους πυρηνικούς χρήστες. Όπως παρατηρούμε πειραματικά, ένα κλάσμα των κόμβων-πυρήνων κυριαρχεί σχεδόν στο υπόλοιπο δίκτυο, δηλαδή ένα κλάσμα $\delta(n)n$ των κόμβων είναι υπεύθυνο για την κυριαρχία του $(1 - a(n))n$ των περιφερειακών 'εμπλεκομένων' κόμβων. Με τον όρο 'εμπλεκόμενος', αναφερόμαστε σε χρήστες που ακολουθούν τουλάχιστον τ άτομα. Στα δεσμευμένα υποδίκτυα, ακόμη και όσο ένα υπογραμμικό κλάσμα κόμβων είναι υπεύθυνοι για την κυριαρχία ενός πολύ υψηλού κλάσματος των κόμβων [16, 17, 18]. Αυτοί οι σημαντικοί κόμβοι που διαθέτουν μεγάλο αριθμό εισερχόμενων συνδέσεων ή *ακόλουθοι* είναι επίσης γνωστοί ως *διασημότητες ή επηρεαστές/διάσημοι του δικτύου*. Η κύρια ιδέα πίσω από αυτό το πείραμα είναι να συγκεντρωθούν αποτελεσματικά οι (επισημασμένοι) παράγοντες επιρροής του δικτύου και, στη συνέχεια, να χρησιμοποιηθούν τα ενδιαφέροντά τους για να μάθουν τα ενδιαφερόντα του υπόλοιπου δικτύου. Τα ενδιαφέροντα μπορούν να εμφανίζονται με φθίνουσα σειρά σε σχέση με τις σχετικές βαθμολογίες τους στον χρήστη ενδιαφέροντος ως μηχανισμός προτάσεων. Η προσέγγιση του προβλήματος της συμπερίληψης των ετικετών χρήστη κάτω από το φακό των επηρεαστών έχει πολλαπλά οφέλη. Πρώτα απ ́όλα, όπως παρατηρούμε αργότερα, ο πυρήνας είναι μικρός αλλά πολύ εκφραστικός σε σχέση με τις πληροφορίες που παρέχει. Επιπλέον, οι επιρροείς τείνουν να εκτίθενται — κυρίως για εμπορικούς λόγους [52, 37, 49, 100, 19] — τις πληροφορίες προφίλ τους (φίλοι και ενδιαφέροντα) δημόσια, έτσι οι πληροφορίες μπορούν να συγκεντρωθούν εύκολα. Χρησιμοποιούμε τους δειγματοληπτικούς κόμβους του δείκτη για να κατασκευάσουμε το διμερές γράφημα που περιέχει τους επηρεαστές μαζί με τους χρήστες που τους ακολουθούν και τις ακμές μεταξύ τους. Ο μέγιστος αριθμός ακμών σε ένα τέτοιο διμερές γράφημα επιρροής είναι σημαντικά μικρότερος από το να πρέπει να κοιτάξετε ολόκληρο το κοινωνικό δίκτυο.

Χρησιμοποιούμε το διμερές γράφημα χρήστη-επηρεαστή για να αρχικοποιήσουμε τις τιμές των βαθμολογιών ενδιαφέροντος των χρηστών ως τον μέσο όρο των ενδιαφερόντων των επηρεαστών που ακολουθούν. Η βασική πρόκληση εδώ είναι πώς να κάνετε τους χρήστες να αλληλεπιδρούν χωρίς να χρειάζεται να δουν τις γειτονιές τους.

Εντοπισμός Διασήμων Χρηστών. Το πρόβλημά μας είναι παρόμοιο με το πρόβλημα της μέγιστης κάλυψης (ΜΚ) στη συνδυαστική βελτιστοποίηση, καθώς ορίζουμε ένα ποσοστόχο και προσπαθούμε να μεγιστοποιήσουμε τους καλυμμένους χρήστες με τον αριθμό των εν λόγω διασημοτήτων. Το πρόβλημα ΜΚ έχει αποδειχθεί ότι είναι NP-Hard [;, 34] και ο άπληστος αλγόριθμος που προχωράει σε γύρους και επιλέγει τον κόμβο με τον μέγιστο αριθμό ακάλυπτων γειτόνων αποδίδει μια βέλτιστη αναλογία προσέγγισης $1 - 1/e$. Η εκτέλεση του άπληστου αλγόριθμου ad-hoc έχει πολύ υψηλό υπολογιστικό κόστος καθώς αυξάνεται ο αριθμός των κόμβων. Για αυτόν τον λόγο, βασίζουμε σε ένα πιρούνι του αρχικού αλγορίθμου που ονομάζουμε Bucketed Greedy Bucketed MC (BGMC). Στον BGMC, έχουμε κόμβους K ως άνω φράγμα που θέλουμε να χρησιμοποιήσουμε στην κάλυψη μας. Ταξινομούμε τους κόμβους σύμφωνα με τους βαθμούς τους και τους βάζουμε σε $\log(n/k)/\log\gamma$ non-uniform buckets $V_1, \ldots, V_r, \ldots$ μεγέθους $\lceil \gamma K \rceil, \ldots, \lceil \gamma^r K \rceil - \lceil \gamma^{r-1} K \rceil, \ldots$, για κάποιο $\gamma > 1$. Στη συνέχεια ξεκινάμε περιορίζοντας τις γειτονιές των κορυφών σε V_1 και τρέχουμε τον άπληστο αλγόριθμο μέγιστης κάλυψης σε αυτό. Εάν καλύψουμε όλους τους κόμβους ή εξαντλήσουμε τις επιλογές K που επιστρέφουμε. Διαφορετικά, συνεχίζουμε το ίδιο χρησιμοποιώντας το σύνολο V_2 και ούτω καθεξής, αφαιρώντας τους ήδη καλυμμένους κόμβους σε κάθε επανάληψη. Αν και είναι προφανές ότι ο αλγόριθμος ΒΓΜ˜ δεν αποδίδει γενικά ένα σύνολο λύσεων που ισούται με τη συμβατική άπληστη λύση και έχει αυστηρά μικρότερη αναλογία προσέγγισης, ο αλγόριθμος αποδίδει εξαιρετικά καλά αποτελέσματα όταν εκτελείται σε ΔΚΔ. Πιο συγκεκριμένα, για μια τιμή κατωφλίου $\tau = 4$, ένας πληθυσμός $n^{0.7}$ διασήμων κυριαρχεί περίπου το 70% του δικτύου, όπως φαίνεται στο Σχ. 7.2.

Δεδομένα

Τα στατιστικά του συνόλου δεδομένων μπορούν να βρεθούν στον Πίνακα ;;. Επιλέξαμε τα ακόλουθα δίκτυα για να αξιολογήσουμε τη μέθοδο μας. Τα περισσότερα από τα σύνολα δεδομένων είναι δημόσια διαθέσιμα στο SNAP [60] [7], με εξαίρεση τα dblp [90] και δβλπ-δψν [27] όπου τα δεδομένα παρέχονται από τον συντάκτη του [90, 27]. Στο σύνολο δεδομένων facebook, καταργήσαμε τους εξερχόμενους συνδέσμους από τον κόμβο του εγώ για να δείξουμε το μοτίβο κυριαρχίας μη ασήμαντα (διαφορετικά ο κόμβος εγώ κυριαρχεί το 100% των χρηστών όλη την ώρα). Στο σύνολο δεδομένων ποκες, εξαγάγαμε τις ετικέτες του Ποκες χειροκίνητα χρησιμοποιώντας τις πληροφορίες που παρέχονται από τα προφίλ χρηστών. Πιο συγκεκριμένα, διατηρήσαμε τη στήλη `hobbies` και διατηρήσαμε τα 280 πιο κοινά χόμπι. Καταργήσαμε επίσης τους κόμβους που δεν έχουν αποκαλύψει τις πληροφορίες προφίλ και τις συνδέσεις τους και είχαν το χαρακτηριστικό `public` ίσο με 0.

Πειραματική Διάταξη

Πραγματικά Δεδομένα. Για να ελέγξουμε την απόδοση του μοντέλου πειραματικά, πραγαματοποιήσαμε πειράματα ταξινόμησης με πολλαπλές ετικέτες (εύρεσης ενδιαφερόντων) όπου μας δίνεται ένα γράφημα μερικώς σεσημασμένο και στοχεύουμε να προβλέψουμε τις

[7] http://snap.stanford.edu/data/

ετικέτες που λείπουν στο γράφημα. Πραγματοποιούμε μια σύγκριση ιδίας-εισόδου-ιδίας-εξόδου, όπου η είσοδος μας αποτελείται από το διμερές γράφημα χρήστη-επηρεαστή και τις ετικέτες επιρροής και η επιθυμητή έξοδος είναι οι παράμετροι-πιθανότητες που πρέπει να προβλεφθούν. Η χρήση της παραπάνω διάταξης μπορεί να χρησιμοποιηθεί και σε συστήματα *συστάσεων* όπου οι πιθανότητες προσδιορίζουν μια κατάταξη (ranking) για το συγκεκριμένο χρήστη.

Πιο συγκεκριμένα, για κάθε κόμβο u και για κάθε ετικέτα $1 \leq i \leq d$ αποδίδουμε μια 'βαθμολογία' $\phi_{iu} \in [0, 1]$ που αντιπροσωπεύει την πιθανότητα ότι ο χρήστης θα υιοθετήσει αυτήν την ετικέτα (ενδιαφέρον). Συγκεντρώνουμε τους επηρεαστές του δικτύου χρησιμοποιώντας το ευρετικό BGMC και διατηρούμε το διμερές γράφημα μεταξύ των διασήμων και του υπόλοιπου δικτύου. Χρησιμοποιούμε μια τιμή κατωφλίου $\tau = 4$ και έναν εκθέτη $p = 0.7$ όπως φαίνεται στο Σχήμα 7.2. Στη σκηνή μας, υποθέτουμε ότι μόνο οι ετικέτες διασημοτήτων είναι γνωστές σε εμάς. Οι πληροφορίες του κοινού από τα προφίλ τους είναι ένα καλά υποστηριζόμενο σενάριο στην επιστημονική βιβλιογραφία, δεδομένου ότι οι επηρεαστές έχουν *οικονομικά κίνητρα* να το πράξουν. Επιπλέον, σε σενάρια πραγματικής ζωής, αυτός ο περιορισμένος αριθμός προφίλ μπορεί να εξορύσσεται μέσω κλήσεων σε REST APIs. Εκτελούμε πειράματα με $k \in \{\lceil \sqrt{n} \rceil, \lceil \log n \rceil\}$ γείτονες, με και χωρίς κανονικοποίηση (όπου χρησιμοποιούμε την αρχική κατάσταση ως σταθμισμένη επιπλέον γνώμη).

Στα πειράματά μας χρησιμοποιήσαμε το LSH για να συμπεράνουμε τους k πλησιέστερους γείτονες [8]. Πρώτον, συγκρίνουμε τη μέθοδο μας με το μοντέλο Random HK που περιγράφεται στο [36] το οποίο είναι το μοντέλο που μοιάζει περισσότερο με τη δουλειά μας. Αντί να κοιτάζει τους ΚΚΓ, το Ρανδομ ΗΚ επιλέγει ένα τυχαίο υποσύνολο με k γείτονες σε ακτίνα $\varepsilon > 0$ του χρήστη. Δεύτερον, εκπαιδεύουμε τις ενσωματώσεις node2vec [43], GraphWave [29] και NodeSketch [111] στο ίδιο γράφημα και στη συνέχεια ταιριάζουμε σε ένα μοντέλο λογιστις ρεγρεσσιον πολλαπλών ετικετών. Αυτό το είδος αναφοράς είναι σχεδόν τυπικό, όπως συζητάμε στην ενότητα Σχετική εργασία, στην εξόρυξη γραφημάτων.

Επιλέξαμε το node2vec ως μια κλασική προσέγγιση που βασίζεται σε τυχαία πεζοπορία, το GraphWave ως προσέγγιση που βασίζεται σε μετασχηματισμό κυματιδίων και το NodeSketch που είναι μια νέα μέθοδος που βασίζεται σε recursive linear sketching. Τα αποτελέσματα παρατίθενται στον Πίνακα ;;.

Συζήτηση

Αναφέρουμε αποτελέσματα καθόδου σε όρους AUC-ROC και RMSE σε όλα τα πειράματά μας: Στο σύνολο δεδομένων στο facebook έχουμε την καλύτερη απόδοση όσον αφορά το RMSE και έχουμε AUC κοντά στις άλλες μεθόδους. λιγότερο από 1% για όλες τις ετικέτες και παρόμοια αποτελέσματα για το τοπ-50 % και το τοπ-1. Στα dblp-dyn, fb-pages και github [9] Το σύνολο δεδομένων ξεπερνά τις άλλες μεθόδους — με εξαίρεση το ΑΥ͂-PO͂ στους τοπ-50% στο dblp-dyn όπου έχουμε μείωση 4%. Επιπλέον, στο σύνολο δεδομένων fb-pages, το GraphWave επιτυγχάνει ένα πολύ μικρό RMSE, ωστόσο αποδίδει πολύ χαμηλό

[8] Παρόμοια αποτελέσματα ελήφθησαν με ακριβείς μεθόδους.

[9] Το σύνολο δεδομένων περιέχει μία ετικέτα, επομένως τα αποτελέσματα AUC-ROC παραμένουν τα ίδια.

AUC-ROC. Τέλος, στο δίκτυο pokec, το GraphWave και το Random HK δεν κλιμακώνουν στους πόρους μας. Επιπλέον, το μοντέλο μας εκτελείται εντός 34 δευτερολέπτων με $k = \lceil \log n \rceil$ γείτονες και 377 δευτερόλεπτα με $k = \lceil \sqrt{n} \rceil$ γείτονες. Ταυτόχρονα, οι μέθοδοι ενσωμάτωσης που μπορούσαμε να τρέξουμε χρειάστηκαν δεκάδες λεπτά. Το βήμα PCA δεν επηρεάζει το χρόνο εκτέλεσης που χρειάζεται μόνο 1 δευτερόλεπτο, καθώς εκπαιδεύεται μόνο στους πολύ σημαντικούς κόμβους που είναι $n^{0.7}$, οι οποίοι είναι αμελιταίοι. Επιτυγχάνουμε AUC-ROC 91.84% και ΡΜΣΕ 0.025 όπου ξεπερνάμε το ΝοδεΣκετση σε όρους RMSE (6 φορές χαμηλότερο) και ξεπερνάμε από άποψη AUC-ROC κατά 0.3%. Τέλος, το node2vec έχει υψηλότερο ρυθμό AUC-ROC (με μικρό περιθώριο) σε σύγκριση με το μοντέλο μας με $k = \lceil \sqrt{n} \rceil$ γείτονες.

Συμπεράσματα

Εμπνευσμένοι από τις ισχυρές ομοφιλικές ιδιότητες των ΔΚΔ, παρουσιάζουμε το μοντέλο μας (NNIM). Αυτό το μοντέλο, αν και πολύ απλό να γίνει κατανοητό, παρουσιάζει σημαντική δυσκολία στο συμπερασμό παραμέτρων με απευθείας ΕΜΠ. Υιοθετούμε τη μεθοδολογία του ΕΜ για να αναπτύξουμε έναν αλγόριθμο για συμπερασμό παραμέτρων του μοντέλου που βασίζονται στην προσέγγιση μέσου πεδίου για να συναγάγουμε εξισώσεις μέσου πεδίου που μοιάζουν με τα παραδοσιακά μοντέλα ΔΔΑ. Το μοντέλο μας συγκλίνει αποδεδειγμένα σε πεπερασμένο χρόνο με ταχύτητα σύγκλισης που οριοθετείται αυστηρά από $k^{-t/2}$ για μεγάλα n. Ως εφαρμογή, μελετάμε το πρόβλημα του συμπεράσματος των ενδιαφερόντων των χρηστών σε ΔΚΔ από διάσημους κόμβους. Πιο συγκεκριμένα, δοκιμάζουμε χρησιμοποιούμε ένα υπογραμμικό σε μέγεθος σύνολο διασήμων ως "trend - setters" για να προετοιμάσουμε το μοντέλο μας και στη συνέχεια να εκτελέσουμε το NNIM σε ολόκληρο το δίκτυο. Δοκιμάζουμε τη μέθοδο μας σε δίκτυα διαφόρων μεγεθών και αξιολογούμε την ακρίβεια και την ποιότητα κατάταξης των μοντέλων. Έχουμε παρόμοια και τις περισσότερες φορές καλύτερα αποτελέσματα από τις πρόσφατες μεθόδους ενσωμάτωσης κόμβων και σχετικά μοντέλα ΔΚΔ.

Αντίκτυπος

Το έργο αυτό μπορεί να χωριστεί σε δύο ξεχωριστούς πυλώνες: θεωρητικό και πρακτικό. Ο πρώην πυλώνας έχει να κάνει με την εισαγωγή προβλημάτων στοχαστικής δυναμικής απόψεων και την ανάπτυξη αποτελεσματικών αλγορίθμων για την εξαγωγή παραμέτρων, καθώς και να παρέχει θεωρητικές εγγυήσεις για το καλό της προσέγγισης και της σύγκλισης. Οι θεωρητικές συνεισφορές από μόνες τους δεν παρουσιάζουν προβλέψιμες κοινωνικοοικονομικές συνέπειες.

Ο τελευταίος πυλώνας αυτού του εγγράφου βασίζεται στην πρόβλεψη των ενδιαφερόντων των χρηστών από τους επηρεαστές στα ΔΚΔ. Η εύρεση των διαδήμων σε ένα δίκτυο από την άποψη των δομικών ιδιοτήτων τους και η χρήση των ενδιαφερόντων τους για την επινόηση των ενδιαφερόντων του υπόλοιπου δικτύου έχει τόσο θετικές όσο και αρνητικές κοινωνικοοικονομικές συνέπειες. Από τη μία πλευρά, τα δεδομένα που παρέχονται από αυτούς

τους χρήστες είναι πιο πιθανό να είναι δημόσια — δεδομένου ότι οι διάσημοι χρήστες συνήθως εκθέτουν τέτοια δεδομένα για το δικό τους κέρδος — και μπορούν να ληφθούν μέσω κλήσεων REST API, οι οποίες μπορούν να αποτελέσουν τον κύριο οδηγό για στοχευμένες προτάσεις στο δίκτυα για τα οποία δεν γνωρίζουμε τις προτιμήσεις της πλειονότητας των χρηστών. Η έρευνά μας, δείχνει πειραματικά ότι τα αποτελέσματα μπορούν να επιτευχθούν εξετάζοντας μόνο αυτούς τους εξαιρετικά επιδραστικούς χρήστες ως 'διαμορφωτές ενδιαφέροντος' (διαμορφωτές τάσεων). Επιπλέον, το δεύτερο μέρος που περιλαμβάνει την εκτέλεση του μοντέλου NNIM για την προσομοίωση της ανταλλαγής απόψεων από εξαιρετικά ομοφιλικούς χρήστες. Αυτή η διαδικασία μπορεί να επιτρέψει στοχευμένες προτάσεις για ένα πολύ μεγάλο μέρος των χρηστών του δικτύου, δεδομένου ότι απαιτεί ένα απλό υποσύνολο των δεδομένων για να λειτουργήσει. Για να προσθέσουμε ένα παράδειγμα, σε ένα δίκτυο χρηστών με $n = 10^6$, ένα κλάσμα $n^{0.7}$ - αντιστοιχεί στο 1.58% των χρηστών. και το κλάσμα γίνεται ακόμη χαμηλότερο καθώς αυξάνουμε n. Από την άλλη πλευρά, αναγνωρίζουμε ότι η ενεργοποίηση συστάσεων χωρίς να γνωρίζουμε ποιο περιεχόμενο προτιμά ήδη ένας χρήστης — το οποίο μπορεί να είναι προσωπικές πληροφορίες για τον χρήστη — εξετάζοντας μόνο τις συνδέσεις του με κόμβους με μεγάλη επιρροή στο δίκτυο ενδέχεται να μην χρησιμοποιείται σωστά από εξωτερικούς πράκτορες.

Μελλοντικές Προεκτάσεις

Αυτή η εργασία μπορεί να επεκταθεί σε πολλές ενδιαφέρουσες μελλοντικές κατευθύνσεις. Πρώτα απ ́όλα, η χρήση της δομής πυρήνα-περιφέρειας σε αλγόριθμους επιτάχυνσης μπορεί να επεκταθεί και σε άλλα προβλήματα. Παραδείγματα προβλημάτων είναι τα συντομότερα μονοπάτια όλων των ζευγαριών (εύρεση μετρήσεων κεντρικότητας σε ένα δίκτυο), κατάταξη χρηστών σε ένα δίκτυο (π.χ. PageRank) και αλγόριθμοι που βασίζονται σε τυχαίους περιπάτους. Επιπλέον, η συγκεκριμένη κατανόηση της δομής πυρήνα-περιφέρειας μέσω γενετικών μοντέλων είναι επίσης μια ανοιχτή γραμμή εργασίας.

Επιπλέον, τα έργα μας παρέχουν μια στατιστική εξήγηση για τη δυναμική της γνώμης, επεκτείνοντας την υπάρχουσα θεωρητική κατανόηση των διαδικασιών σχηματισμού γνώμης [12, 11]. Η επέκταση αυτής της γραμμής εργασίας, λαμβάνοντας υπόψη γενικότερες ρυθμίσεις (π.χ. εκθετικές οικογένειες) και εξελικτικές διαδικασίες θα μπορούσε να αποφέρει σημαντικά αποτελέσματα σε εργασίες εκμάθησης γραφημάτων και αλγόριθμους συμπερασμού.

Chapter 1

Introduction

"The masses have never thirsted after truth. They turn aside from evidence that is not to their taste, preferring to deify error, if error seduce them. Whoever can supply them with illusions is easily their master; whoever attempts to destroy their illusions is always their victim. An individual in a crowd is a grain of sand amid other grains of sand, which the wind stirs up at will."

— Gustave Le Bon, The Crowd: A Study of the Popular Mind

1.1 Motivation

The wealth of nowadays' networks is tremendous. One can observe networks almost everywhere: social networks, traffic networks, biological networks, production networks and particle interaction networks are some very vivid examples. The tendency of the various life forms, under nature or society, to bond and cooperate gives rise to rich patterns which govern our daily lives.

It is well understood that most large-scale Online Social Networks (OSN) exhibit the so-called *core-periphery structure* (see e.g., [48, 114, 88, 115, 99, 82, 109] and the references therein). Namely, their nodes are naturally partitioned into a *core set* C of nodes that are tightly connected with each other, and a *periphery set* U, where the nodes are sparsely connected, but are relatively well-connected to the core. In most cases, the core nodes almost dominate the rest of the network, in the sense that a small fraction of δn high-degree nodes dominate an $(1-a)n$ fraction of the network's engaged nodes (where "engaged" refers to nodes with degree above than a certain threshold). If we restrict to engaged nodes only, even a sublinear fraction of nodes dominate almost everything (see also [16, 17, 18]). These influential core nodes, which posses a large number of incoming connections, or *followers*, are also known (and serve) as the *celebrities* or the *influencers* of the network. Influencers tend to publicly expose — mainly for commercial reasons [52, 37, 49, 100, 19] — their profile information (friends and interests), thus information can be gathered easily, for example through *REST API calls*.

Another major driving force shaping the structure of social network is *homophily*, i.e., the property under which connected individuals in a social network have similar interests [73, 72]. Modern large-scale OSN seem to exhibit strong homophilic trends, which was a major part of our motivation (see also Chapter 2).

1.2 Approach and Contribution

In this Diploma Thesis, we leverage homophilic trends and the core-periphery structure of modern OSN to obtain scalable and accurate learning methods for predicting the interests of a network's peripheral users. Our approach is to identify and use the influencers of the network as *steady-state trend-setters* and let the network around them evolve according to an iterative process initialized from an aggregation of the influencers' features. The influencers' sublinear number allows for a quite fast initialization (in worst-case strongly subquadratic-time) of the users' interests. Inspired by *coevolutionary opinion formation* [46, 11], we next treat the network as the result of a natural *interest exchange* dynamical process, where each peripheral user updates her features according to the interests of her k-nearest neighbors in the periphery, until *consensus* is reached (see also Chapter 4).

We use the interest space of the network generated by this process to infer the probability that a peripheral user adopts certain interests (a task equivalent to multilabel classification). Key to the algorithm's scalability is that throughout the process, each peripheral user interacts only with her k-nearest neighbors.

More specifically, a key part of our approach is the *Nearest Neighbor Influence Model* (NNIM), a stochastic iterative process according to which users evolve their binary interest vectors. At each timestep, each peripheral user samples a new binary interest vector based on the interests of her k nearest neighbors (wrt. their interest vectors) in the periphery. The general structure of NNIM is inspired by the Hegselmann-Krausse model [46]. However, NNIM is stochastic and is used as a *generative model*, aiming to explain, through homophily, the coevolution of the network structure and the peripheral user interests (see Chapter 7).

From a bird's view, our prediction method aims to recover the latent NNIM interest vectors of the peripheral users that maximize the likelihood that NNIM evolves as observed. Although the idea is simple, its efficient implementation requires significant effort and care (see Section 7.1.1 and Chapter 3). We use Variational Expectation-Maximization, due to the latent nature of NNIM, since direct maximization of the log-likelihood is intractable. As a result, we obtain a simplified mean-field approximation of NNIM (see Algorithm 1, Theorem 5 and (7.21)), which is similar to the classical opinion dynamics equations, thus establishing a connection between stochastic and deterministic opinion dynamics. We prove (see Theorem 7) that our algorithm converges in a finite number of steps and establish an upper bound between the total variation distance, the number of iterations, and the number k of neighbors used in the interest exchange processes (which affects the running time). Our algorithm efficiently scales to networks with millions of

nodes.

Our user interest prediction method scales smoothly to networks with millions of nodes, with an *almost linear-time complexity*, for appropriate choices of hyperparameters (see Table 7.2). We evaluated our method experimentally on six standard network benchmarks taken from [60, 90, 27] with quite different characteristics (see Table 7.1). Our experimental results suggest that our method performs similarly (or often outperforms) sophisticated node embedding and traditional opinion dynamics methods in terms of AUC-ROC and RMSE, whilst being able to run up to 100 times faster than the best known node embedding methods in networks with up to 10^6 nodes (see Table 7.3).

Conceptually, our work draws ideas from (and contributes to) three major research directions (see also the comparison to previous work in Section 7.4). From an algorithmic perspective, we take advantage of the core-periphery structure of OSN to speed up inference in large-scale networks. Moreover, we introduce and analyze a natural stochastic generalization of coevolutionary opinion dynamics, which we eventually utilize for user interest prediction. As a result, we obtain a new truly scalable user prediction approach with excellent accuracy. Our methodology can be extended to a variety of problems in combinatorial optimization and machine learning, where inference from the entire network leads to prohibitive running times.

Chapter 7 contains the main contribution of the Thesis. More specifically, in terms of practical contributions, we open-source the code[1] of the thesis[2]. The code is implemented using NumPy[3], NetworkX[4], Scikit-learn[5], and Annoy[6]. In terms of theoretical contributions we contribute all the theorems presented in Chapter 7. The Extended Abstract in Greek contains the same information as Chapter 7, together with some additional information, in the English Version. The data used for the experiments in this thesis are *anonymized* and openly available on the Internet. We redirect the interested experimentalist to their original sources; for the avoidance of second-hand bias.

1.3 Thesis Structure

The current Thesis is written in two languages: English and Greek. The parts of the Thesis that have been written in Greek contain an overview of our contribution in Greek and are presented in the prelude. Following, there is a depth-one overview of the work that has been used in order to arrive at the results presented as the main contribution. Finally, Chapter 7 (Part C) contains the main contribution of our Thesis. This Thesis' results have been submitted to a conference, and are currently under review [87].

The depth-one related works to our Thesis are presented in Parts A, B, C, and Ap-

[1] Licensed under the MIT License
[2] https://shorturl.at/fxS34
[3] https://numpy.org
[4] https://networkx.github.io
[5] https://scikit-learn.org
[6] https://github.com/spotify/annoy

pendix A. The Thesis contains all the theoretical and technical tools to understand the contribution of the thesis. Broken in parts, the constituent Chapters link with the contribution as follows

Part A: Online Social Networks. This part gives an overview of the basic characteristics of Online Social Networks (Homophily, Scale-free degree distribution, Small-world properties, Densification Laws and Core-periphery structure). After reading it, the reader should be able to understand the *motivation and the hypothesis testing* behind the proposed model.

Part B: Learning. This part gives a brief overview of the various learning techniques involved with our contribution. To begin with, Chapter 3 does a basic introduction in Generative Models and briefly presents methods for performing statistical inference (MLE, EM and EM Variants). The main purpose of this Chapter is to assist the reader through understanding how the inference algorithm for our generative model works. Afterwards, Chapter 4 gives an overview about classical models in Opinion Dynamics (FJ, LIP-FJ, DeGroot, HK, Random HK, Network HK), and *coevolutionary opinion formation*. Since Opinion Dynamics are equivalent to Dynamical Systems, we need to study their properties as Dynamical Systems, that is to account for convergence properties, convergence rates and clustering behaviour. For these reasons, Chapter 5 briefly gives an overview of dynamical systems and their behaviour for $t \to \infty$. It also examines techniques for proving Global Asymptotic Stability (GAS) and makes a reference to the second largest eigenvalue theorem of a k-regular graph. Results from this Chapter are used to prove the two main theorems of our contribution (Theorem 7 and Theorem 5). For implementation matters of the proposed model, Chapter 6 gives an overview of Unsupervised Feature Learning Techniques which are used (implementation-wise) in the Thesis. The techniques involve dimensionality reduction (PCA, Random Projections (JL Transform), MinHash) and nearest neighbor search (KD Trees, Ball Trees, LSH, DCI/PDCI).

Technical Tools. Appendix A gives an overview to technical theoretical tools that are used to help establish the theoretical basis of our paper.

Further Reading. For a more concrete and more-in-depth overview of the tools used within the thesis we redirect the interested reader to the excellent classical textbooks in Machnine Learning [98, 13, 32].

Part A

Online Social Networks

Chapter 2

Characteristics of Social Networks

"Ομοιος ομοίω αεί πελάζει"

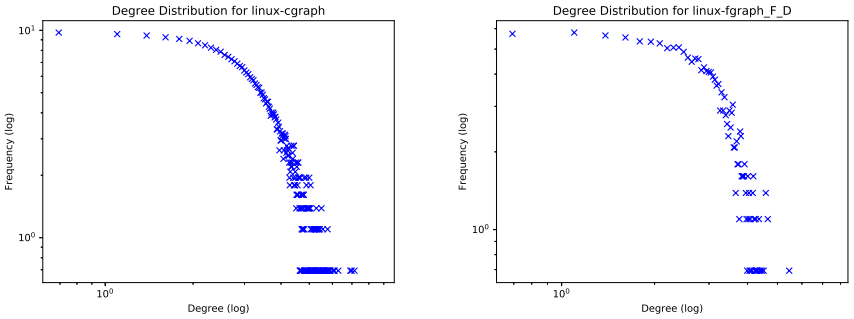
— Plato, Symposium, c. 385–370 BC

2.1 Motivation

A research question that has been out for many decades is concerned with the understanding of structural properties and patterns inside real-world networks. Questions like *"What does a 'normal' network look like?"*, *"How does the network evolve?"* and *"What abnormalities can arise in a real-world network?"* are crucial to a wide range of applications in economics, epidemiology, sociology and computer science. Attention has also been shifting from "node-centric" approaches — i.e. approaches where properties of individual nodes in the network are examined — to "network-centric" — that is to study the network properties treating the network as an entity. Below, we present some properties of social networks, such homophily, scale-free distributions, core-periphery structure, shrinking diameters and small-world properties.

2.2 Homophily

Homophily from Ancient Greek "homou" (same) and Greek "philia" (friendship) is the tendency of be friends with similar others. A proverb for this property is well known: "birds of a feather flock together". Homophily has been well observed in multiple instances of network and in various forms. Modeling of homophilic processes is usually done through a Blau Space [73, 73] which is a multidimensional coordinate system where the socio-demographic variables come as different dimensions. Examples of dimensions include age, sex, years of education, salary, geographic location and so on. The organizing force in Blau space is the homophily principle, which argues that the flow of information from person to person is a declining function of distance in Blau space. Persons located at great distance in Blau space are very unlikely to interact, which creates the conditions for social differences in any characteristic that is transmitted through social communication. The homophily



(a) Call Graph (b) FD Graph

Figure 2.1: Power law degree distributions in software projects. Log-log plots of the Linux Kernel 20.3M-long codebase.

principle thus localizes communication in Blau space, leading to the development of social niches for human activity and social organization. Mathematically, the smaller the distance $\|\boldsymbol{x}_u - \boldsymbol{x}_v\|$ is between the users, the higher is the probability of them being connected.

Individuals in homophilic relationships share common characteristics (beliefs, values, education, etc.) that make communication and relationship formation easier. The opposite of homophily is heterophily or intermingling.

2.3 Scale-free Degree Distributions

Many real-life social networks have degree distributions that come in the form of a power law, that is a function $p(k)$ such that for all constants b the property $p(bk) = g(b)p(k)$ holds. That is the fraction of the nodes $p(k)$ with degree k behaves as $k^{-\gamma}$ for large values of k. The value of the parameter γ is usually between 2 and 3 [53]. The main contributing factors which explain the emergence of scale free distributions are growth, preferential attachment and latent features. The "growth" part refers to new nodes joining the existing network over an extended period of time, the "preferential attachment" part refers to the "rich nodes getting richer" [7], and the latent variables [95, 4] refer to sets of latent variables that lead to power-law-like degree distributions such as in [55, 54]. It is important to highlight that scale-free properties emerge in multiple domains, such as statistical physics through molecular interactions, software through the degree distribution of calls after static analysis [67, 86], the World Wide Web [61], airline networks and many more.

The most notable characteristic in a scale-free network is the relative commonness of vertices with a degree that greatly exceeds the average. The highest-degree nodes are often called "hubs" (in our work these nodes will be called influencers or celebrities), and are thought to serve specific purposes in their networks. In this work, we show how a sublinear fraction of these nodes forms the opinions of a linear fraction of the nodes.

Scale-free networks have higher tolerance to faults. This characteristic emerges largely

due to the structure itself. The influencer nodes are followed by nodes of smaller degree, these nodes by nodes of (even) smaller degree and so son. If a failure occurs uniformly at random, since most nodes have small degrees, the overall network will not be affected extensively.

Moreover, scale-free networks have a clustering coefficient that is as well scale-free. More precisely, the local clustering coefficient for a node u is defined as the fraction of somenone's friends that are friends with one another.

$$\mathrm{LCC}(u) = \frac{|E \cap (N(u) \times N(u))|}{|N(u) \times N(u)|} \tag{2.1}$$

The global clustering coefficient is defined as

$$\mathrm{GCC} = \frac{\text{Number of closed triplets}}{\text{Number of all triplets}} = \frac{3 \times \text{Number of triangles}}{\text{Number of all triplets}} \tag{2.2}$$

Hence, the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs. This is a main drive for people to form communities, which are small groups in which most people know most people. A node, depending on its position on the distribution (or its "fame") tends to belong to more communities the more famous it is. This fact, combined with homophily, serves as a drive on how "influencers shape interests" in a network. For example, a famous footballer like Christiano Ronaldo endorses many people to be interested in "football" and a politician drives people towards certain political affiliations. We will come back to this idea later in this thesis, where we will examine sampling and fractional domination.

2.3.1 Examples of Scale-free Distributions

Zipf distribution. Zipf's Law originated in the field of natural language processing [95, 4]. Each word is associated by a rank (how frequent the word is) and the distribution of the word is proportional to the inverse rank. Zipf's law is most easily observed by plotting the data on a log-log graph, with the axes being log (rank order) and log (frequency). Formally, let:

- n be the number of elements
- k be the rank of a word
- s be the exponent

Then

$$f(k) = \frac{1}{H_{n,s}k^s} \qquad 1 \leq k \leq n \tag{2.3}$$

where $H_{n,s} = \sum_{k=1}^{n} \frac{1}{k^s}$.

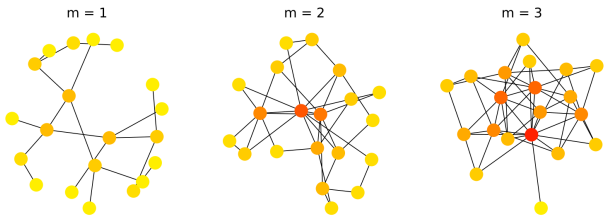


Figure 2.2: Barabási-Albert Model for $T = 3$ iterations. Source: Wikipedia.

Preferential Attachment. The "Preferential Attachment" mechanism refers to the principles of the rich nodes getting richer and the poor nodes getting poorer. It has been shown that such mechanisms generate power law distributions. One example is the Barabási-Albert (BA) model [7]. In the BA model, we start with a network with n_0 nodes. At each step $t \geq 1$ a node v_t arrives and is connected to $n \leq n_{t-1}$ nodes with probability that is proportional to the degree of the existing nodes, that is

$$\Pr[(v_t, s) \in E_t] = \frac{\deg_{G_{t-1}}(s)}{\sum_{z \in V_{t-1}} \deg_{G_{t-1}}(z)} \qquad s \in V_{t-1} \tag{2.4}$$

Influencers tend to quickly accumulate even more links, while nodes with only a few links are unlikely to be chosen as the destination for a new link. The new nodes have a "preference" to attach themselves to the already heavily linked nodes. The BA model demonstrates a power law of k^{-3}.

2.4 Small-world

A small-world network is a type of mathematical graph in which most nodes are not neighbors of one another, but the neighbors of any given node are likely to be neighbors of each other and most nodes can be reached from every other node by a small number of hops or steps. Specifically, a small-world network is defined to be a network where the average length of the shortest path between any two nodes u, v is proportional to $\log n$, where n is the number of nodes in the network. Small-world phenomena have been found to hold extensively in multiple social networks. Travers and Milgram [108] with their famous "six-degrees of separation" experiment demonstrated that people that phenomenally seem very "far" apart (from an acquaintance viewpoint) are separated by at most six steps. Furthermore, the work of Watts and Strogatz (Watts-Strogatz model) provided a very simple random graph model [110] where we are given a graph G with vertices $v_1, \ldots, v_n$ and each vertex is connected to the k vertices right of it (modulo n). Then each vertex the $k/2$ rightmost edges are being rewired with probability β uniformly at random with probability $\frac{1}{n-1}$. In limiting case for $\beta \to 1$ the Watts-Strogatz model has an average path length of $\frac{\log n}{\log k}$ and in the cases where $\beta \in (0, 1)$ the path length drops rapidly as β increases. Figure 2.3 shows how the Watts-Strogatz model behaves

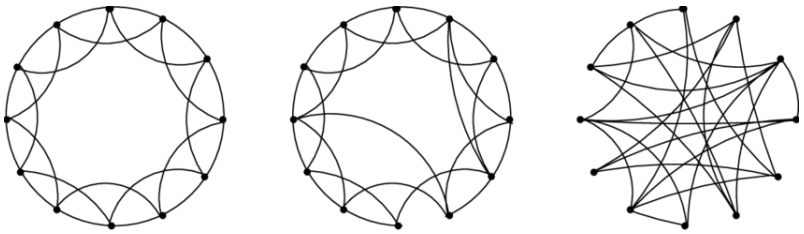


Figure 2.3: Behaviour of the Watts-Strogatz model on a graph with $n = 12$ vertices. In the leftmost figure the value of β is 0 indicating complete order. In the middle figure the value of β has increased near $1/2$ and small-world phenomena start to appear. Finally, in the rightmost figure consists of the case of $\beta = 1$ where the $k/2$ rightmost edges do rewire uniformly at random, each with probability $\frac{1}{n-1}$.

for increasing values of β (from left to right). Later, Kleinberg [56] verified small-world phenomena using a random graph model where $n = \nu^2$ vertices were positioned in a square lattice, each vertex u was connected to its 4-neighborhood of distance 1, and each vertex u was connected to one vertex v outside of its 4-neighborhood with probability proportional to $\|\boldsymbol{x}_u - \boldsymbol{x}_v\|_1^{-a}$ where $a > 0$ is a clustering exponent. The agents are requested to deliver a message from a start s to a terminal t and each agent forwards the message to the neighbor which is closest to the target distance (by L1 distance). What Kleinberg realized is that the expected delivery time T depended on the clustering exponent a and not on the size ν of the network[1].

2.5 Densification Laws and Shrinking Diameters

In the work of Leskovec, Kleinberg and Faloutsos [59] the authors discover that in many real-world graphs the diameter of the graphs decreases as new $\omega(n)$ edges are added on the graph. More specifically, they observe that the average path length between the nodes of the graph shrinks contrary to conventional wisdom that the average path length will increase in terms of n. The Community Guided Attachment model they propose is a tree of height H with $n = b^H$ leaves, which represent a communities-within-communities structure[2]. Then they connect edges between leaves u and v depending on the function $h(u, v)$, that is the height of the subtree rooted at $\mathrm{LCA}(u, v)$. The probability of connection is

$$f(u, v) = c^{-h(u,v)}$$

The authors prove that the average out degree $\bar{d}$ is $n^{1-\log_b(c)}$ for $c \in [1, b)$, $\log_b(n)$ for $c = b$ and $\Theta(1)$ if $c > b$. Then the number of edges is $\bar{d}n = n^a$ indicating that when $c \in [1, b)$ the network obeys the densification law with $a = 2 - \log_b(n)$. Then, they extend the model by allowing new nodes to join the existing structure in the form that at each

[1] The same applies if each node has $p \geq 1$ short-range connections and $q \geq 1$ long-range connections.
[2] Implying scale-free properties.

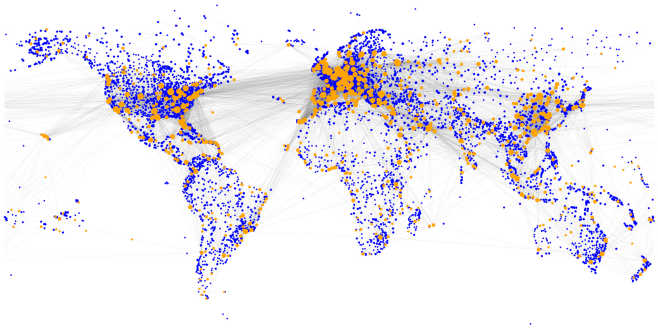


Figure 2.4: Core-periphery structure in an airline network. Source: [48].

timestep t, b leaves are added to the existing leaves. Moreover, the connections can now happen between every pair of nodes and have probabilities equal to

$$f(u, v) = c^{-\gamma\delta(u,v)}$$

where $\delta(u, v)$ is the shortest path between the nodes on the tree. Now, observe that for two leaves $\delta(u, v) = 2h(u, v)$ since the tree is perfect. Hence, in order to accord with the original model one has to set $\gamma = \frac{1}{2}$. In the same fashion, the authors show that the model obeys the same behaviour with the exception of the case when $c > b^2$ where a power law appears.

Finally, they introduce the Forest Fire model which is described by a graph process G_t such that.

- G_1 contains only one node.
- For $t \geq 2$ a new node v joins G_t.
- Two numbers x and y are generated following geometric distibution with means $p/(1-p)$ and $rp/(1-rp)$. The new node selects x outlinks and y inlinks and chooses nodes $w_1, \ldots, w_{x+y}$ that have not been visited. If there are less than x outlinks or y inlinks available, occupy the available links.
- The links are added accordingly and nodes $w_1, \ldots, w_n$ are marked as visited.

The authors show that the model obeys all the desiderata for a real-network and shrinks as new edges are added to the network. Firstly, the network has power-law in-degree and out-degree distributions, has a copying flavour (a newcomer copies the neighbors of his/her ambassador), obeys a densification power law (a user is most engaged to the community of his/her ambassador) and a *shrinking diameter*.

2.6 Core-periphery structure

The core-periphery structure is a structure model in real-world networks which was firstly introduced by Wallersteiin in reference [109] and builds on an actual observation

that nodes in a network belong to two categories[3]

The former category is the core C and the latter is called the periphery P. The nodes of the core set are tightly connected to themselves, the peripheral nodes are connected to vertices of the core and the nodes of the peripheral set are loosely connected to one another (in the ideal case, no connections occur betweeen the peripheral nodes). Intuitively, if p_{CC} is the probability that two nodes in the core are connected, $p_{CP} = p_{PC}$ is the probability that a node of the periphery is connected to a node at the core and p_{PP} is the probability that two peripheral nodes are connected then

$$p_{CC} > p_{CP} > p_{PP}$$

Another way to view the core-periphery structure is through the continuous model of Jia and Benson [48] according to which each node is associated with a real "coreness" score θ_u. More specifically, if $\theta_u > 0$ then the node is part of the core and if $\theta_u < 0$ then the node is a peripheral one. The probability of and edge (u, v) appearing between nodes u and v is given as

$$p(u, v) = \frac{1}{1 + \exp(-\theta_u - \theta_v)}$$

It is clear that again $p_{CC} > p_{CP} > p_{PP}$ is satisfied since the sum $\theta_u + \theta_v$ decreases as we move from core-core (positive), to core-periphery (near zero) and then to periphery-periphery (negative). Moreover, Jia and Benson add spatial features $\{\boldsymbol{x}_u\}_{u \in V}$ and extend their model to include them as

$$p(u, v) = \frac{1}{\epsilon \|\boldsymbol{x}_u - \boldsymbol{x}_v\|_2 + \exp(-\theta_u - \theta_v)}$$

However, the qualitative notion that social networks can have a core–periphery structure has a long history in disciplines such as sociology, international relations, and economics. Observed trade flows and diplomatic ties among countries fit this structure. For instance, the airline network studied by Jia and Benson depicted in Figure 2.4 discerns hubs of high airline traffic — such as Europe — and peripheral nodes of low traffic — such as Oceania. Krugman [58] argues that when transportation costs are low enough manufacturers concentrate in a single region known as the core and other regions (the periphery) limit themselves to the supply of agricultural goods.

[3]Original work by Wallerstein discerned between three categories core, semi-periphery and periphery. "World-system" refers to the inter-regional and transnational division of labor, which divides the world into core countries, semi-periphery countries, and the periphery countries. Core countries focus on higher skill (high capital production), and peripheral countries focus on labor-intensive production and raw materials. This system subsequently augments the power of the core countries toward the rest. Besides, the system has dynamic characteristics, in part as a result of revolutions in transport technology, and individual states can gain or lose their core (semi-periphery, periphery) status over time, since the world-systems are rooted in capitalist and imperialist economies.

Finally, the core-periphery structure of networks combined with the scale-free properties of the networks motivates us to further investigate the structure of the influencers[4] inside real-world networks. In modern follower-based networks, someone may be interested in covering a large portion of the network using the influential nodes. Practically, one may have easier access to public profile information of celebrities — such as sports players, show-biz people and politicians — than from ordinary people. Identifying these "core-players" in an online social network can be a key for designing efficient algorithms. The high-level idea is to sample these nodes using a simple sampling procedure which can scale to millions of nodes, use these nodes — for example obtain the subgraph which they span or the bipartite subgraph between the influencers and the rest of the network — and design an efficient algorithm. The computational motivator behind it is that on a network with n nodes the $O(n^2)$ worst-case cost of traversing all the edges is prohibiting for large network applications. More specifically, we argue that in a real-world network, the number of influencers is sublinear with respect to the total number of users. The quantifying objective is the one of *coverage*, that is how many nodes can these influencers cover. The coverage objective is directly connected to the notion of the *Almost Dominating Set*.

Definition 1. *A subset $S \subseteq V$ of a vertex set of a graph $G(V, E)$ with $|V| = n$ vertices is called an a-Almost Dominating Set if and only if it dominates at least an a of the vertices, that is at least a fraction a of the nodes has neighbors in S.*

In our study we will see how a sublinear fraction of the nodes that is for example $n^{0.7}$ nodes can dominate more than 75% of the network. For instance in a network of 8 million users, the $n^{0.7}$-fraction is 0.8% of the total nodes. Undoubtedly, using information from only 0.8% of the nodes can be a huge advantage for designing algorithms. For instance, one can calculate approximate all-pairs shortest paths via precomputing all-pair-shortest paths to the fractional subgraph and get very good approximations. Our applications will mainly focus on interest prediction and the list can go on.

[4]The terms "influencers", "celebrities" and "hubs" are used interchangeably throughout the text.

Part B

Generative Models

Opinion Dynamics

Dynamical Systems

Unsupervised Learning

Chapter 3

Generative Models

3.1 Motivation

Learning in general can be of two main types. The first one is *discriminative learning* where we do not pose any assumptions on the underlying distribution $\mathcal{D}$ of the data. In discriminative learning, our goal is to build a good predictor and not the distribution itself. However, if we know the distribution $\mathcal{D}$ which is modeled by a set of parameters $\boldsymbol{\theta}$, we can do much more compared to discriminative learning. First of all, we can generate samples from the distribution given that we know its parameters. Secondly, we can learn — given an adequate number of samples — the *parameters* of the distribution such that the PDF $p(\boldsymbol{x}|\boldsymbol{\theta})$ of the distribution fits the known samples in the best possible way. The samples may either be observed directly or generated through a multi-level cause-result procedure modeled by a Bayesian network.

Below, we discuss parameter learning schemes, starting from the simplistic Maximum Likelihood Estimator, and then continuing with latent variable model learning through Expectation-Maximization and its variants. For more information about generative models we redirect the interested reader to [98] and [13].

3.2 Maximum Likelihood Estimation

To introduce the notion of Maximum Likelihood Estimation (MLE) we will start by giving a simple example. Suppose that we have a coin that can either come heads or tails when tossed and each outcome is labeled with 0 and 1 respectively and that we have observed n independent coin tosses $X_1, \ldots, X_n$ via simulating experiments. It is easy to assume that our coin is modeled by a Bernoulli distribution with probability μ of being 1. We now want to find the best possible μ which explains the probability of observing $\boldsymbol{X} = (X_1, \ldots, X_n)$. We therefore define the function

$$p(\boldsymbol{X}|\mu) = \prod_{i=1}^{n} p(X_i|\mu) = \prod_{i=1}^{n} \mu^{X_i}(1-\mu)^{1-X_i} \tag{3.1}$$

And the function

$$\mathcal{L}(\mu) = \sum_{i=1}^{n} X_i \log \mu_i + (1 - X_i) \log(1 - \mu_i) \tag{3.2}$$

For finding the best μ that describes the function, it suffices to set the derivative with respect to μ equal to 0. Therefore, after solving for μ, we have

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3.3}$$

The procedure described above is the most characteristic example of MLE. Moreover, as we have described in the section about Concentration Bounds this estimate is very close to the actual value μ of the distribution. To be more specific, the CH Bound states that with probability of at least $1 - \delta$, the following holds

$$|\hat{\mu} - \mu| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \tag{3.4}$$

Indeed, if we let $n \to \infty$ then $\hat{\mu} \to \mu$ a.a.s., which means that our estimator is *asymptotically consistent*. In general the MLE seeks the optimal set of parameters $\boldsymbol{\theta}$ that maximize the joint distribution $p(\boldsymbol{X}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$.

3.3 Inference of Latent Variable Models through Expectation-Maximization

3.3.1 Latent Variable Models

In generative models of random variables we usually assume that our data is sampled from a specific distribution $\mathcal{D}$ for which all the variables are known, if provided a sample from $\mathcal{D}$. However, this is not always the case. Imagine that an unobserved variable Z depends on the observed variable X through some distribution for which we want to learn the parameters. Sometimes it is convenient to express such phenomena through networks of latent variables. Below, we give some examples:

Example 1: Gaussian with Normally Distributed Mean. Assume that $X \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is deterministic and μ is a random variable distributed as $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ where μ_0 and σ_0 are known parameters.

Example 2: Gaussian Mixture Models. Consider k Gaussians $\mathcal{N}(\mu_1, \sigma_1^2), \ldots, \mathcal{N}(\mu_k, \sigma_k^2)$. Then consider the categorical variable $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ for which $p(Z_i = 1) = \pi_i$ and $\sum_{i=1}^{k} \pi_i = 1$. We then observe the variable X such that $p(X|Z_i = 1) = \mathcal{N}(\mu_i, \sigma_i^2)$. The PDF of the data X is given as

$$p(X) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mu_i, \sigma_i^2) \tag{3.5}$$

Our goal is to infer the means μ_k and the variances σ_k^2 as well as the mixture components π_i only by observing data from $p(X)$.

Example 3: A toy problem in OSN. Given an OSN $G(V, E)$ each user u has a hidden binary attribute X_u which is Bernoulli-distributed with parameter p. For every pair of users $\{u, v\} \in E$ we observe and edge which appears with probability $p(u, v) = \sigma(X_u X_v)$ independently of the other edges where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function. By observing the edges $e_1, \ldots, e_m$ of the network we desire to find the probability that an attribute X_u is 1.

Generalilzation. More generaly, given a set of (independent) observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ our goal is to infer the parameters $\boldsymbol{\theta}$ which maximize the log-likelihood of the observed data, namely

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{X}|\boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{3.6}$$

Contrastingly to a classical MLE problem, the above problem poses severe computational barriers since the summation/integration with respect to $\boldsymbol{Z}$ inside the logarithm is in general

1. Usually not tractable in P-time.
2. A closed-form for the solution is very difficult to be found, like in the case of Gaussian Mixtures.

For that reason, we assume that the latent variables $\boldsymbol{Z}$ are modeled by a variational distribution $Q(\boldsymbol{Z})$ which is non-zero at the domain of $\boldsymbol{Z}$. We now rewrite the log-likelihood under this assumption, i.e. by dividing and multiplying by $Q(\boldsymbol{Z})$

$$\mathcal{L}(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{Z}} Q(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{Q(\boldsymbol{Z})} = \log \mathbb{E}_{Q(\boldsymbol{Z})} \left[\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{Q(\boldsymbol{Z})} \right] \tag{3.7}$$

By Jensen's Inequality for the function $f(x) = \log(x)$ — which is concave since $f''(x) = -1/x^2 < 0$ — we have that

$$\mathcal{L}(\boldsymbol{\theta}) \geq \mathbb{E}_{Q(\boldsymbol{Z})} \left[\log \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{Q(\boldsymbol{Z})} \right] = \mathbb{E}_{Q(\boldsymbol{Z})} \left[\log p(\boldsymbol{X}, \boldsymbol{Z}) \right] + \mathbb{E}_{Q(\boldsymbol{Z})} \left[-\log Q(\boldsymbol{Z}) \right] \tag{3.8}$$

The quantity $\mathcal{L}_Q = \mathbb{E}_{Q(\boldsymbol{Z})} \left[\log p(\boldsymbol{X}, \boldsymbol{Z}) \right]$ is called the Evidence Lower Bound (ELBO) whereas the quantity $\mathcal{H}_Q = \mathbb{E}_{Q(\boldsymbol{Z})} \left[-\log Q(\boldsymbol{Z}) \right]$ is the entropy of $Q(\boldsymbol{Z})$. It can be well understood from this form that optimization with respect to $Q(\boldsymbol{Z})$ on the right hand side

is a well-defined problem since an optimal solution of $\mathcal{L}_Q + \mathcal{H}_Q$ is at most the optimal value of the actual likelihood $\mathcal{L}$. This EM approach for learning latent variable models was first introduced by Dempster and Rubin in their classical paper [26]. In general, the choices that are available for selecting the function $Q(\boldsymbol{Z})$ are ample, there are some main categories to take into account [13]:

Classical EM. The function $Q(\boldsymbol{Z})$ is defined as the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_0)$ given the previous values of the parameters. In a Gaussian Mixture Model the posterior probability $p(z_i = 1|\boldsymbol{x})$ of a sample x belonging to the i-th Gaussian is given as

$$\gamma_i = p(z_i = 1|x, \boldsymbol{\theta}_0) = \frac{p(x|z_i = 1)p(z_i = 1)}{\sum_i p(x|z_i = 1)p(z_i = 1)} = \frac{\pi_i \mathcal{N}(x|\mu_{i0}, \sigma_{i0}^2)}{\sum_{i=1}^k \pi_i \mathcal{N}(x|\mu_{i0}, \sigma_{i0}^2)} \tag{3.9}$$

Hence the Evidence Lower Bound for this sample is calculated as

$$\begin{aligned} \mathcal{L}_{Q,j} &= \sum_{i=1}^{k} p(z_i = 1, \boldsymbol{z}_{-i} = 0|x, \boldsymbol{\theta}_0) \log p(x, z_i = 1, \boldsymbol{z}_{-i} = 0|\boldsymbol{\theta}) \\ &= \sum_{i=1}^{k} \gamma_i \log \left(\pi_i \mathcal{N}(x|\mu_i, \sigma_i^2)\right) \end{aligned} \tag{3.10}$$

And for a set of samples $\boldsymbol{X} = (x_1, \ldots, x_n)$

$$\mathcal{L}_Q = \sum_{j=1}^{n} \sum_{i=1}^{k} \gamma_{ji} \log \left(\mathcal{N}(x_j|\mu_i, \sigma_i^2)\right) \tag{3.11}$$

The optimal parameters can be found by setting $\nabla_{\boldsymbol{\theta}} \mathcal{L}_Q = \mathbf{0}$ hence

$$n_i = \sum_{j=1}^{n} \gamma_{ji} \tag{3.12}$$

$$n = \sum_{i=1}^{k} n_i \tag{3.13}$$

$$\pi_i = \frac{n_i}{n} \tag{3.14}$$

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n} \gamma_{ji} x_j \tag{3.15}$$

$$\sigma_i^2 = \frac{1}{n_i} \sum_{j=1}^{n} \gamma_{ji} (x_j - \mu_i)^2 \tag{3.16}$$

The algorithm is ran iteratively until the values stop to change. A pictorial representation of the results of EM on Gaussian Mixtures for 2D Gaussians is shown in Figure 3.1 Hence the Classical EM algorithm can be summarized via the following procedure

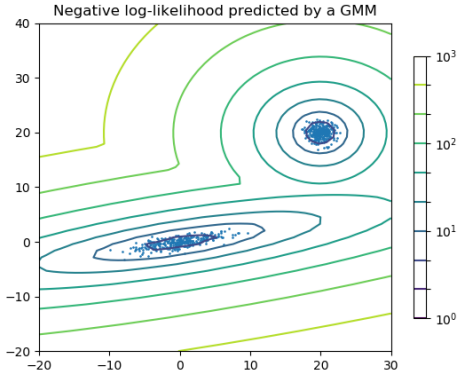


Figure 3.1: Expectation-Maximization for a Mixture of $k = 2$ Gaussians.

$$\text{E-Step: } \mathcal{L}_Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_0) \log p(\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{\theta}) \tag{3.17}$$

$$\text{M-Step: } \boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \tag{3.18}$$

The procedure runs iteratively until it (empirically) converges. Although at first glance, the procedure seems to converge empirically and provide good fits for the desired parameters it has not been proven analytically and remains a demanding open problem. A recent result by Daskalakis, Tzamos and Zampetakis [22] has given global convergence guarantees for the EM algorithm for a Gaussian Mixture Models.

To prove the correctness of the EM procedure we first prove the following lemma

Lemma 1. *Let $p(x)$ and $q(x)$ be two probability distributions defined over a domain A with cross-entropy*

$$\mathcal{H}(p||q) = \mathbb{E}_{q(x)} \left[-\log p(x) \right] \tag{3.19}$$

Then the cross-entropy attains a minimum exactly when $p(x) = q(x)$ for all $x \in A$

Proof. For this optimization problem we define the Lagrangian function with respect to $p(x)$ as

$$L(\lambda, p) = \int_A -q(x) \log p(x) dx + \lambda \left(\int_A p(x) dx - 1 \right) \tag{3.20}$$

Since the minimization should be done subject to the normalization constraint. Differentiation with respect to $p(x)$ yields

$$\int_A \left[-\frac{q(x)}{p(x)} + \lambda \right] dx = 0 \tag{3.21}$$

The integral should be 0 for every x and every integrand, therefore the integrand must be 0. So

$$q(x) = \lambda p(x) \tag{3.22}$$

Integrating with respect to x both sides leaves out $\lambda = 1$ and therefore $p(x) = q(x)$. Furthermore, the second derivative of the entropy is $p(x)/q^2(x) > 0$, hence the function is convex. Therefore the minimum value is attained. The same holds when $p(\boldsymbol{x}), q(\boldsymbol{x})$ are functions of multiple variables. □

Using the above result we can now state the correctness of the EM algorithm

Theorem 1. *If $\mathcal{L}_Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ increases at every iteration then the actual likelihood $\mathcal{L}(\boldsymbol{\theta})$ increases.*

Proof. By using the above lemma directly we have that $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \geq \mathcal{H}(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0)$. Using the relation between ELBO and the actual likelihood and this inequality we arrive at

$$\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}_0) \geq \mathcal{L}_Q(\theta|\boldsymbol{\theta}_0) - \mathcal{L}_Q(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) \tag{3.23}$$

□

Variational EM (Mean-field). When the posterior function $p(\boldsymbol{Z}|\boldsymbol{X})$ is either difficult to be computed online for the iterative maximization proceedure or the resulting equations, the use of a general *variational distribution* $Q(\boldsymbol{Z})$ is called into action. More specifically, inspired by work on statistical physics, the idea of *Variational EM* or *Mean-field Variational Inference* was introduced in [50, 101]. According to this method the latent variables are independent with respect to each other and are distributed with PDFs Q_i parametrized by variational parameters $\boldsymbol{\phi}_i$ that is

$$Q(\boldsymbol{Z}) = \prod_i Q_i(Z_i|\boldsymbol{\phi}_i) \tag{3.24}$$

such that $Q(\boldsymbol{Z})$ approaches the *true posterior distribution* in the *statistical sense*, namely their Kullback-Leibler Divergence[1] approaches zero. An example of Variational EM appears for parameter learning of the Multiplicative Attribute Graph Model [55, 54], where each user $u \in V$ of a social network $G(V, E)$ has a d-dimensional binary feature vector $\boldsymbol{F}_u = (F_{u1}, \ldots, F_{ud})$ where each coordinate follows $\text{Be}(\mu_i)$ independently from the other coordinates and the probability that two users u and v are connected is given by

$$p(u, v) = \prod_{i=1}^{d} \Theta_i[F_{ui}, F_{vi}] \tag{3.25}$$

where $\{\Theta_i\}_{1 \leq i \leq d}$ is a family of real valued 2×2 matrices with components less than 1. The authors choose a variational distribution such that $F_{ui} \sim \text{Be}(\phi_{ui})$, that is

[1] For two probability distributions $p(x), q(x)$ defined on A where $q(x) \neq 0$ the Kulback-Leibler Divergence is defined as

$$D(p||q) = \int_{x \in A} p(x) \log \frac{p(x)}{q(x)} dx$$

It can be proven that the Kullback-Leibler divergence is 0 iff $p(x) \equiv q(x)$ for all $x \in A$.

$$Q = \prod_{u \in V} \prod_{i=1}^{d} Q_{ui}(F_{ui}) \tag{3.26}$$

After each expectation step where the ELBO $\mathcal{L}_Q$ is optimized with respect to ϕ_{ui} using gradient descent, the actual parameters μ_i are updated by freezing ϕ_{ui} and optimizing with respect to μ_i, where the likelihood to be optimized is

$$\mathcal{L}_{Qi}(\mu_i) = \sum_{u \in V} \mathbb{E}_{Q_{ui}} \left[\log p(F_{ui}|\mu_i)\right] = \sum_{u \in V} \left[\phi_{ui} \log \mu_i + (1 - \phi_{ui}) \log(1 - \mu_i)\right] \tag{3.27}$$

which attains a maximum when

$$\mu_i = \frac{1}{|V|} \sum_{u \in V} \phi_{ui} \tag{3.28}$$

Our contribution utilizes the Variational EM approach in a similar manner for performing inference in Stochastic Opinion Dynamics models.

Pseudo EM. Many times, good fits can be found via a much simpler procedure. Instead of calculating $\mathcal{L}_Q = \mathbb{E}_{Q(\boldsymbol{Z})} \left[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})\right]$ one can complete the missing data $\boldsymbol{Z}$ directly in the actual likelihood with their expected values given the old parameters $\boldsymbol{\theta}_0$ and then maximize with respect to the new parameters $\boldsymbol{\theta}$. This approach is known as *Pseudo EM* and is discussed in references [42], [97] and [70]. Formally

$$\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \log p\left(\boldsymbol{X}, \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}_0)} \left[\boldsymbol{Z}\right]\right) \tag{3.29}$$

To adduce an example, consider a set $D = D_{good} \cup D_{bad}$ of n samples which follow $\mathcal{N}(\mu, 1)$, where μ has to be estimated. The set D_{good} contains observed samples and the set D_{bad} comprises only of missing samples. We first complete the bad data with their old expected value $\mu^{(t)}$ and then compute the MLE of the completed data. From basic statistics, we know that

$$\mu^{(t+1)} = \frac{1}{n} \sum_{x \in D} x = \frac{1}{n} \sum_{x \in D_{good}} x + \frac{1}{n} \sum_{x \in D_{bad}} x = \frac{1}{n} \sum_{x \in D_{good}} x + \frac{n_{bad}}{n} \mu^{(t)} \tag{3.30}$$

In the steady state $\mu^{(t+1)} = \mu^{(t)} = \mu^*$ and therefore $\mu^* = \frac{1}{n_{good}} \sum_{x \in D_{good}} x$. Even though this trivial example serves to demonstrate the Pseudo EM technique — albeit the result is trivial — the actual strength can be observed for example in samples from a Multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where at each coordinate samples are missing at different locations.

Chapter 4

Opinion Dynamics

"It is better to change an opinion than to persist in a wrong one."
— Socrates

4.1 Motivation

In everyday life, people tend to communicate with one another by exchanging information and shaping their opinions. Given the tremendous size of social networks, interactions upon which the peoples' opinions change are mostly local meaning that each person has his/her own dynamical social circles discussion with which results in steering opinions. With the general term "opinions" we mean almost everything that can be modeled as a real number in $[0, 1]$. Examples are numerous: tendency to voting certain candidates, opinions about controversial topics, likeliness of publishing papers at certain venues and trash talking. In this context, the various people (or agents) of the network discuss until they reach a *consensus state* where their opinions do not change in the future.

Opinion dynamics processes have been extensively studied in scientific literature. The most influential models are the ones due to Friedkin and Johnsen (FJ) [38], DeGroot [25] and Hegselmann and Krausse (HK) [46]. Alternations of these models have also been introduced — such as in [35, 36, 2] — both from an optimization perspective and a dynamical systems perspective. Below we are going to present some main models in opinion dynamics as well as results regarding their properties.

4.2 Models of Opinion Dynamics

Notational Conventions and Basic Definitions. We consider a system U of $|U| = n$ agents each of which is associated with a function $x_u^{(t)} \in [0, 1]$ which represents his/her opinion. The vector of all agents at time $t \geq 0$ is denoted by $\boldsymbol{x}^{(t)}$. We say that a model converges asymptotically to a point $\boldsymbol{x}^*$ — which is called a *consensus* — if and only if $\lim_{t\to\infty} \|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_\infty = 0$. Each agent $u \in U$ has a dynamical neighborhood $N^{(t)}(u)$ and a transition function (model) which aggregates information from each agent $v \in N^{(t)}(u)$

in order to produce the next state, that is

$$x_u^{(t+1)} = f\left(\left\{x_v^{(t)}\right\}_{v\in N^{(t)}(u)}\right) \tag{4.1}$$

Equivalently, an opnion dynamics model can be seen as a series of graphs $\{G^{(t)}\}$ where each graph $G^{(t)}(U, E^{(t)})$ has an edge set $E^{(t)} = \bigcup_{u\in U} N^{(t)}(u)$. Convergence to consensus is similary defined in the combinatorial sense such that $\lim_{t\to\infty} \left|E^{(t)} \ominus E^*\right| = 0$ where $\ominus$ denotes the symmetric difference between two sets, that is $A \ominus B = (A \setminus B) \cup (B \setminus A)$. Having defined the above notions, we are now ready to explore some fundamental models of opinion dynamics.

The DeGroot Model [25]. In the DeGroot Model we are given a stochastic matrix P and an update rule of

$$\boldsymbol{x}^{(t+1)} = P\boldsymbol{x}^{(t)} \tag{4.2}$$

of the agents' opinions. Note that the update rule is identical to a Markov Chain and hence — from Markov Chains theory — we know that if the matrix P is aperiodic and irreducible then the model converges to $\boldsymbol{x}^*$. The convergence rate of the DeGroot model is associated with the second largest eigenvalue of P, that is the total variation distance decreases as $O(\lambda_2^t)$ where $\lambda_2 < 1$ is the second largest eigenvalue of P.

The Friedkin-Johnsen (FJ) Model [38]. In the FJ model, there is an underlying weighted and undirected social network $G(V, E)$ with $|V| = n$ nodes. An edge between two agents i and j exists in case the agents know one another in some way. Each agent's opinion $x_u^{(t)}$ is a real number in $[0, 1]$ and each agent has an initial opinion about a matter, which he does not change. Each edge has a non-negative weight w_{ij} associated with the strength of communication between the agents i and j. Two agents do not communicate if and only if $w_{ij} = 0$. At each round $t + 1$ the agent updates his/her opinion using information from round t as

$$x_u^{(t+1)} = \frac{\sum_{v\neq u} w_{uv} x_v^{(t)} + w_{uu} s_u}{\sum_{v\in V} w_{uv}} \tag{4.3}$$

The above process can be represented by a stochastic matrix $\hat{W}$ such that each row is normalized by its sum, that is

$$\hat{w}_{uv} = \frac{w_{uv}}{\sum_{v\in V} w_{uv}} \qquad u \in V \tag{4.4}$$

We define the matrices A and B such that

$$b_u = \hat{w}_{uu} \qquad A_{ij} = \hat{w}_{ij}(1 - \delta_{ij}) \tag{4.5}$$

where δ_{ij} is the Kronecker delta function which is 1 if and only if $i = j$ and 0 otherwise. The system can be rewritten in vector notation as

$$\boldsymbol{x}^{(t+1)} = A\boldsymbol{x}^{(t)} + B\boldsymbol{s} \tag{4.6}$$

The initial opinions of the agents — namely $\boldsymbol{s}$ — are constant in each iteration and repeated averaging will not bring all the agents to the same opinion. By that construction the intrinsic beliefs of each agent as well as the connections with one another create the basis for reaching different opinions in the consensus state. By construction, the system is GAS since A is substochastic — hence it has a spectral radius[1] — strictly less than 1 — and $w_{uu} < 1$ and the consensus point is equal to

$$\boldsymbol{x}^* = (I - A)^{-1}B\boldsymbol{s} \tag{4.7}$$

where $|I-A| \neq 0$ since at consensus $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t+1)} = \boldsymbol{x}^*$. Observe that the final opinions for each agent are given as a linear combination of the initial opinions of the agents. The FJ model has also been studied from a game-theoretical perspective by Bindel et al. [12]. They posed the question of how someone would asses the cost of not reaching a consensus point. For that reason, they assumed that the update rule is derived as an optimal point to the following *social cost* function

$$C_u(t) = C_u(x_u^{(t+1)}, \boldsymbol{x}_{-u}^{(t)}) = \sum_{v \in N^{(t)}(u)} w_{uv}\left(x_v^{(t)} - x_u^{(t+1)}\right)^2 + w_{uu}\left(x_u^{(t+1)} - s_u\right)^2 \tag{4.8}$$

where each agent u is a selfish agent. We are interested in studying what is the optimal strategy of each agent given that the opinions of the other agents do not change. By setting $\frac{\partial C_u(t)}{\partial x_u^{(t+1)}} = 0$ leads to the FJ update rule. From a stochastic viewpoint, the quadratic disagreement cost can be viewed as *likelihood maximization* for means of Gaussian opinions. We will come back later to this point, when we will study stochastic opinion dynamics when extra samples are added to the likelihood as regularization. The function $C_u(t)$ is indeed convex and it can be shown that a Nash Equilibrium — namely a state when each player is not benefited from changing strategy given that the other players have fixed strategies is reached upon reaching $\boldsymbol{x}^* = (I-A)^{-1}B\boldsymbol{s}$. Indeed, the Nash Equilibrium in this setting is defined as

Definition 2 (Nash Equilibrium (NE))**.** *Let $\boldsymbol{x}$ be a set of opinions for all the agents. Then $\boldsymbol{x}$ is a Nash Equilibrium if for every player u and for every strategy $y \in [0, 1]$ the following holds*

$$C_u(y, \boldsymbol{x}_{-u}) \geq C_u(x, \boldsymbol{x}_{-u}) \tag{4.9}$$

where $\boldsymbol{x}_{-u}$ denotes the opinions of the agents, excluding agent u.

[1] The spectral radius of a matrix A is defined as its maximum eigenvalue

Algorithm 1 Best response dynamics

Every player has initial opinion $x_u^{(0)}$

$t \leftarrow 1$

while consensus is not reached **do**

Every player update his/her opinion via

$$x_u(t) = \text{argmin}_{x\in[0,1]} C_u(x, \boldsymbol{x}_{-u}^{(t-1)})$$

$t \leftarrow t + 1$

end while

Recalling this definition, we can easily show that if every agent chooses the best response (cost minimizer) regarding his/her individual cost $C_u(t)$ then the system reaches a NE at the steady state. This class of dynamic behaviours — namely when everyone chooses the strategy that minimizes his/her cost — are called *best response dynamics*. We give the general description of the best response dynamics at Algorithm 1.

Of course, the best response strategy can generate multiple opinion dynamics models if one takes into account that the weights $w_{uv} = w_{uv}^{(t)}$ are a function of time. Besides, one can extend this philosophy to address problems that involve limited information [35], that is problems at which each agent u cannot attend to his/her whole neighborhood and instead chooses to one of the agents — WLOG let him/her be v — in his/her neighborhood v with probability

$$p_{uv} = \frac{w_{uv}}{\sum_{z\in N(u)} w_{uz}} \tag{4.10}$$

Then the agent suffers a cost due to v equal to

$$(1 - a_u)(x_u - x_v)^2 + a_u(x_u - s_u)^2 \tag{4.11}$$

where $a_u = \frac{w_{uu}}{\sum_{v\in N(u)} w_{uv}}$, where the neighborhood includes u as well. The expected cost is equal to the cost in the deterministic case and therefore under expectation the NE is reached under the same stationary state $\boldsymbol{x}^* = (I - A)^{-1}B\boldsymbol{s}$. Therefore, we can define the stochastic game $I = (P, \boldsymbol{s}, \boldsymbol{a})$ of n agents as the game defined by

- P which is a row-stochastic matrix with entries p_{uv}
- $\boldsymbol{s} \in [0, 1]^n$ which is the vector of internal/initial opinions
- $\boldsymbol{a} \in (0, 1]^n$ which is the vector of self-confidence with entries a_u

The authors of [35] study the Follow the Leader (FTL) strategy which is a classical game-theoretical strategy based on "play the best you have observed" motto. Letting $W_{uv}^{(t)}$ be the set of random variables for which "u meets v at time t" parametrized by $p_{uv}^{(t)} = p_{uv}$ such that $\sum_v W_{uv}^{(t)} = 1$ the FTL strategy states an update rule of the form

$$x_u^{(t+1)} = \text{argmin}_{x\in[0,1]} = \sum_{\tau=0}^{t}(1-a_u)\left(x - x^{(\tau)}_{W_u^{(\tau)}}\right)^2 + a_u(x-s_u)^2 \tag{4.12}$$

Finanlly, the authors show that for the $I = (P, \boldsymbol{s}, \boldsymbol{a})$ stochastic game the FTL strategy satisfies

$$\mathbb{E}\left[\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\|_\infty\right] \leq C\sqrt{\log n}\frac{(\log t)^{3/2}}{t^{\min(1/2,\rho)}} \tag{4.13}$$

where $\rho = \min_u a_u$ and C is a constant. For a more detailed explaination of the work, we redirect the interested reader to [51].

The Hegselmann-Krausse Model (HK) [46]. In the Hegselmann-Krausse (HK) model each agent constructs his/her influence neighborhood

$$N^{(t)}(u) = \left\{v \in U \mid \left\|x_u^{(t)} - x_v^{(t)}\right\| \leq \varepsilon\right\} \tag{4.14}$$

and updates his/her opinion according to the following rule

$$x_u^{(t+1)} = \frac{1}{|N^{(t)}(u)|}\sum_{v\in N^{(t)}(u)|} x_v^{(t)} \tag{4.15}$$

A more general formulation of the HK model attributes a different radius ε_u for each agent u. The model is shown to converge in reference [46] using combinatorial arguments. However, recent work in [81] has shifted attention towards treating such models using control systems theory. Moreover, recent work done in [36] has introduced two new models related to the classical HK model.

The Network-HK Model [36]. An underlying network structure $H(V, E)$ is added such that the neighborhoods of the HK now become

$$N^{(t)}(u|H) = \left\{v \in U \mid \left\|x_u^{(t)} - x_v^{(t)}\right\| \leq \varepsilon \cap \{u, v\} \in E(H)\right\} \tag{4.16}$$

The update rule remains the same. This model is also shown to converge, in the same paper.

The Random-HK Model [36]. The model is similar to the HK model however now we choose a random subset of k elements from the neighborhood and aggregate the results. The convergence result is now proven using randomized analysis such that the system reaches consensus under expectation.

General Averaging Dynamics. Moreover, the form of the HK model inspires a whole family of dynamics with zero-input which are based on a similar "smoothing" procedure. We give the formal definition below

Definition 3 (Averaging Dynamics)**.** *An opinion dynamics model is said to belong to the class of Averaging Dynamics if and only if it has an update rule of*

$$x_u^{(t+1)} = \frac{1}{|N^{(t)}(u)|} \sum_{v \in N^{(t)}(u)} x_v^{(t)} \tag{4.17}$$

for every agent $u \in U$*.*

This class of dynamics is very interesting since it serves as a generalization of the classical moving average filters from signal processing.

4.3 Coevolutionary Opinion Formation Games

The family of game-theoretic models of opinion formation that are most related to ours is the *Coevolutionary Opinion Formation Games* (COFG) introduced in [11]. According to COFG, the agents of the network evolve their opinions along with their neighborhoods (such as in the HK model). This extends the work of Bindel et al. [12] which seeks the minimization of the disagreement cost of agents, but with the network fixed, eventually arriving at a game-theoretical understanding of the FJ model. The authors of [11] generalize the social cost function imposed by Bindel et al. and tightly bound the Price of Anarchy (PoA)[2] and interpret it it as a way to attribute value to how much nodes value their intrinsic and their friends' opinions.

In a CG, there are n players each of which has an intrinsic opinion s_i and expresses an opinion z_i (where in general $s_i \neq z_i$). Each player's goal is to minimize the cost $C_i(\boldsymbol{z})$ which is a function of s_i and the expressed opinions $\boldsymbol{z} = (z_1, \ldots, z_n)$ of all players. The cummulative social cost is defined as $C(\boldsymbol{z}) = \sum_{i=1}^{n} C_i(\boldsymbol{z})$. The authors consider two games. The former one is the *symmetric CG* where each player's cost function is given as

$$C_i(z_i, \boldsymbol{z}_{-i}) = \sum_{j \neq i} f_{ij}(z_i - z_j) + w_i g_i(z_i - s_i) \tag{4.18}$$

where f_{ij} and g_i are real (fixed) valued functions that are convex, continuously differentiable and symmetic, that is $f_{ij}(-x) = f_{ij}(x)$ and $g_i(x) = g_i(-x)$ and $g(0) = 0$. In the symmetric setting $f_{ij} = f_{ji}$ which makes the game symmetric wrt to pairs of players. The work of Bindel et al. sets $g(x) = x^2$ and $f_{ij}(x) = w_{ij}x^2$ where $w_{ij} = w_{ji}$ represent the weight of the edge $\{i, j\}$ between players i and j. The existence of a unique pure NE can be shown via the potential function

[2]The PoA of a game measures how the efficiency of a system degrades due to selfish behavior of its agents. It is a general notion that can be extended to diverse systems and notions of efficiency. Given a game $G = (\mathcal{N}, \mathcal{S}, u)$ with a set $\mathcal{N}$ of players strategy sets $\mathcal{S}$ for each player $i \in \mathcal{N}$, utility functions $u_i : \mathcal{S} \to \mathbb{R}$,a welfare function $W : \mathcal{S} \to \mathbb{R}$, such as the utilitarian objective $W(s) = \sum_{i \in \mathcal{N}} u_i(s)$ or the egalitarian objective $W(s) = \min_{i \in \mathcal{N}} u_i(s)$, and a set $\mathcal{E} \subseteq \mathcal{S}$ of equilibria, the PoA is defined as $\text{PoA} = \frac{\max_{s \in \mathcal{S}} W(s)}{\min_{e \in \mathcal{E}} W(e)}$.

$$\phi(\boldsymbol{z}) = \sum_i w_i g_i(z_i - s_i) + \sum_{i<j} f_{ij}(z_i - z_j) \tag{4.19}$$

Moreover, to extract the PoA bound the authors consider the set

$$\mathcal{H}_{x,y,f} = \left\{ (\lambda, \mu) \middle| f(x) + \frac{y-x}{2} f'(x) \leq \lambda f(y) + \mu f(x) \text{ for all} x, y \geq 0,\ f \text{ is a weight function} \right\} \tag{4.20}$$

the set

$$\mathcal{H}_{u,v,g} = \left\{ (\lambda, \mu) \middle| g(u) + (v-u) g'(u) \leq \lambda g(v) + \mu g(u) \text{ for all} x, y \geq 0,\ g \text{ is a weight function} \right\} \tag{4.21}$$

and the sets $\mathcal{A}_1, \mathcal{A}_2$ which are given as $\mathcal{A}_1 = \bigcup_f \mathcal{H}_{x,y,f}$ and $\mathcal{A}_2 = \bigcup_g \mathcal{H}_{u,v,g}$. Finally the authors show that for any $(\lambda, \mu) \in \mathcal{A}_1 \cap \mathcal{A}_2$ the value $\lambda/(1-\mu)$ is an upper bound onf the PoA and $\zeta = \min_{(\lambda,\mu)\in\mathcal{A}_1\cap\mathcal{A}_2} \frac{\lambda}{1-\mu}$ is the best upper bound.

Their proof is based on the technique of *Local Smoothness* introduced by Roughgarden and Schoppmann [92] to which we make a quick reference. For each function C_i one has to prove that for $\mu < 1,\ \lambda > 0$ and for every $\boldsymbol{z}$, and for a fixed profile $\boldsymbol{o}$ that

$$\sum_i C_i(z_i, \boldsymbol{z}_{-i}) + (o_i - z_i) \frac{\partial C_i(z_i, \boldsymbol{z}_{-i})}{\partial z_i} \leq \lambda C(\boldsymbol{o}) + \mu C(\boldsymbol{z}) \tag{4.22}$$

Then the authors make use of the following result of [93] to prove their PoA bounds

Theorem 2. *Let σ denote a correlated equilibrium. If (4.22) holds for every outcome $\boldsymbol{z}$ with respect to a fixed outcome $\boldsymbol{o}$ then the ratio $\mathbb{E}_{\boldsymbol{z}\sim\sigma}[C(\boldsymbol{z})]$ to $C(\boldsymbol{o})$ is at most $\lambda/(1-\mu)$. If $\boldsymbol{o}$ is the optimal outcome then the PoA is at most $\lambda/(1-\mu)$.*

When the functions are convex and differentiable the PoA bound is always at most 2. Finally, the authors provide a general lower bound construction for the symmetic CG.

In the K-NN CG each player looks at her k nearest neighbors (with consistent tie-breaking) with respect to s_i and forms the set $K(\boldsymbol{z}, i)$ and suffers a cost of

$$C_i(z_i, \boldsymbol{z}_{-i}) = \sum_{j \in K(\boldsymbol{z}, i)} (z_j - z_i)^2 + \alpha k (z_i - s_i)^2 \tag{4.23}$$

The authors show that the K-NN game has a PoA of at most a constant for $\alpha > 1$, where the constant improves together with the increase of α. The social outcomes become better when nodes are "narrow minded" and give larger weight to their opinions ($\alpha \to \infty$). Contrary to Bindel et al. the authors show show that if nodes can choose their neighbors based on their k nearest neighbors the the PoA can be bounded. Finally, the authors show that for small α the PoA is at least $1/\alpha^2$, which explains why PoA deteriorates upon the agents being more "broad-minded".

The connection to our work is the cost formulation as *negative log-likelihood*. Indeed, in the stochastic case of Chapter 7 with agents having stochastic opinions the negative log-likelihood cost resembles the costs introduced in [12, 11]. More specificaly, our work considers binary opinions over a latent setting and the maximization of the ELBO (after removing the double stochasticity) reduces to a cost similar to its game-theoretic counterpart. Finally, when the agents have Gaussian opinions with unity covariance matrix and the k-nearest neighbors are considered with respect to the expressed opinions z_i, the cost function of (4.23) is equivalent to the negative log-likelihood of the NNIM model.

Chapter 5

Dynamical Systems on the Steady State

5.1 Lyapunov Stability and Lyapunov Functions

The study of dynamical systems the last century has been of prime importance in mathematical and engineering sciences. As a computer scientist, I have one more reason to care about: algorithms. It is without doubt that dynamical systems and algorithms are two faces of the same coin. On the one hand, a (discrete) dynamical system is defined as

$$\boldsymbol{x}^{(t+1)} = f(\boldsymbol{x}^{(t)}, \boldsymbol{u}(t)) \tag{5.1}$$

Where f is a transition function, $\boldsymbol{x}^{(t)}$ is the state of the system and $\boldsymbol{u}(t)$ is the input of the system which we usually apply from outside. If the temporal variable t is continuous, the system is similarly

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}^{(t)}, \boldsymbol{u}(t)) \tag{5.2}$$

Similarly, an algorithm $\mathcal{A}$ is usually an iterative process f that evolves over discrete time with the prospect of reaching some solution. Such algorithms are primarily learning ones, with the most famous of them being the Steepest Descent [98] update that is

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta(t)\nabla f(\boldsymbol{x}^{(t)}) \tag{5.3}$$

for finding the local optima of a function f. Steepest Descent updates also occur when minimizing functionals of an input function, through the involvement of Euler Derivatives. These problems usually arise in computer vision in terms of smoothing algorithms, such as Gaussian smoothing or the infamous Perona-Malik anisotropic diffusion process [102]. On the other hand, classical dynamical systems refer usually to electro-mechanical systems and usually take the linear form of

$$\boldsymbol{x}^{(t+1)} = A(t)\boldsymbol{x}^{(t)} + B(t)\boldsymbol{u}(t) \tag{5.4}$$

where $A(t), B(t)$ are known matrices. While the exact behaviour of the dynamical system is usually of little interest, most care about the limiting behaviour of a system. For example, an electrical system is observed and measured after an adequate amount of time has passed since its initialization which is called the *steady state*. Obviously, systems which tend to attain infinitely large values in the steady state are not interesting in any way, and the main goal of control systems theory is to counterbalance the systems' uncontrolled behaviour towards instability with a controller, so that a final objective can be achieved subject to the control law $\boldsymbol{u}(t)$.

For that reason, Aleksander Lyapunov dedicated his Doctoral Thesis [68] in the study of the stability of dynamical systems. His work primarily focused on the study of the stability of non-linear systems through analytical methods. Lyapunov's Method is based on the differential equation that describes the dynamical system and gives out information about its behaviour without the need for analytically solving the initial differential equation. Before reciting Lyapunov's method we are going to introduce some definitions about stability of dynamical systems. For that reason, we, for now, assume that a system is described by the model

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}^{(t)}) \tag{5.5}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{5.6}$$

with a solution of $\phi(t, \boldsymbol{x}_0)$. We now define the *equilibrium point* to be a root of the function f.

Definition 4 (Equilibrium Point)**.** *The point $\boldsymbol{x}^*$ is called an equilibrium point of Eq. 5.5 iff $f(\boldsymbol{x}^*) = 0$.*

For instance in the system $\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}^{(t)}$ has a unique equilibrium at $\mathbf{0}$ if and only if $|A| \neq 0$. Next, we give a general definition of stability

Definition 5 (General Definition of Lyapunov Stability)**.** *The equilibrium point $\boldsymbol{x}^*$ is said to be stable iff for every $\varepsilon > 0$ there exists some $\delta = \delta(\varepsilon)$ such that if $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq \delta$, then $\|\phi(t, \boldsymbol{x}_0) - \boldsymbol{x}^*\| \leq \epsilon$ for all $t \geq 0$.*

Furthermore, we are interested on the system behaviour as $t \to \infty$, i.e. study if the solution $\boldsymbol{x}^{(t)}$ is sufficiently near $\boldsymbol{x}^*$. We define a system for which the solution is sufficiently near the equilibrium point — regarless of the initial condition $\boldsymbol{x}_0$ — for all t and additionally $\boldsymbol{x}^{(t)} \to \boldsymbol{x}^*$ as $t \to \infty$.

Lyapunov's idea for assesing the (asymtotic) stability of dynamical systems was to define a "generalized energy" of the system $V(\boldsymbol{x}^{(t)})$ which will gradually decrease[1] until the system reaches the equilibrium point $\boldsymbol{x}^*$. The function $V(\boldsymbol{x}^{(t)})$ is a scalar function of the state vector $\boldsymbol{x}^{(t)}$. We below give the formal definition of a Lyapunov Function

[1] For a stable system

Definition 6 (Lyapunov Function)**.** *The not time-varying Lyapunov Function V satisfies the following conditions for all $t > 0$ and for all $\boldsymbol{x}$ near $\mathbf{0}$ where $\boldsymbol{x}^* = \mathbf{0}$ is an equilibrium point.*

1. $V(\boldsymbol{x}) \in C^1(\mathbb{R})$.
2. $V(\mathbf{0}) = 0$.
3. $V(\boldsymbol{x}) > 0$ *for all* $\boldsymbol{x} \neq \mathbf{0}$.
4. $\dot{V}(\boldsymbol{x}) < 0$ *for all* $\boldsymbol{x} \neq \mathbf{0}$ *where* $\dot{V}(\boldsymbol{x}) = \nabla V(\boldsymbol{x})^T \dot{\boldsymbol{x}}$.

Note that for a discrete system the only modification we need to consider is $V(\boldsymbol{x}^{(t+1)}) < V(\boldsymbol{x}^{(t)})$ for all $\boldsymbol{x}^{(t)} \neq 0$ as the condition for the monotonicity of the Lyapunov Function and leave the rest conditions the same. We are now ready to present the main result of Lyapunov [68]

Theorem 3 (Lyapunov's Theorem (GAS))**.** *Let $\dot{\boldsymbol{x}} = f(\boldsymbol{x})$ be a dynamical system with equilibrium point $\boldsymbol{x}^* = \mathbf{0}$ such that a Lyapunov Function V can be determined. Then the equilibrium point $\boldsymbol{x}^* = 0$ is globally asymptotically stable (GAS).*

Example 1: Warm-up. As a warm-up example, consider the system

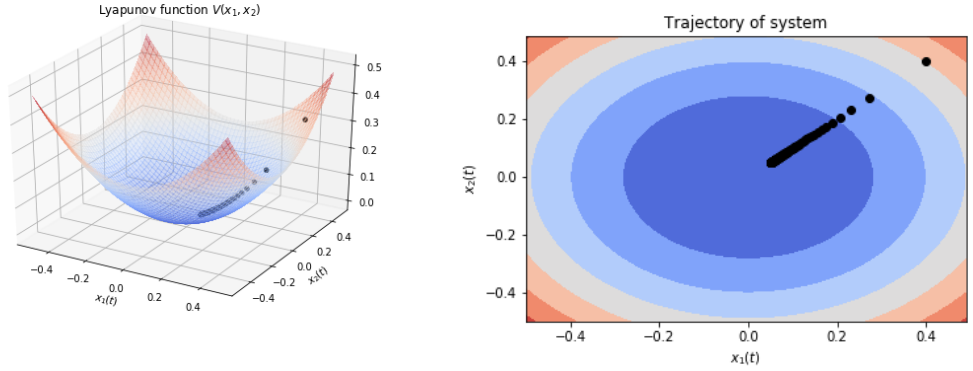
$$\dot{x}_1 = x_2 - x_1^3 - x_1 x_2^2 \tag{5.7}$$
$$\dot{x}_2 = -x_1 - x_1^2 x_2 - x_2^3 \tag{5.8}$$

and the function $V = x_1^2 + x_2^2$. Obviously, the point $(x_1, x_2)^T = (0, 0)^T$ is an equilibrium point. The function V satisfies the first three properties of the definition for a Lyapunov function and moreover $\nabla V = (2x_1, 2x_2)$. Plugging in the system definition and doing the algebra we arive at $\dot{V} = -2(x_1^2 + x_2^2)^2 < 0$ for all $x_1, x_2 \neq 0$. Therefore the system is GAS.

Determining Lyapunov Functions. Even though at first glance, the method is descent for determining the stability of a dynamical system, yet finding an appropriate Lyapunov Function is usually a very difficult problem. On the one hand, for an LTI system $\dot{\boldsymbol{x}} = A\boldsymbol{x}$ one defines a Lyapunov function $V = \boldsymbol{x}^T P \boldsymbol{x}$ such that $P > 0$ and $A^T P - PA = -Q$ for some Q, where $Q > 0$. On the other hand, as we will investigate later many opinion dynamics systems such as the Hegselmann-Krausse model require much more sophisticated Lyapunov Functions. Some popular methods for determining the Lyapunov Functions is the gradient method [20] and the Krasovskii method [7].

5.2 Markov Chains

One characteristic example of a dynamical system is a Markov Chain. A Markov chain is a stochastic model describing a sequence of possible events in which the probability of



(a) 3D Surface Plot

(b) Contour Plot

Figure 5.1: Lyapunov Function $V(x_1, x_2)$ of the example system. The trajectory $\gamma = \{(x_1, x_2) \mid \dot{x}_1 = x_2 - x_1^3 - x_1 x_2^2,\ \dot{x}_2 = -x_1 - x_1^2 x_2 - x_2^3,\ x_{10} = x_{20} = 0.4,\ t \geq 0\}$ is also provided as a scatter plot. Since $\dot{V} < 0$ for all $(x_1, x_2)^T \neq (0, 0)^T$ the system moves towards the base of the paraboloid $V = x_1^2 + x_2^2$

each event depends only on the state attained in the previous event. Markov Chains are used widely in modeling population dynamics, communication systems and particle interactions. Markov chains are the basis for general stochastic simulation methods known as Markov Chain Monte Carlo (MCMC), which are used for simulating sampling from complex probability distributions, and have found application in Bayesian statistics and artificial intelligence. In this Thesis, we will refer to an important theorem for Markov Chains regarding the convergence time of a Markov Chain. A Markov Chain is modeled by an initial distribution $\boldsymbol{\pi}(0)$ over a state space $\mathbb{X}$ and a stochastic transition matrix P such that the probability of being at a state $x \in \mathbb{X}$ at time t is given by

$$\boldsymbol{\pi}^{(t+1)} = \boldsymbol{\pi}^{(t)} P \tag{5.9}$$

Asymptotic Behaviour of Markov Chains. If the Markov chain is irreducible and aperiodic[2], then there is a unique stationary distribution $\boldsymbol{\pi}$. Our focus is how fast does the sequence $\boldsymbol{\pi}^{(t)}$ converge to $\boldsymbol{\pi}$, hence we are interested in the total variation distance between $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\pi}$ that is

$$d_{TV}(t) = \|\boldsymbol{\pi}^{(t)} - \boldsymbol{\pi}\|_{TV} = \sup_{A \subseteq \mathbb{X}} \left| \sum_{x \in A} \pi(x, t) - \sum_{x \in A} \pi(x) \right| \tag{5.10}$$

In algorithm analysis we are interested in the time such that every subset $\mathbb{A} \subset \mathbb{X}$ of the states is near to the respective stationary distribution with respect to some error tolerance $\eta > 0$. For that reason, we define the *mixing time* of a Markov Chain as the minimum

[2]A Markov chain is said to be irreducible if it is possible to get to any state from any state.
A Markov Chain is aperiodic iff there is only 1 step required to pass from a state twice.

time t_0 such that the probability that the total variation distance between the current probability distribution and the stationary distribution is less than η. More formally

Definition 7 (Mixing Time)**.** *The mixing time $t_0 = t_0(\eta)$ of a Markov Chain is defined as*

$$t_0(\eta) = \inf \left\{ t \geq 0 \mid \left\| \boldsymbol{\pi}^{(t)} - \boldsymbol{\pi}^* \right\|_{TV} \leq \eta \right\} \tag{5.11}$$

Moreover, it has been shown by Perron and Frobenius that the convergence rate of a stationary Markov Chain is exponential with the second largest eigenvalue of the matrix P[3], that is

$$d_{TV}(t) = O(\lambda_2^t) \tag{5.12}$$

Generally, if P is a stochastic matrix with second eigenvalue $\lambda = \frac{1}{\alpha} < 1$ and $t \geq \left\lceil \frac{\log(1/\eta)}{\log \alpha} \right\rceil$ then the total variation distance will be at most η. One particularly interesting case is when the stochastic matrix P is proportional to the adjacency matrix A of a k-regular graph[4] such that $P = \frac{1}{k}A$. Essentially, we are interested to understand the *worst-case* behaviour of such a Markov Chain. That is we want to find the second largest eigenvalue of the "slowest matrix" P^*. From an optimization perspective

$$\begin{aligned}
\text{maximize}_P \quad & \lambda_2(P) && (5.13)\\
\text{subject to} \quad & P \text{ is row-stochastic} && (5.14)\\
& P = \frac{1}{k}A && (5.15)\\
& A \text{ is an adjacency matrix of a } k\text{-regular graph} && (5.16)
\end{aligned}$$

We give the following lemma to simplify the above optimization problem.

Lemma 2. *Let A and B be two square $m \times m$ matrices such that $B = \kappa A$ for some $\kappa \in \mathbb{R}^*$. Then if λ is an eigenvalue of A then $\kappa\lambda$ is an eigenvalue of B.*

Proof. The proof is directly inferred from the characteristic polynomial and the identity $|\kappa A| = \kappa^m |A|$. Therefore

$$\chi_B(\lambda) = |\lambda I - B| = |\lambda I - \kappa A| = \kappa^m \left| \frac{\lambda}{\kappa} I - A \right| = \kappa^m \chi_A \left(\frac{\lambda}{\kappa} \right) \tag{5.17}$$

The result comes from the fact that $\chi_B(\lambda) = 0 \iff \chi_A(\lambda/\kappa) = 0$. □

This simple lemma simplifies the above problem to the following one

$$\begin{aligned}
\text{maximize}_A \quad & \lambda_2(A) && (5.18)\\
\text{subject to} \quad & A \text{ is an adjacency matrix of a } k\text{-regular graph} && (5.19)
\end{aligned}$$

[3]The eigenvalues are less or equal to 1. If a chain is stationary that implies that $\boldsymbol{\pi} = \boldsymbol{\pi}P$ therefore P has its largest eigenvalue equal to 1. Hence the convergence rate is specified by the second largest eigenvalue

[4]In a k-regular graph every node has exactly k neighbors.

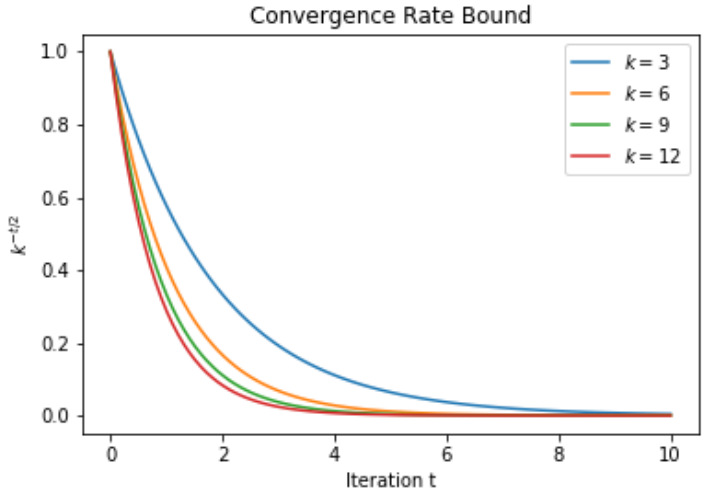


Figure 5.2: Worst-Case Convergence Behaviour of a Markov Chain with a k-regular transition Matrix for various values of $k \geq 3$.

The above problem was posed by Alon[5] and eventually solved by Friedman in [39] which states that the largest possible eigenvalue is $2\sqrt{k-1} + o(1)$ for $n \to \infty$. For completeness purposes, we cite the Chernoff-type Bound proved by Friedman in [39]:

Theorem 4 (Alon-Friedman's Theorem)**.** *Let A be the adjacency matrix of a k-regular graph G with $k \geq 3$ and second largest eigenvalue λ_2. Let $\varepsilon > 0$ be any positive number and $\tau = \lceil (\sqrt{k-1} + 1)/2 \rceil - 1$. Then*

$$\Pr[\lambda_2 \leq 2\sqrt{k-1} + \varepsilon] \geq 1 - O(n^{-\tau}) \tag{5.20}$$

That is $\lambda_2 = 2\sqrt{k-1} + o(1)$ a.a.s. for $n \to \infty$.

Proof. The proof is lengthy (60 pages) and beyond the scope of this Thesis. We redirect the interested reader to the respective paper. □

Corollary 1. *The total variation distance of a Markov Chain with n states and a stochastic matrix P of a k-regular graph decreases as $o(k^{-t/2})$ for very large n for $k \geq 3$.*

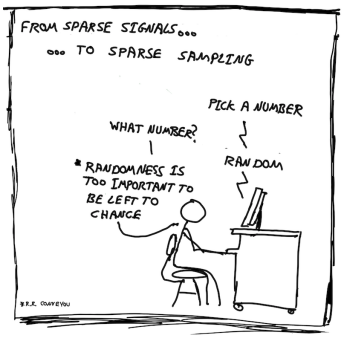
Proof. We apply Alon-Friedman's Theorem to the adjacency matrix $A = kP$ and then scale the eigenvalues. The eigenvalue of the new matrix is $o(\sqrt{k})$ a.a.s. for very large n so the convergence rate follows to be $o(k^{-t/2})$. The behaviour for various values of $k \geq 3$ is shown in Figure 5.2.

□

[5] It remained unsolved for many years.

Chapter 6

Unsupervised Feature Learning



6.1 Dimensionality Reduction

In dimensionality reduction, our aim is to bring data which live in a high-dimensional space to a low dimensional space, while avoiding losing valuable information, and run our algorithms on. Running algorithms in low-dimensional spaces avoids "falling" to the so-called "curse of dimensionality" where algorithms become inefficient in high dimensions. The process of dimensionality reduction is closely tied to the ideas of lossy compression in information theory. In this section, we are going to describe famous methods for performing dimensinoality reduction and are closely related to this dissertation's work. For a more detailed introduction to dimensionality reduction, we redired the interested reader to [98]. Moreover an excellent overview is given at [93].

6.1.1 Principal Components Analysis

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ be vectors in a d dimensional real vector space with zero mean, that is $\bar{\boldsymbol{x}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i = \boldsymbol{0}$. We would like to reduce the dimensions via a linear transformation $\boldsymbol{x} \mapsto W\boldsymbol{x}$ where $\boldsymbol{y} = W\boldsymbol{x}$ is a lower dimensional vector of dimension $d' \ll d$. Moreover, given a compressed vector $\boldsymbol{y}$, we seek a transformation matrix U such that $\tilde{\boldsymbol{x}} = U\boldsymbol{y}$ is d dimensional and the squared error is minimized. Therefore the PCA objective seeks $W \in \mathbb{R}^{d,d'}$, $U \in \mathbb{R}^{d',d}$ such that the objective

$$L_{PCA}(U,W) = \sum_{i=1}^{m} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2^2 = \sum_{i=1}^{m} \|\boldsymbol{x}_i - UW\boldsymbol{x}_i\|_2^2 \tag{6.1}$$

is minimized. From the above objective, it is evident that $W = U^T$ and that $U^TU = I$ since if we fix U, W and let $R = \{UW\boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}^d\}$ then the vectors of R can be represented as $V\boldsymbol{y}$ where $V^TV = I$ and $\boldsymbol{y} \in \mathbb{R}^{d'}$. Hence

$$\|\boldsymbol{x} - V\boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{y}^T(V^T\boldsymbol{x}) \tag{6.2}$$

Setting the gradient wrt to $\boldsymbol{y}$ to zero, we obtain that $\boldsymbol{y} = V^T\boldsymbol{x}$. Therefore for every $\boldsymbol{x}$, the vector that asserts minimum reconstruction error is $VV^T\boldsymbol{x}$. Since this holds for all U and W we conclude that $U^TU = I$ and $W = U^T$. Now the optmization objective simplifies to

$$L_{PCA}(U) = \sum_{i=1}^{m} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_2^2 = \sum_{i=1}^{m} \|\boldsymbol{x}_i - UU^T\boldsymbol{x}_i\|_2^2 \tag{6.3}$$

subject to $U^TU = I$ and $W = U^T$. Expanding the identity $\|\boldsymbol{x} - UU^T\boldsymbol{x}\|_2^2$ inside the objective we get

$$\|\boldsymbol{x} - UU^T\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}\|_2^2 - \text{tr}(U^T\boldsymbol{x}\boldsymbol{x}^TU) \tag{6.4}$$

Hence the minimization objective transforms to the maximization objective

$$\max_{U^TU=I} \text{tr}\left(U^T \sum_{i=1}^{m} \boldsymbol{x}_i\boldsymbol{x}_i^T U\right) \tag{6.5}$$

We now let $A = \sum_{i=1^m} \boldsymbol{x}_i\boldsymbol{x}_i^T$ which is symmetric and can be decomposed as $A = VDV^T$ where $VV^T = I$, the diagonal elements of D are the eigenvalues of A and the columns of V are the eigenvectors of A. Therefore $\text{tr}(U^TAU) = \text{tr}(U^TVDV^TU) = \text{tr}(B^TDB) = \sum_{i=1}^{d} \lambda_j \sum_{j'=1}^{d'} b_{jj'}^2$. But $B^TB = I$ and hence the problem reduces to finding the d' largest eigenvalues of A and the corresponding eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d'}$ which will become the columns of U. Finally $W = U^T$.

6.1.2 Random Projections

Imagine we are given two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and we are asked to find a way such that in a lower dimensional space $\mathbb{R}^{d'}$ with $d' < d$ the Eucledian distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is negligibly affected. In other words if two points are nearby in $\mathbb{R}^d$ they should be nearby in $\mathbb{R}^{d'}$. It turns out that there is a dimensionality reduction technique called the Johnson-Lindenstrauss (JL) transform that achieves that [3, 98, 13]. Assume that we are given a random vector $\boldsymbol{r}$ and define the projection function $f(\boldsymbol{x}|\boldsymbol{r})$ to be $f(\boldsymbol{x}|\boldsymbol{r}) = \langle \boldsymbol{x}, \boldsymbol{r} \rangle$. Clearly the function f is random and if we choose $\boldsymbol{r} \sim \mathcal{N}(\boldsymbol{0}, I)$ then for two vectors $\boldsymbol{x}, \boldsymbol{y}$ we have that

$$g(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r}) = f(\boldsymbol{x}|\boldsymbol{r}) - f(\boldsymbol{y}|\boldsymbol{r}) = \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{r} \rangle \tag{6.6}$$

with mean

$$\mathbb{E}_{\boldsymbol{r}} \left[g(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r}) \right] = 0 \tag{6.7}$$

and variance

$$V_{\boldsymbol{r}}(g(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r})) = \mathbb{E}_{\boldsymbol{r}} \left[g(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r})^2 \right] = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{6.8}$$

Suppose that we repeat (independently) the experiment for d times with vectors $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_d$ where we gen d unbiased estimates of the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. If one uses the concentration bound for the χ^2-variables, then approximating the actual distance within a factor of $1 + \epsilon$ requires setting $d = O(\log n/\epsilon^2)$. More specifically, the final JL transform is defined by a matrix A with entries $a_{ij} \sim \mathcal{N}(0, 1)$ and the mapping $f(\boldsymbol{x}|A) = \frac{1}{\sqrt{d'}} A\boldsymbol{x}$[1], similarly for two vectors $\boldsymbol{x}, \boldsymbol{y}$ we have that

$$\|f(\boldsymbol{x}|A) - f(\boldsymbol{y}|A)\|_2^2 = \frac{\sum_{i=1}^{d'} (a_i^T(\boldsymbol{x} - \boldsymbol{y}))^2}{d'} \tag{6.9}$$

as well as for every $\epsilon \in (0, 3)$ the concentration inequality implies

$$\Pr \left[\frac{\|(1/\sqrt{d'})A\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|^2} > \epsilon \right] \leq 2 \exp(-\epsilon^2 d'/6) \tag{6.10}$$

Hence, for a set of n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ choosing

$$\epsilon = \sqrt{\frac{6 \log(n/\delta)}{d'}} \leq 3 \tag{6.11}$$

asserts that with probability of at least $1 - \delta$ the following holds

$$\sup_{\boldsymbol{x}_i} \left| \frac{\|A\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} - 1 \right| < \epsilon \qquad \forall 1 \leq i \leq n \tag{6.12}$$

Note that the value of ϵ does not depend on the initial dimensionality d of the vectors.

6.1.3 Similarity for Sets: MinHash

One of the very well-known problems in database systems is the one of finding how similar two collections of items A and B are. A reasonable similarity measure is the one of the Jaccard distance $J(A, B) = |A \cap B|/|A \cup B|$ (or the intersection-over-union ratio). In the case of real valued vectors, random projections project from a high dimensional space to a low dimesional space using a random matrix A such that in expectation distance is preserved and hence, due to measure concentration, the original distance is preserved w.h.p.. In the case of sets, one should desire the same but for the Jaccard distance. Having

[1] $f : \mathbb{R}^d \to \mathbb{R}^{d'}$

a way to measure Jaccard in expectation combined with independent trials would yield a similar algorithm for sets. It turns out that there's such algorithm which is called the MinHash of a set that generates a set's signature randomly. First of all, the MinHash algorithm for a set U, chooses a permutation $\pi : U \to U$ of the elements of U usiformly at random (each permutation occurs with probability $\frac{1}{|U|!}$ and maps each set $S \subseteq U$ to $\text{MinHash}(S) = \text{argmin}_{x \in S} \pi(x)$. The MinHash function can serve as an *unbiased* estimator of $A, B \subseteq U$. Via a simple counting argument one can deduce that $\Pr[\text{MinHash}(A) = \text{MinHash}(B)] = J(A, B)$

6.2 Nearest Neighbor Search

One of the fundamental problems in Data Science is the one of finding the nearest (or the k nearest) neighbors of a point $\boldsymbol{x}$, where the possible answers come from a set $A = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^d$ that contains n points. Formally the nearest neighbor problem is defined by a function $q : \mathbb{R}^d \to \mathbb{R}^d$ such that[2]

$$q(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{z} \in A} \{\|\boldsymbol{x} - \boldsymbol{z}\|\} \tag{6.13}$$

where $\|\cdot\|$ is a norm function. From now on, for ease of demonstration we will assume that it is the Eucledian norm. Similar results can be obtained for other norms since norms are equivalent. The problem seems very easy from a first glimpse, however, as we process, we will conclude that actually the opposite is correct: nearest neighbor search is a difficult problem! To give a first scent why this is true, we will lie in the power of the nearest neighbor classifier. A consider K classes $A_1, A_2, \ldots A_K \subseteq R^d$ of points and a classifier that for each query point $\boldsymbol{x}$ assigns the class at which the closest point to $\boldsymbol{x}$ belongs, with ties broken consistently. It turns out that, however large the classes A_1, A_2 are the real space $\mathbb{R}^d$ can be partitioned into two sets D_i for $1 \leq i \leq K$ such that

$$D_i = \left\{ \boldsymbol{x} \in \mathbb{R}^d | \exists \boldsymbol{z} \in A_i, \forall \boldsymbol{y} \in \bigcup_{j \neq i} A_j : \|\boldsymbol{x} - \boldsymbol{y}\| \succeq \|\boldsymbol{x} - \boldsymbol{z}\| \right\} \tag{6.14}$$

the maximum cardinality of points $n = \sum_{i=1}^{k} |A_i|$ that a classifier $h \in \mathcal{H}$ can correctly separate, where $\mathcal{H}$ is the space of all the possible suitable classifiers (hypothesis class), is also known as the VC-dimension of the classifier. The nearest neighbor classifier has an infinite VC-dimension since it can "separate" arbitrarily many points.

6.2.1 Brute-force Approach

The most straightforward approach to find the k Nearest Neighbors is brute-force. We calculate all the distances in $O(nd)$ time and then find the k nearest with a partition

[2] To be totally correct, argmin is a set of at least one element and therefore q must lie in it. However, here we assume that we consistently select one element from the set and therefore we can treat it as a function

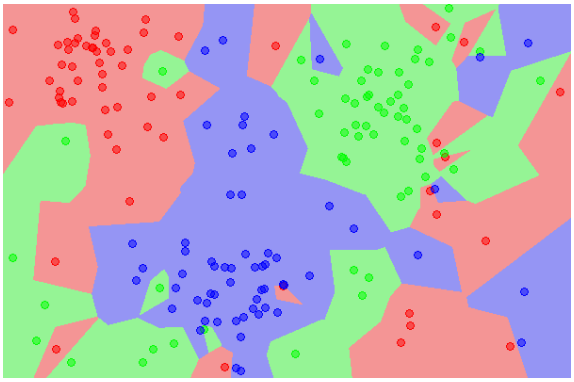


Figure 6.1: Nearest neighbor classifier example with 3 classes.

in linear time, yielding a total runtime of $O(nd)$. Obviously, if we have m queries then the total complexity is $O(nmd)$; which is impractical for most real-world scenaria. Below we will discuss how we can find nearest neighbors — either exactly or approximately — via *data structures* like Ball Trees, KD Trees, Locality Sensitive Hashing or Dynamic Continuous Indexing.

6.2.2 Intrinsic Dimensionality

Contrary to the well known notion of dimensionality — also called the *ambient dimensionality* — nearest neighbor search problems are mainly dependent on the intrinsic structure of the data. For that reason, one refers to this quantity as the *intrinsic dimensionality of the data*. Below we give the definition

Definition 8. *A set of points $D \subseteq \mathbb{R}^d$ has intrinsic dimensionality Δ if for all $r > 0$ and $\alpha > 1$ and $p \in D$ any ball $B(p, r)$ satisfies*

$$|B(p, \alpha r)| \leq \alpha^{\Delta} |B(p, r)| \tag{6.15}$$

In other words, every ball of radius $r > 0$ contains $O(r^{\Delta})$ points from D.

To give an intuitive explanation, if the data points are uniformly distributed on a manifold, then $\Delta \approx d$. To adduce some examples, the d dimensional integer lattice $\mathbb{Z}^d$ has $\Delta = d$, and if one embedded a set D to a set D' in a higher dimensional space, then intrinsic dimensionality would be retained. Moreover, as a simple thought experiment take the set $D = \mathbb{Z}^d$ and a query point $\boldsymbol{x}$ that has $1/2$ on all of its coordinates. Then $\boldsymbol{x}$ has 2^d candidate nearest neighbors (wrt to the Eucledian norm) since distance $d_{\boldsymbol{x}} : \mathbb{Z}^d \to \mathbb{R}$ with

$$d_{\boldsymbol{x}}(\boldsymbol{y}) = \sqrt{\sum_{i=1}^{d} \left(\frac{1}{2} - y_i \right)^2} \tag{6.16}$$

attains a minimum at exactly 2^d points.

6.2.3 KD Trees

The KD Tree is a divide-and-conquer data structure that allows efficient queries when the intrinsic dimensionality (density) of the space is low. More specifically the construction of the KD Tree builds a tree given a set of S of points and has two main parts

1. Base: When $|S| = 1$ then return the point (as a leaf of the KD Tree)

2. Recursion: Pick a dimension $1 \leq i \leq d$, find the median point (wrt to this dimension) and partition the space into two sets L, R. Then recurse on L and R making two nodes v_L, v_R in the tree.

It is straightforward that, for example, in 1 dimension, the average-case query time is $O(\log n)$ whereas the construction time is $O(n \log n)$. In the more general case, querying an axis-parallel range in a balanced KD tree takes $O(n^{1-1/d} + m)$ time, where m is the number of the reported points.

6.2.4 Ball Tree

The Ball Tree data structure follows a similar philosophy to KD Trees. More specifically, at each step, the data are partitioned into two balls B_1, B_2 with centers $\boldsymbol{c}_1, \boldsymbol{c}_2 \in \mathbb{R}^d$. Each point is assigned to the ball with the nearest center to it. Finally, each leaf node in the tree defines a ball and enumerates all data points inside that ball. If d^* is the dimension of greatest spread and $\boldsymbol{m}$ is the median point along d^* then the sets L, R that are created contain points lying to the left and right of $\boldsymbol{m}$ respectively. Querying the nearest neighbors of a point $\boldsymbol{z}$ is done through a depth-first traversal of the Ball Tree. A max-first priority queue Q is maintained and at each node B we do one out of three operations

1. If for all $\min_{\boldsymbol{x} \in B} \|\boldsymbol{x} - \boldsymbol{z}\|_2 > \max_{\boldsymbol{y} \in Q} \|\|\boldsymbol{y} - \boldsymbol{x}\|_2$ then return Q.

2. If B is a leaf then do an exhaustive scan of the points that lie inside B and update Q.

3. If B is an internal node with children B_L, B_R then recurse on B_L, B_R starting from the set whose center is closer to the query point. The order of search usually prunes the search space considerably.

The performance of the Ball Tree is similar to the KD Tree, except from a slighter advantage in higher dimensionalities. The general rule of thumb for choosing the appropriate between Brute-force, KD Trees and Ball Trees is that when $d < 20$ the KD Tree is efficient, for small n Brute-force can be used and for the rest of the cases the Ball Tree is preferred.

6.2.5 Locality Sensitive Hashing

Another method of obtaining nearest neighbors is through hashing. To get the reader familiar with the ideas behind Locality Sensitive Hashing (LSH) [41] we first need to define Locality Sensitive functions. A hash function $h : A \to B$ is (d_1, d_2, p_1, p_2)-locality-sensitive if and only if

1. $\|\boldsymbol{x} - \boldsymbol{y}\| \leq d_1 \implies \Pr[h(\boldsymbol{x}) = h(\boldsymbol{y})] \geq p_1$
2. $\|\boldsymbol{x} - \boldsymbol{y}\| \geq d_2 \implies \Pr[h(\boldsymbol{x}) = h(\boldsymbol{y})] \leq p_2$

For instance, if we use MinHash as h and the Jaccard distance as $\|\cdot\|$ then for $0 \leq d_1 < d_2 \leq 1$, the family is $(d_1, d_2, 1 - d_1, 1 - d_2)$-locality-sensitive. The same doctrine can be followed for other distance measures such as the Hamming distance, the Eucledian distance and the Cosine distance. Now we make the above more general and define a family $\mathcal{H} = \{h_i | h_i : A \to B, 1 \leq i \leq n\}$ of (independent) hash functions that is (d_1, d_2, p_1, p_2)-locality-sensitive. Defining the AND operator as all the results from the h_i's falling in the same bucket. Since the h_is are independent then AND is (d_1, d_2, p_1^n, p_2^n)-locality-sensitive. In the same way the OR operator is $(d_1, d_2, 1 - (1 - p_1)^n, 1 - (1 - p_2)^n)$-locality-sensitive. We can chain AND and OR "gates" as boolean circuits to create the desired probabilities, such that the "good" probability is high and the "bad" probability is low. The nearest neighbor search using LSH one can chain k hash functions as AND and l hash functions as OR to find a neigboring point with probability at least $1 - (1 - p_1^k)^l$. Similarly the failure probability is at most $1 - (1 - p_2^k)^l$. Letting $k = \log n/(\log(1/p_2)$ and $l = n^\rho$ where $\rho = \log p_1 / \log p_2$ one can obtain a point within distance $(1 + \epsilon)R$ from the query point with preprocessing $O(n^{1+\rho})$, space complexity of $O(n^{1+\rho})$ and query time of $O(n^\rho(kT + l))$ if T time is needed to compute $h_i(\boldsymbol{x})$.

6.3 Dynamic Continuous Indexing

Dynamic Continuous Indexing (DCI) is a recent idea on nearest neighbor search that appeared in [63] and [64]. In DCI/PDCI we choose m random directions $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ and project the n points $D = \{\boldsymbol{x}_1 \ldots, \boldsymbol{x}_n\}$ onto them. Given a query point $\boldsymbol{y}$ we define the candidate points along each random direction to be the nearest points with respect to the projections

$$\boldsymbol{z}_j = \text{argmin}_{\boldsymbol{x} \in D} \, |\langle \boldsymbol{v}_i, \boldsymbol{x} \rangle - \langle \boldsymbol{v}_i, \boldsymbol{y} \rangle| \qquad j \in [m] \tag{6.17}$$

The DCI algorithm proceeds as follows using a central min-heap Q and datastructures that efficiently answer nearest neighbor queries in one dimension. [3]

1. Start by initializing an empty heap Q and then add the first round of shortest projected distances by querying the data structures in $O(\log n)$ time.

[3] Without loss of generality assume that all distances are unique

2. Between all projections find the $\boldsymbol{z}_{j*}$ which has the minimum distance compared to the other $\boldsymbol{z}_j$'s, and corresponds to direction j^*.

3. At the direction j^* find the next nearest point (different from $\boldsymbol{z}_{j*}$) and add it to the min-heap Q.

4. Repeat until there is a point which has visited candidate projections among all directions $j \in [m]$.

The correctness of the method is thightly related to the idea of the JL transform, where the projection along a random dimension serves as an unbiased estimator of the distance. The construction time is $O(m(dn + n \log n))$, takes $O(nm)$ space and queries can be answered at $O(dk \max\{\log(n/k), (n/k)^{1-m/\Delta}\} + mk \log m \max\{\log(n/k), (n/k)^{1-1/\Delta})$. Intuitively an increase in the intrinsic dimensionality Δ can be compestated by an increase in the number m of projection directions.

Part C

Contribution

Chapter 7

Stochastic Opinion Dynamics for Interest Prediction in Social Networks

"We should not ignore the fact that in the real world consensus is usually not reached. Recognizing this, most traditional social network scientists do not focus on an equilibrium of consensus. They are instead more likely to be concerned with explaining the lack of consensus (the variance) in beliefs and attitudes that appears in actual social influence contexts."

— David Krackhardt, A plunge into Networks [57]

7.1 The Nearest Neighbor Influence Model

We present the NNIM model and the inference algorithm in Algorithm 1. We assume that the network $G(C \cup U, E, \hat{\boldsymbol{X}})$ consists of a core C, a periphery U with size $|U| = n$, and a matrix of initial features $\hat{\boldsymbol{X}}$ with an d-dimensional binary vector $\hat{\boldsymbol{X}}_c$ for each $c \in C$ which represents the trends that c endorses throughout the iterative process. The core members serve as *steady-state-trend-setters* meaning that their interests do not change throughout the process. NNIM proceeds in steps, where we use the letter t to denote time steps. Each peripheral user $u \in U$ has a d-dimensional vector at time t, denoted by $\boldsymbol{X}_u^{(t)}$. Each $u \in U$ initializes her vector as a Bernoulli trial with a probability equal to the maximum likelihood estimation (sample mean) given the members of the core she follows. At each step $t \geq 1$ each member of the periphery $u \in U$ observes her k-nearest neighbors with respect to the Hamming Norm $\sum_{i=1}^{d} \mathbf{1}\left\{X_{ui}^{(t)} \neq X_{vi}^{(t)}\right\}$, which quantifies how much the agent disagrees with another agent $v \in U$, and constructs the stochastic set $\mathcal{K}^{(t)}(u)$. Afterwards the agent constructs the vector $\boldsymbol{\xi}_u^{(t+1)}$ which is the average of the observed opinions inside the set $\mathcal{K}^{(t)}(u)$ including the user herself, as $\boldsymbol{\xi}_u^{(t+1)} = \frac{1}{k} \sum_{v \in \mathcal{K}^{(t)}(u)} \boldsymbol{X}_v^{(t)}$. Then each agent updates her opinion $\boldsymbol{X}_u^{(t+1)}$ at time $t+1$ drawing a Bernoulli sample from $\boldsymbol{Be}(\boldsymbol{\xi}_u^{(t+1)})$,

Algorithm 2 Generative Model (`NNIM` procedure) and Inference (`NNIM_Inference` procedure). The functions `FindStochNN` and `FindNN` query the k nearest neighbors of a node $u \in U$ at time t based on their stochastic vectors (with respect to the Hamming Distance) or their expected values (with respect to the L2 Norm) respectively. In EM jargon, finding the k nearest neighbors is analogous to an E-Step and updating the variational and macroscopic parameters is analogous to an M-Step.

procedure Initialize($\hat{\boldsymbol{X}}, C, U$)
  for $u \in U$ **do**
    $\boldsymbol{\xi}_u^{(0)} = \frac{1}{|N^+(u) \cap C|} \sum_{v \in N(u) \cap C} \hat{\boldsymbol{X}}_v$
    $\boldsymbol{X}_u^{(0)} \sim \boldsymbol{Be}(\boldsymbol{\xi}_u^{(0)})$
  end for
end procedure

procedure NNIM($\hat{\boldsymbol{X}}, C, U, k$)
  Initialize($\hat{\boldsymbol{X}}, C, U$)
  $t \leftarrow 0$
  while no consensus **do**
    for $u \in U$ **do**
      $\mathcal{K}^{(t)}(u) \leftarrow$ FindStochNN(u, k, t)
      $\boldsymbol{\xi}_u^{(t+1)} = \frac{1}{k} \sum_{v \in \mathcal{K}^{(t)}(u)} \boldsymbol{X}_u^{(t)}$
      $\boldsymbol{X}_u^{(t+1)} \sim \boldsymbol{Be}(\boldsymbol{\xi}_u^{(t+1)})$, $t \leftarrow t + 1$
    end for
  end while

end procedure
procedure Initialize_Infer($\hat{\boldsymbol{X}}, C, U$)
  for $u \in U$ **do**
    $\boldsymbol{\phi}_u^{(0)} = \frac{1}{|N^+(u) \cap C|} \sum_{v \in N(u) \cap C} \hat{\boldsymbol{X}}_v$
  end for
end procedure

procedure NNIM_Inference($\hat{\boldsymbol{X}}, C, U, k$)
  Initialize_Infer($\hat{\boldsymbol{X}}, C, U$)
  $t \leftarrow 0$
  while no consensus **do**
    for $u \in U$ **do**
      $K^{(t)}(u) \leftarrow$ FindNN(u, k, t)
      $\phi_u^{(t+1)} \leftarrow \frac{1}{k} \sum_{v \in K^{(t)}(u)} \boldsymbol{\phi}_v^{(t)}$
    end for
    $t \leftarrow t + 1$
    $\boldsymbol{\mu}^{(t+1)} \leftarrow \frac{1}{|U|} \sum_{u \in U} \boldsymbol{\phi}_u^{(t+1)}$
  end while
end procedure

independently for each coordinate. The process continues until consensus is reached in expectation. This way, at each step t, a *stochastic temporal graph* G_t is created, where each agent has a neighborhood that corresponds to her k-nearest neighbors, in place of the actual OSN (see the `NNIM` procedure in Algorithm 1 for details).

Intuitively, NNIM aims to explain the space of user interests in the network by homophily. So, NNIM treats the k nearest neighbors of a user wrt. her interests as her highly homophilic nodes. To test our hypothesis that NNIM explains well the interests of the peripheral users, we compare the neighborhood of the ground social network with the k-nearest neighbors for each $u \in C \cup U$ according to NNIM. Given the un-initialized directed social network $G(C \cup U, E, \hat{X})$ (where each user has a binary interest vector), we define $\boldsymbol{\alpha}_w = \frac{1}{|N^+(w)|+1} \sum_{v \in N^+(w) \cup \{w\}} \hat{\boldsymbol{X}}_v$ and $\boldsymbol{\beta}_w = \frac{1}{k_w} \sum_{v \in \mathcal{K}(w)} \hat{\boldsymbol{X}}_v$, where $N^+(w)$ is the set of users that w follows and k_w is either $|N^+(w)| + 1$ or $\lceil \log n \rceil$ (depending on the column of Table 7.1). These vectors represent the average feature vector over a user's

Table 7.1: Dataset Statistics and Homophilic Index are reported. We count directed edges where the network is undirected. The Homophilic Index is calculated after dimensionality reduction with PCA so that 95% of the original variance is explained after the transformation.

Name	Network Type	Nodes	Edges	Homophilic Index $k_u = \lvert N^+(u)\rvert + 1$	Homophilic Index $k_u = \lceil \log n \rceil$	d
facebook [60, 62]	ego	1.03K	27.8K	93.24	91.03	576
dblp-dyn [27]	co-authors	1.23K	4.6K	82.02	83.56	43
fb-pages [60, 94]	page-page	22.5K	342K	91.69	92.31	4
github [60, 94]	developer	37.7K	578K	85.48	84.41	1
dblp [90]	co-authors	41.3K	420K	82.54	85.62	29
pokec [60, 103]	social	1.6M	30.6M	66.10	67.72	280

ground-truth neighborhood and her k_w-nearest neighbors in the ground network. For each user, we measure the Root Mean Squared Error $\text{RMSE}(\boldsymbol{\alpha}_w, \boldsymbol{\beta}_w) = d^{-1/2}\|\boldsymbol{\alpha}_w - \boldsymbol{\beta}_w\|$ for each node $w \in C \cup U$. Then, we take a degree weighted average, where the weight of each node is $(1 + |N^+(w)|)/(|E| + |C \cup U|)$, and measure the distance from 100%. This degree-weighted average puts emphasis on the nodes by order of "prestige" in the network G. We call this quantity the *Homophilic Index* (HI) of G. Intuitively, the HI measures how much the aforementioned two neighborhoods look similar in the feature space. We report the HI for the studied datasets in Table 7.1.

7.1.1 Model Inference through Variational Expectation-Maximization

For the inference problem we are interested in determining the parameters the peripheral nodes in the NNIM model, namely the probability vectors $\{\boldsymbol{\xi}_u^{(t)}\}_{u\in U}^{t\geq 0}$ of the feature vectors $\{\boldsymbol{X}_u^{(t)}\}_{u\in U}^{t\geq 0}$ given the initial state of the cores' interests. We start by forming the optimization objective (log-likelihood) at each step t. Initially, according to our setting we assume that we know the initial values of the peripheral user interests as the samples with probabilities equal to the sample average of the influencers of the core she is following, as delineated in the procedure `INITIALIZE_INFER` of Algorithm 1. In reference [12], Bindel et al. view the opinion formation problem for the FJ model under a *game-theoretical viewpoint* where each agent suffers a quadratic convex cost for not reaching consensus at a given time t. Similarly, in our case at each time t is the (instantaneous) log-likelihood that better explains the distribution of the agents parametrized by $\boldsymbol{\xi}^{(t+1)}$ is needed to be maximized, given the previous state of the agents $\boldsymbol{X}^{(t)}$, that is

$$\mathcal{L}_{\xi}^{(t+1)}\left(\boldsymbol{\xi}^{(t+1)}\right) = \log \sum_{\boldsymbol{X}^{(t)}} \Pr\left[\boldsymbol{X}^{(t)} | \boldsymbol{\xi}^{(t+1)}\right] \tag{7.1}$$

We observe the initial opinions $\boldsymbol{X}^{(0)}$ of the network and then the opinion vectors are latent, thus inference requires summation over exponentially many events. The opinion vectors are assumed to have the *Markov property*, namely the opinions at a given time are affected only by the previous step. Observe that the stochastic nature of the model imposes intractability on the likelihood functions $\mathcal{L}_{\xi}^{(t+1)}$ since it requires a summation over the exponentially-many latent variables $\boldsymbol{X}^{(t)}$ which have binary outcomes. For simplicity, we assume that the interest distribution is approximated by a *variational distribution* $Q^{(t)}$ that makes the latent variables $\{\boldsymbol{X}^{(t)}\}_{t\geq 1}$ independent, and approaches the actual parameters $\{\boldsymbol{\xi}^{(t)}\}_{t\geq 1}$ having a form of $Q^{(t)} = \prod_{u\in U}\prod_{i=1}^{d}\left(\phi_{iu}^{(t)}\right)^{X_{iu}^{(t)}}\left(1-\phi_{iu}^{(t)}\right)^{1-X_{iu}^{(t)}}$ where $\boldsymbol{\phi}_u^{(t)}$ are the variational parameters that are the "empirical counterparts" of the actual parameters $\boldsymbol{\xi}_u^{(t)}$[1]. Using Jensen's Inequality on the likelihood function $\mathcal{L}_{\xi}^{(t+1)}$, that is $\mathcal{L}_{\xi}^{(t+1)} \geq \mathbb{E}_{Q^{(t)}}\left[\log \Pr[\boldsymbol{X}^{(t)}|\boldsymbol{\xi}^{(t+1)}]\right] + \mathbb{E}_{Q^{(t)}}\left[-\log Q(\boldsymbol{X}^{(t)})\right]$, we obtain two terms, the first of which (Evidence Lower Bound/ELBO) we maximize, since the second term (Entropy) is positive. Maximizing the ELBO $\mathcal{L}_{Q,\xi}^{(t+1)} = \mathbb{E}_{Q^{(t)}}\left[\log \Pr[\boldsymbol{X}^{(t)}|\boldsymbol{\xi}^{(t+1)}]\right]$ is a tractable problem [47, 54] and can be used as a proxy for approximating the actual interest distribution. Now, the ELBO can be expressed as

$$\mathcal{L}_{Q,\xi}^{(t+1)} = \mathbb{E}_{Q^{(t)}}\left[\sum_{i=1}^{d}\sum_{u\in U}\sum_{v\in U}\mathbf{1}\left\{v\in\mathcal{K}^{(t)}(u)\right\}\left(X_{iv}^{(t)}\log\xi_{iu}^{(t+1)} + \left(1-X_{iv}^{(t)}\right)\log\left(1-\xi_{iu}^{(t+1)}\right)\right)\right]$$

Computing the expectation over the stochastic set $\mathcal{K}^{(t)}(u)$ of the k-nearest neighbors exactly still poses computational barriers. However, the aforementioned expectation can be estimated by observing that choosing the k-nearest neighbor random vectors can be approximated a.a.s. by choosing the k-nearest neighbors with respect to their parameter vectors. To found our claim, we first prove the following helper lemma about the behaviour of the distance between two random Bernoulli vectors with respect to an L-Lipschitz Norm. We state the following Lemma using Talagrand's Inequality (see Appendix A)

Lemma 3. *Let $\|\cdot\| : [0,1]^d \to \mathbb{R}_+$ be a L-Lipschitz and convex norm. Let $\boldsymbol{X} = (X_1,\dots X_d) \sim \boldsymbol{Be}(\boldsymbol{p})$ and $\boldsymbol{Y} = (Y_1,\dots,Y_d) \sim \boldsymbol{Be}(\boldsymbol{q})$ be two random vectors such that the components of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent with respect to each other. Then for every $\eta \in (\eta_0, +\infty)$*

$$\Pr\left[\|\|\boldsymbol{p}-\boldsymbol{q}\| - \|\boldsymbol{X}-\boldsymbol{Y}\|\| \geq \eta\right] \leq c_1\exp(-c_2(\eta-\eta_0)^2/L^2) \tag{7.2}$$

for some constants c_1, c_2 and $\eta_0 = \mathbb{E}\left[\|\boldsymbol{Z}\|\right] - \|\mathbb{E}\left[\boldsymbol{Z}\right]\| > 0$ where $\boldsymbol{Z} = \boldsymbol{X}-\boldsymbol{Y}$.

Proof. The proof is derived easily from Talagrand's Inequality and the triangle inequality. Let $\boldsymbol{Z} = \boldsymbol{X}-\boldsymbol{Y}$ be the difference of the random vectors with expected value $\boldsymbol{r} = \mathbb{E}\left[\boldsymbol{Z}\right] = \boldsymbol{p}-\boldsymbol{q}$. Since $\|\cdot\|$ is L-Lipschitz and $\|\boldsymbol{Z}\|_\infty \leq 1$, Talagrand's Inequality we obtain

[1]This approach is also known as *mean field approximation* [13, 50, 101].

$$\Pr[|\|\boldsymbol{Z}\| - \mathbb{E}\left[\|\boldsymbol{Z}\|\right]| \geq \eta] \leq c_1 \exp(-c_2\eta^2/L^2) \tag{7.3}$$

for every $\eta > 0$ and some constants $c_1, c_2 > 0$. By triangle inequality

$$|\|\boldsymbol{r}\| - \|\boldsymbol{Z}\|| \leq \eta_0 + |\|\boldsymbol{Z}\| - \mathbb{E}\left[\|\boldsymbol{Z}\|\right]| \tag{7.4}$$

where $\eta_0 = \mathbb{E}\left[\|\boldsymbol{Z}\|\right] - \|\boldsymbol{r}\| > 0$ (due to convexity). Let $\eta \in (\eta_0, \infty)$ and the events $A = \{\omega \in \Omega \mid \|\boldsymbol{Z}(\omega)\| - \|\boldsymbol{r}\|| \geq \eta\}|$ and $B = \{\omega \in \Omega \mid |\|\boldsymbol{Z}(\omega)\| - \mathbb{E}\left[\|\boldsymbol{Z}\|\right]| \geq \eta - \eta_0\}$. By Eq. 7.4 we obtain that $A \subseteq B \implies \Pr[A] \leq \Pr[B]$. Invoking Eq. 7.3 for $t = \eta - \eta_0 > 0$ we obtain

$$\Pr[A] \leq \Pr[B] \leq c_1 \exp(-c_2(\eta - \eta_0)^2/L^2) \tag{7.5}$$

□

We now specialize the result for the Hamming Norm. The Hamming Norm is $2\sqrt{d}$-Lipschitz (in $[0,1]^d$) with respect to the Eucledian Norm and convex as well. The previous Lemma specializes to the following Corollary

Corollary 2. *Let $\boldsymbol{X}, \boldsymbol{Y}$ be two random Bernoulli vectors that satisfy the hypotheses of Lemma 3 with respect to the Hamming Norm. Then for every $\epsilon > 0$ the following is true*

$$\Pr[|\|\boldsymbol{X} - \boldsymbol{Y}\| - \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{Y}\|\right]| \geq (1+\epsilon)d/2] \leq 2\exp\left(-\frac{\epsilon^2 d}{2}\right) \tag{7.6}$$

Proof. Our proof is based on Lemma 3 and McDiarmid's Inequality. From Lemma 3 the quantity η_0 wrt. to the Hamming Norm can be given as

$$\eta_0 = \sum_{i=1}^{d}(p_i(1-q_i) + q_i(1-p_i)) - \sum_{i=1}^{d}(p_i - q_i)^2 \tag{7.7}$$

The extremum is found at $p_i = q_i = 1/2$ and therefore $\eta_0 \leq d/2 = O(d)$. Hence the bound is simplified as

$$\Pr[|\|\boldsymbol{X} - \boldsymbol{Y}\| - \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{Y}\|\right]| \geq \eta] \leq c_1 \exp(-c_2(\eta - d/2)^2/(4d)) \tag{7.8}$$

Now, via McDiarmid's Inequality [30] one can determine the constants. More specifically, for the Hamming Norm, the above equality specializes to

$$\Pr[|\|\boldsymbol{X} - \boldsymbol{Y}\| - \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{Y}\|\right]| \geq \eta] \leq 2\exp\left(-\frac{2\left(\eta - \frac{d}{2}\right)^2}{d}\right) \tag{7.9}$$

It is easy to observe that, in order to keep the error probability less than δ one should choose

$$d \leq \frac{2\eta}{1 + \log(2/\delta)} \tag{7.10}$$

Moreover, setting $\eta = (1+\epsilon)d/2$ since $\eta \in (\eta_0, \infty)$ one can get

$$\Pr[|\|\boldsymbol{X}-\boldsymbol{Y}\| - \mathbb{E}\left[\|\boldsymbol{X}-\boldsymbol{Y}\|\right]| \geq (1+\epsilon)d/2] \leq 2\exp\left(-\frac{\epsilon^2 d}{2}\right) \tag{7.11}$$

Setting $d \to \infty$ we observe that $\exp(-\epsilon^2 d/2) \to 0$. In the same way, to keep the error less than or equal to δ, one must set

$$d = \Omega\left(\frac{\log(2/\delta)}{\epsilon^2}\right) \tag{7.12}$$

□

Using Corollary 2, we are able to state that the k-nearest neighbor set of a user wrt. to the stochastic vectors is near to the k-nearest neighbor set wrt. to their expected values, for appropriate choices of k.

Theorem 5. *Let U be a collection of n Bernoulli d-dimensional vectors that are pairwise independent, and $k \leq C(4n\exp(-d\epsilon^2) + \log n)$ neighbors for $C > 1$ and $\epsilon > 0$. If $\mathcal{K}(u)$ is the set of stochastic k-nearest neighbors of $u \in U$ with respect to the Hamming Norm and $K(u)$ is the set of k-nearest neighbors of $u \in U$ with respect to their parameters measured in the (squared) Euclidean Norm, then the probability that the two sets contain the same elements is $1 - O(1/n)$.*

Proof. We are given n independent Bernoulli Variables in d-dimensions, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, distributed with parameter vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$. We fix $\boldsymbol{X}_1$ and construct the random set $\mathcal{K}$ of the k-Nearest Neighbors of $\boldsymbol{X}_1$ with respect to a total ordering relation $\prec$. We also construct the set K which contains the k-Nearest Neighbors of $\boldsymbol{\phi}_1$ as expectations. Our aim is to provide an exponential bound on the error probability of the symmetric difference $\mathcal{K} \ominus K = (\mathcal{K} \setminus K) \cup (K \setminus \mathcal{K})$, namely on $P_e = \Pr_k[|\mathcal{K} \ominus K| \geq 1]$, in order to state that $\mathcal{K} \approx K$ with probability tending to 1.

We first compute a Chernoff bound for the probability $\tilde{p}$ that the k-th neighbor v is mistaken for an adversary $v' \notin \mathcal{K}$ given that the $k-1$ neighbors are correctly included in both $\mathcal{K}$ and K. In order for this to happen the following two inequalities must hold

$$\|\boldsymbol{X}_1 - \boldsymbol{X}_v\| \leq \|\boldsymbol{X}_1 - \boldsymbol{X}'_v\| \tag{7.13}$$

$$\|\boldsymbol{\phi}_1 - \boldsymbol{\phi}'_v\| \leq \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_v\| \tag{7.14}$$

Adding both inequalities we get that

$$\begin{aligned}
\tilde{p} \leq & \Pr[|\|\boldsymbol{X}_1 - \boldsymbol{X}_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_v\| \leq \|\boldsymbol{X}_1 - \boldsymbol{X}'_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}'_v\|] \\
\leq & \Pr[|\|\boldsymbol{X}_1 - \boldsymbol{X}_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_v\|| \leq \eta, \eta \leq |\|\boldsymbol{X}_1 - \boldsymbol{X}'_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}'_v\|| \leq \eta'] \\
= & \Pr[|\|\boldsymbol{X}_1 - \boldsymbol{X}_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_v\|| \leq \eta] \Pr[\eta \leq |\|\boldsymbol{X}_1 - \boldsymbol{X}'_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}'_v\|| \leq \eta'] \\
\leq & \Pr[|\|\boldsymbol{X}_1 - \boldsymbol{X}_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_v\|| \leq \eta] \Pr[|\|\boldsymbol{X}_1 - \boldsymbol{X}'_v\| - \|\boldsymbol{\phi}_1 - \boldsymbol{\phi}'_v\|| \leq \eta'] \\
\leq & 4\exp(-\epsilon^2 d) = \tilde{p}_b
\end{aligned} \tag{7.15}$$

Now, the probability of having $1 \leq \ell \leq k$ neighbors mistaken is

$$\Pr[|\mathcal{K} \ominus K| = \ell] \leq \binom{n-k}{\ell} \tilde{p}_b^{\ell} \leq \frac{(n-k)^{\ell}}{\ell!} \tilde{p}_b^{\ell} = ((n-k)\tilde{p}_b)^{\ell} \left(\frac{e}{\ell}\right)^{\ell} = \left(\frac{(n-k)e\tilde{p}_b}{\ell}\right)^{\ell} \tag{7.16}$$

Letting

$$\ell \leq k \leq C\left(4n \exp\left(-\epsilon d^2\right) + \log n\right) \tag{7.17}$$

We have that

$$\Pr[|\mathcal{K} \ominus K| = \ell] \leq \frac{1}{n^C} \to 0 \text{ as } n \to \infty \tag{7.18}$$

By a union bound the total error probability

$$P_e = \Pr[|\mathcal{K} \ominus K| \geq 1] \leq \sum_{\ell=1}^{k} \Pr[|\mathcal{K} \ominus K| = \ell] \leq \frac{k}{n^C} \leq C\left(4n^{1-C} \exp(-\epsilon^2 d) + \frac{\log n}{n^C}\right) \tag{7.19}$$

Clearly for $C > 1$ we have that $\lim_{n\to\infty} P_e = 0$.

□

Therefore, via selecting an appropriate value of k we have that the set $\mathcal{K}^{(t)}(u)$ approaches its "expected" set $K^{(t)}(u)$ in the parameter space, where the distances are considered between the parameter vectors of the users, and the variational parameters approach the actual parameters due to Hoeffding's Inequality [98] since $\Pr_{Q^{(t)}}\left[|\xi_{iu}^{(t)} - \phi_{iu}^{(t)}| \geq \epsilon\right] \leq 2\exp(-2k\epsilon^2)$. Under the above result, the ELBO is almost surely approximated as

$$\mathcal{L}_{Q,\xi}^{(t+1)} \approx \sum_{i=1}^{d} \sum_{u \in U} \sum_{v \in K^{(t)}(u)} \left[\phi_{iv}^{(t)} \log \phi_{iu}^{(t+1)} + \left(1 - \phi_{iv}^{(t)}\right) \log\left(1 - \phi_{iu}^{(t+1)}\right)\right] \tag{7.20}$$

The optimal solution to the concave optimization problem of Eq. (7.20) is [98, p. 295]

$$\boldsymbol{\phi}_u^{(t+1)} = \frac{1}{k} \sum_{v \in K^{(t)}(u)} \boldsymbol{\phi}_v^{(t)} \tag{7.21}$$

This system of equations rise by observing the instantaneous likelihood at each time t.[2]. Moreover, in order to make our model more "stubborn" to the initial opinions of the agents we can impose regularization functions $\omega(t)$ such that the negative cross-entropy between the current opinions and the initial opinions is maximized, that is

[2]Similar results, modulo an additive logit term can be deduced if one attempts to maximize the complete likelihood across all steps. Our approach can be viewed as *Pseudo EM* on the complete likelihood [42, 97, 70] which sequentially maximizes the likelihood function of samples with observed and missing data via iteratively replacing the missing data with their expected values and maximizing the known likelihood given the completed data.

$$\omega(t) = \alpha \sum_{u \in U} \sum_{i=1}^{d} \left[\phi_{iu}^{(0)} \log \phi_{iu}^{(t)} + \left(1 - \phi_{iu}^{(0)}\right) \log \left(1 - \phi_{iu}^{(t)}\right) \right] \tag{7.22}$$

where α is the regularization parameter. Intuitively, we introduce one more sample to our model that is modeled by the initial conditions. Differentiating the likelihood we arrive at the recurrence relation

$$\boldsymbol{\phi}_u^{(t+1)} = \frac{1}{k+\alpha} \sum_{v \in K^{(t)}(u)} \boldsymbol{\phi}_v^{(t)} + \frac{\alpha}{k+\alpha} \boldsymbol{\phi}_u^{(0)} \tag{7.23}$$

This equation is similar to the opinion dynamics model where each agent is "stubborn" — namely stuck to her initial opinion — with a weight α as an input, such as in the Friedkin-Johnsen (FJ) Model [38]. While, the system of Eq. 7.21 is shown to converge in this paper, convergence is not guaranteed for Eq. 7.23.

Additionally, it can be proven that the ELBO improves at each iteration given the optimal update rule of (7.21). We give the following result for the improvement of the Evidence Lower Bound (ELBO) at each timestep given the optimal update rule.

Theorem 6. *Let $\mathcal{L}_{Q,\xi}^{*(t+1)}$ be the optimal Evidence Lower Bound at time $t+1$ defined as*

$$\mathcal{L}_{Q,\xi}^{*(t+1)} = \sum_{i=1}^{d} \sum_{u \in U} \sum_{v \in K^{(t)}(u)} \left[\phi_{iv}^{(t)} \log \phi_{iu}^{(t+1)} + \left(1 - \phi_{iv}^{(t)}\right) \log \left(1 - \phi_{iu}^{(t+1)}\right) \right] \tag{7.24}$$

where $\phi_{iu}^{(t+1)} = \frac{1}{k} \sum_{v \in K^{(t)}(u)} \phi_{iv}^{(t)}$ for all $i \in [d]$. Then, at each iteration the bound improves that is $\mathcal{L}_{Q,\xi}^{*(t+1)} \geq \mathcal{L}_{Q,\xi}^{*(t)}$ for all $t \geq 0$.

Proof. Define the helper functions $f(x,y) = x \log y + (1-x)\log(1-y)$, and $g(x,y) = -f(x,y) - f(y,x)$. The (concave) function f is known as the Negative Entropy of Bernoulli variables [98] and the (convex) function g is known as the Variation of Information Metric [5]. We have that

$$
\begin{aligned}
\mathcal{L}_{Q,\xi}^{*(t+1)} &= \sum_{u\in U}\sum_{i=1}^{d}\sum_{v\in K^{(t)}(u)} f\left(\phi_{iv}^{(t)}, \frac{1}{k}\sum_{w\in K^{(t)}(u)}\phi_{iw}^{(t)}\right) \\
&\overset{\text{Jensen}}{\geq} \sum_{u\in U}\sum_{i=1}^{d}\frac{1}{k}\sum_{v,w\in K^{(t)}(u)} f\left(\phi_{iv}^{(t)}, \phi_{iw}^{(t)}\right) \\
&= -\sum_{u\in U}\sum_{i=1}^{d}\frac{1}{2k}\sum_{v,w\in K^{(t)}(u)} g\left(\phi_{iv}^{(t)}, \phi_{iw}^{(t)}\right) \\
&\geq -\frac{1}{2}\sum_{u\in U}\sum_{i=1}^{d}\max_{w\in K^{(t)}(u)}\sum_{v\in K^{(t)}(u)} g\left(\phi_{iv}^{(t)}, \phi_{iw}^{(t)}\right) \\
&\geq -\frac{1}{2}\sum_{u\in U}\sum_{i=1}^{d}\sum_{v\in K^{(t)}(u)}\max_{w\in K^{(t)}(u)} g\left(\phi_{iv}^{(t)}, \phi_{iw}^{(t)}\right) \\
&\geq -\frac{1}{2}\sum_{u\in U}\sum_{i=1}^{d}\sum_{v\in K^{(t-1)}(u)} g\left(\phi_{iv}^{(t-1)}, \phi_{iu}^{(t)}\right) \\
&= \mathcal{L}_{Q,\xi}^{*(t)}
\end{aligned}
\tag{7.25}
$$

Where the last inequality holds from the fact that the total distance between agents decreases as time passes and that g is a metric (similarly to the proof of Theorem 12).

□

Ending, we define the "macroscopic" distribution which is parametrized by $\{\boldsymbol{\mu}^{(t)}\}_{t\geq 1}$ and has a Bernoulli density over the interests, with parameter vectors defined as $\boldsymbol{\mu}^{(t)} = \frac{1}{n}\sum_{u\in U}\boldsymbol{\xi}_u^{(t)}$ and displays how the agents behave with respect to trends in general, namely if they adopt (or not) an interest as a whole. Given the calculated parameters $\boldsymbol{\phi}^{(t+1)}$, we can determine the parameters $\boldsymbol{\mu}^{(t+1)}$ using the same variational approach. More specifically, the expected log-likelihood $\mathcal{L}_{Q,\mu}^{(t)}$ of the *macroscopic parameters* $\boldsymbol{\mu}^{(t)}$ under the variational distribution Q is given as $\mathcal{L}_{Q,\mu}^{(t)} = \sum_{u\in U}\sum_{i=1}^{d}\left(\phi_{iu}^{(t)}\log\mu_i^{(t)} + \left(1-\phi_{iu}^{(t)}\right)\log\left(1-\mu_i^{(t)}\right)\right)$. Invoking the expected value according to the variational parameters and setting $\partial\mathcal{L}_{Q,\mu}^{(t)}/\partial\mu_i^{(t)} = 0$ for all $1\leq i\leq d$ and $1\leq t\leq T$. Analogously to (7.21), we obtain the update rule

$$
\boldsymbol{\mu}^{(t)} = \frac{1}{n}\sum_{u\in U}\boldsymbol{\phi}_u^{(t)} \tag{7.26}
$$

Relation to EM. We refer to the above equations as the *mean field equations* since the variational parameters are "approximated" with exactly the same model, but now the process does not involve randomness. From an EM perspective, we can view our algorithm as having two discrete steps: In the E-step we compute the k nearest neighbors of each agent whereas in the M-step we update the variational parameters by averaging and then compute the "macroscopic distribution" by averaging on the new variational

parameters per dimension. The form of (7.21) is very familiar to the classical opinion dynamics equations, like the HK model.

7.1.2 Model Convergence and Convergence Rate

We prove that NNIM converges[3] in finite time and that the convergence rate in total variation distance[4] is strictly dominated by an exponential with base of $1/\sqrt{k}$. We start by proving the following Lemma about the convergence rate

Lemma 4. *The total variation distance $d_{TV}(t)$ of the 1D NNIM model decreases as $o(k^{-t/2})$ a.a.s. for $n \to \infty$ and any $k \in \mathbb{N}$. More specifically, if we fix some small $\delta \in [0,1]$, and $n = \Omega\left(\delta^{-1/\tau}\right)$ agents where $\tau = \lceil(\sqrt{k-1}+1)/2\rceil - 1$ then with probability of at least $1-\delta$ the total variation distance $d_{TV}(t)$ decreases as $o(k^{-t/2})$.*

Proof. Let $\lambda_2(A(t'))$ represent the second largest eigenvalues of the stochastic matrices $A(t')$ and let $\lambda_2^* = \max_{0\le t'\le t-1}\lambda_2(A(t'))$. Then by the Perron-Frobenius Theorem the convergence rate will be dominated by the second largest eigenvalue of the "slowest" matrix, i.e.

$$\|\Phi^* - \Phi(t)\|_{TV} = O\left((\lambda_2^*)^t\right) \tag{7.27}$$

We define the matrix sequence $\{B(t')\}_{0\le t'\le t-1}$ such that $B(t') = kA(t')$. Let $\chi_{B(t')}(\lambda)$ represent the characteristic polynomial of $B(t')$ and $\chi_{A(t')}(\lambda)$ be the characteristic polynomial of $A(t')$ then it is straightforward to show that $\chi_{A(t')}(\lambda) = 1/k^n\chi_{B(t')}(\lambda/k)$. Therefore the eigenvalues of $A(t')$ are connected with the eigenvalues of $B(t')$ with the relation $\lambda(A(t')) = \frac{1}{k}\lambda(B(t'))$. The matrices $\{B(t')\}_{0\le t'\le t-1}$ represent the adjacency matrices of k-regular graphs. Hence our problem resides in determining an upper bound on the second largest eigenvalue of a k-regular graph $G(t')$. This is a well known problem in Spectral Graph Theory once conjectured by Alon [**?**] and recently proved by Friedman in reference [39]. Alon-Friedman's Theorem states that for any $0 \le t' \le t-1$

$$\Pr[\lambda_2(B(t')) \le 2\sqrt{k-1} + \varepsilon] \ge 1 - O(n^{-\tau}) \tag{7.28}$$

for some fixed $\varepsilon > 0$ and $\tau = \lceil(\sqrt{k-1}+1)/2\rceil - 1$. Since the eigenvalues of $B(t')$ are k times larger than the eigenvalues of $A(t')$

$$\Pr[\lambda_2(A(t')) \in o(k^{-1/2})] \ge 1 - O(n^{-\tau}) \tag{7.29}$$

Hence

[3] All our proofs regarding convergence assume that the model has $d = 1$ dimension (unless otherwise stated), and the coordinate indices are discarded for ease of notation. The results can be extended to d dimensions defining the appropriate structures (convex hull) to showcase cluster isolation phenomena as described below.

[4] The total variation distance between two measures μ, ν defined on a countable set Ω and a σ-algebra $\mathcal{F}$ of the subsets of Ω is given as $\|\mu - \nu\|_{TV} = \sup_{A\in\mathcal{F}}|\mu(A) - \nu(A)| = \frac{1}{2}\|\mu - \nu\|_1$.

$$\Pr[|\lambda_2^*| \in o(k^{-1/2})] \geq 1 - O(n^{-\tau}) \tag{7.30}$$

And for the error probability for $d_{TV}(t)$

$$\Pr[d_{TV}(t) \in \Omega(k^{-t/2})] \leq O(n^{-\tau}) \leq \delta \tag{7.31}$$

choosing $n = \Omega\left(\delta^{-\frac{1}{\tau}}\right)$ we have that with probability of at least $1 - \delta$

$$\Pr[d_{TV}(t) \in o(k^{-t/2})] \geq 1 - \delta \tag{7.32}$$

□

To prove the finite time convergence we rely in Lyapunov Stability Theory (see Chapter 5). Intuitively, our goal is to define a *potential function* which is strictly decreasing away from the equilibrium point (consensus point) of the network and that the agents form clusters which after a certain (finite) step, do not interact and hence convergence occurs in finite time. We first start by stating that the ordering of the agents is preserved (in 1D) throughout the process. More specifically, we define the set $\hat{K}^{(t)}(u)$ of the k nearest neighbors of u. In case of ties, these ties are broken arbitrarily. However, as we prove below, the relative ordering of vertices persists from one round to the next, even if ties are broken arbitrarily

Lemma 5 (Persistence of Relative Ordering)**.** *If for two agents u and v at time t_0 the relation $\phi_u^{(t_0)} \leq \phi_v^{(t_0)}$ holds, then $\phi_u^{(t)} \leq \phi_v^{(t)}$ for all $t \geq t_0$ under arbitrary breaking of ties.*

Proof. Order the elements of $\hat{K}^{(t_0)}(u)$ and $\hat{K}^{(t_0)}(v)$ by their distance from 0. We pick the lefmost element $w \in \hat{K}^{(t_0)}(u)$ which is related to the leftmost element $z \in K^{(t_0)}(v)$ by the definition of $\hat{K}^{(t_0)}(u)$ as $\phi_w^{(t_0)} \leq \phi_z^{(t_0)}$. We remove the two points and repeat. We finally sum the resulting inequalities to get the result for $t = t_0 + 1$. The case for every $t \geq t_0$ follows inductively from the previous result. □

However, an arbitrary tie-breaking mechanism, does not guarantee that our algorithm will converge. Hence, we need to devise a *systematic ordering* under which we resolve ties which we will use to prove that our algorithm converges. Below we give such a total ordering relation.

We define the following total ordering on the vertices. We firstly enumerate the vertices with ids $v_1, \ldots, v_n$ according to their distance from 0 and then define for each v_i the total ordering $\prec_{i,t}$ such that, for $t \geq 1$

$$v_j \prec_{i,t} v_\ell \iff \|\boldsymbol{\phi}_{v_j}^{(t)} - \boldsymbol{\phi}_{v_i}^{(t)}\| < \|\boldsymbol{\phi}_{v_\ell}^{(t)} - \boldsymbol{\phi}_{v_i}^{(t)}\| \text{ or } \left(\|\boldsymbol{\phi}_{v_j}^{(t)} - \boldsymbol{\phi}_{v_i}^{(t)}\| = \|\boldsymbol{\phi}_{v_\ell}^{(t)} - \boldsymbol{\phi}_{v_i}^{(t)}\| \text{ and } j < l\right) \tag{7.33}$$

Indeed, $\prec_{i,t}$ is a total ordering relation and it is straightforward to show that it satisfies the connexity, antisymmetry and transitivity properties. The sets $K^{(t)}(u)$ of the k nearest

neighbors are defined with respect to the $\prec_{i,t}$ total ordering relation and therefore *ties are eliminated*. We also define the set $\sigma^{(t)}(u) = \left\{ v \in U \mid \boldsymbol{\phi}_v^{(t)} = \boldsymbol{\phi}_v^{(t)} \right\}$. The next theorem follows as a special case of Lemma 5

Lemma 6. *If for two agents u and v at time t_0 the relation $\phi_u^{(t_0)} \leq \phi_v^{(t_0)}$ holds, then $\phi_u^{(t)} \leq \phi_v^{(t)}$ for all $t \geq t_0$ under the total ordering relation $\prec_{i,t}$*

Proof. The proof is the same as Theorem 5 however now the ties are not broken arbitrarily. Again the leftmost k-th nearest neighbor of u is at most the leftmost k-th neighbor of v. □

However note that $v_j \prec_{i,t} v_\ell \implies v_j \prec_{i,t+1} v_\ell$. Moreover, we observe that when two agents "fuse" together at time t_0, they remain fused for all $t \geq t_0$. Equivalently $t_1 \leq t_2 \iff \sigma^{(t_1)}(u) \subseteq \sigma^{(t_2)}(u)$ for all $u \in U$.

Lemma 7 (Termination Condition in 1D)**.** *The NNIM algorithm converges at time $T \in \mathbb{N} \cup \{\infty\}$ if and only if $|\sigma^{(T)}(u)| \geq k$ for every $u \in U$.*

Proof. ($\Longleftarrow$) This direction is trivial. Let $|\sigma^{(T)}(u)| \geq k$ for all $u \in U$. Then $\sigma^{(T)}(u) \supseteq K^{(T)}(u)$ for all $u \in U$. The result follows by applying the update rule and the definition of $\sigma^{(T)}(u)$.
($\Longrightarrow$) Suppose that the NNIM algorithm converges. Equivalently for every $t \geq T$ and for every $w \in U$ we have $\phi_w^{(t)} = \phi_w^{(T)}$. We will reside in the case that $t = T + 1$ since the rest follows by induction. Suppose that there exists some $u \in U$ such that $|\sigma^{(T)}(u)| < k$. Then the set $K^{(T)}(u) \setminus \sigma^{(T)}(u)$ is non-empty. So

$$
\begin{aligned}
\phi_u^{(T+1)} &= \frac{1}{k} \left(\sum_{v \in K^{(T)}(u) \cap \sigma^{(T)}(u)} \phi_v^{(T)} + \sum_{w \in K^{(T)}(u) \setminus \sigma^{(T)}(u)} \phi_v^{(T)} \right) \\
&= \frac{k - |K^{(T)}(u) \setminus \sigma^{(T)}(u)|}{k} \phi_u^{(T)} + \frac{1}{k} \sum_{w \in K^{(T)}(u) \setminus \sigma^{(T)}(u)} \phi_v^{(T)} \\
&\overset{\text{Conv}}{\Longrightarrow} \frac{|K^{(T)}(u) \setminus \sigma^{(T)}(u)|}{k} \phi_u^{(T)} = \frac{1}{k} \sum_{w \in K^{(T)}(u) \setminus \sigma^{(T)}(u)} \phi_v^{(T)} \\
&\implies \phi_u^{(T)} = \frac{1}{|K^{(T)}(u) \setminus \sigma^{(T)}(u)|} \sum_{w \in K^{(T)}(u) \setminus \sigma^{(T)}(u)} \phi_v^{(T)}
\end{aligned}
$$

which yields a *contradiction* since there are no constraints on the values of $\phi_v^{(T)}$ which impose such a relation. Therefore, for every $w \in U$ the set $\sigma^{(T)}(w)$ contains at least k elements.

□

We also define the *distance* of two sets $W, Z \subseteq U$ as the quantity

$$
\delta_{WZ}^{(t)} = \min_{w \in W, z \in Z} \|\phi_w^{(t)} - \phi_z^{(t)}\| \tag{7.34}
$$

We directly infer the following properties (which are straighforward to prove)

1. $\delta_{WZ}^{(t)} \geq 0$ for all $W, Z \subseteq U$
2. $\delta_{WW}^{(t)} = 0$ for all $W \subseteq U$
3. $\delta_{WZ}^{(t)} = \delta_{ZW}^{(t)}$ for all $W, Z \subseteq U$
4. $\delta_{WZ}^{(t)} \leq \delta_{WV}^{(t)} + \delta_{VZ}^{(t)}$ for all $V, W, Z \subseteq U$

Therefore $\delta_{WZ}^{(t)}$ is a metric. Moreover we define that two (non-overlapping) intervals *split* if and only if the k-nearest neighbor of each of the closest points are less than $\delta_{WZ}^{(t)}$ for some $t \geq 0$.

Lemma 8. *If two non-overlapping intervals split at t_0 then they remain split for all $t \geq t_0$.*

Proof. Let $W, Z \subseteq U$ be two non-overlapping clusters that have split at t_0. Let $\hat{w}, \hat{z}$ be the closest points of W, Z. Without loss of generality let $\phi_{\hat{w}}^{(t_0)} < \phi_{\hat{z}}^{(t_0)}$. Then for all $u \in K^{(t_0)}(w)$ we have that $\phi_u^{(t_0)} \leq \phi_{\hat{w}}^{(t_0)}$. By summing up we get $\phi_{\hat{w}}^{(t_0+1)} \leq \phi_{\hat{w}}^{(t_0)}$. Similarly $\phi_{\hat{z}}^{(t_0)} \leq \phi_{\hat{z}}^{(t_0+1)}$. Therefore the minimum distance increases. Hence the sets remain split at $t_0 + 1$. Inductively the sets remain split for all $t \geq t_0$ □

We define the *splitting time* of W and Z as the minimum t_0 that the split occurs. We also define that a subset of (consecutive) agents $W \subseteq U$ of cardinality at least k is said to be *isolated* if and only if there exists some $t_0 \geq 0$ such that it splits from the left set $l(W) = \{v \in U \setminus W | \phi_v^{(t_0)} < \inf_{w \in W} \phi_w^{(t_0)}\}$ and the right set $r(W) = \{v \in U \setminus W | \phi_v^{(t_0)} > \sup_{w \in W} \phi_w^{(t_0)}\}$. Now we are ready to state the *finite-time-convergence* result. To start with, we write the system in vector format

$$\Phi(t+1) = A(t)\Phi(t) \tag{7.35}$$

Where $\Phi(t)$ is the column vector with elements $\phi_u(t)$ and $A(t)$ is the stochastic matrix defined as

$$A_{uv}(t) = \begin{cases} 0 & v \notin K^{(t)}(u) \\ \frac{1}{k} & v \in K^{(t)}(u) \end{cases} \tag{7.36}$$

We now make use of the following Theorem from [107]

Lemma 9. *Let $\{A(t)\}_t$ be a sequence of stochastic matrices such that there exists some $\gamma \in (0, 1]$ such that $A_{uu}(t) \geq \gamma$ for all $i \in [n]$ and some scalar $\alpha \in (0, 1]$ such that for every set $S \subset [n]$ and its complement $\bar{S} = S \setminus [n]$ there holds $\sum_{u \in S, v \in \bar{S}} A_{uv}(t) \geq \alpha \sum_{u \in \bar{S}, v \in S} A_{uv}(t)$. Then the dynamics $\Phi(t+1) = A(t)\Phi(t)$ has adjoint dynamics $\Pi(t) = \begin{pmatrix} \pi_1(t) & \dots \pi_n(t) \end{pmatrix}^T$ such that $\Pi^T(t+1) = \Pi^T(t)A(t)$ with $\pi_u(t) > p$ for all u and some $1 > p > 0$.*

We can now use the aforementioned Theorem on the NNIM model to prove the following

Lemma 10. *The NNIM dynamics admit adjoint dynamics of the form provided by Theorem 9.*

Proof. Invoking the aforementioned theorem for $\gamma = 1/k$ and $\alpha = 1/n$ since

1. $A_{uu}(t) = \frac{1}{k}$
2. For every element of $A(t)$ the following holds

$$A_{uv}(t) = \frac{1}{k}\mathbf{1}\{v \in K^{(t)}(u)\} \geq \frac{1}{nk}\mathbf{1}\{u \in K^{(t)}(v)\} = \alpha A_{vu}(t) \tag{7.37}$$

Let $S \subset U$ and $\bar{S} = U \setminus S$. Summing the above equation we arrive at

$$\sum_{u \in S, v \in \bar{S}} A_{uv}(t) \geq \alpha \sum_{u \in \bar{S}, v \in S} A_{vu}(t) \tag{7.38}$$

Therefore, by Lemma 9 the NNIM model admits an adjoint sequence.

□

In order to prove that our system converges in finite time T, we will reside in Lyapunov theory. More specifically, we first prove that the system has a globally asymptotically stable point $\lim_{t\to\infty} \Phi(t) = \Phi^*$. For this we provide the following Lemma

Lemma 11 (Global Asymptotic Stability)**.** *The NNIM model is globally asymptotically stable.*

Proof. By Lemma 9, we assert the existence of the adjoint dynamics. We then define the Lyapunov function

$$V(t) = \sum_{i=1}^{n} \pi_u(t) \|\phi_u(t) - \Pi^T(t)\Phi(t)\|_2^2 \tag{7.39}$$

Our approach will follow the methodology presented in [107] and [81]. Note that $V(t) \geq 0$ for all $t \geq 0$. Letting $H(t) = A^T(t)\text{diag}(\pi_u(t+1))A(t)$ and doing the matrix operations expressing $V(t)$ as a quadratic form the function $V(t)$ can be written as

$$V(t) = V(t+1) + \frac{1}{2}\sum_{u,v} H_{uv}(t)(\phi_u^{(t)} - \phi_v^{(t)})^2 \tag{7.40}$$

since $H^T(t) = H(t)$. The elements of $H(t)$ are

$$H_{uv}(t) = \frac{1}{k^2}\sum_{w} \pi_w(t+1)\mathbf{1}\{u \in K^{(t)}w\}\mathbf{1}\{v \in K^{(t)}(w)\} \tag{7.41}$$

Combining everything we arrive at

$$V(t+1) = V(t) - \frac{1}{2k^2}\sum_{w} \pi_w(t+1) \sum_{u,v \in K^{(t)}(w)} (\phi_u^{(t)} - \phi_v^{(t)})^2 \leq V(t) \tag{7.42}$$

Hence the function $V(t)$ is decreasing globally in $[0,1]^d$. Hence there exists some point Φ^* such that $\lim_{t\to\infty}\Phi(t)=\Phi^*$.

□

Another point for proving that our algorithm is indeed practical is to prove that convergence occurs in finite time. Until now we have assumed that the finish time $T_f = \inf\{t \geq 0 \mid \Phi(t+1) = \Phi(t)\}$ may also be infinite. More specifically, we prove the following.

Lemma 12. *The NNIM Model converges in finite time.*

Proof. Eliminating recurrence via observing that the sum telescopes, we arrive at

$$V(t) = V(0) - \frac{1}{2k^2}\sum_{t=0}^{T}\sum_{w}\pi_w(t+1)\sum_{u,v\in K^{(t)}(w)}(\phi_u^{(t)}-\phi_v^{(t)})^2 \tag{7.43}$$

Using the definition for $\sigma^{(t)}(w)$ we can rewrite the above as

$$V(t) = V(0) - \frac{1}{2k^2}\sum_{t=0}^{T}\sum_{w}\pi_w(t+1)\sum_{u,v\in K^{(t)}(w)\setminus\sigma^{(t)}(w)}(\phi_u^{(t)}-\phi_v^{(t)})^2 \tag{7.44}$$

Since $V(t) \geq 0$ for every T, the negative difference term should vanish as $T\to\infty$. More specifically

$$\lim_{T\to\infty}\frac{1}{2k^2}\sum_{t=0}^{T}\sum_{w}\pi_w(t+1)\sum_{u,v\in K^{(t)}(w)\setminus\sigma^{(t)}(w)}(\phi_u^{(t)}-\phi_v^{(t)})^2 = 0 \tag{7.45}$$

Note that $\pi_w(t+1) > p$ for some $p\in(0,1)$ by the definition of the adjoint dynamics and $k>0$, hence we have a sum of squares with positive coefficients vanishing as $T\to\infty$. In order for this to happen, every individual term of the sum must go to 0. Therefore, for every $w\in U$, by the squeeze theorem

$$\lim_{T\to\infty}\sum_{u,v\in K^{(T)}(w)\setminus\sigma^{(T)}(w)}(\phi_u^{(T)}-\phi_v^{(T)})^2 = 0 \tag{7.46}$$

Again by the same argument for all $u,v\in\lim_{T\to\infty}K^{(T)}(w)$ for all $w\in U$

$$\lim_{T\to\infty}(\phi_u^{(T)}-\phi_v^{(T)}) = 0 \tag{7.47}$$

By the definition of NNIM the update process is continuous hence

$$\lim_{T\to\infty}\phi_u^{(T)} = \lim_{T\to\infty}\phi_v^{(T)} \tag{7.48}$$

as well as by the monotonicity of $V(t)$ we know that there exists some $\phi_w^*\in[0,1]$ such that

$$\lim_{T\to\infty} \phi_u^{(T)} = \lim_{T\to\infty} \phi_v^{(T)} = \phi_w^* \tag{7.49}$$

Hence $\lim_{T\to\infty} \phi_u^{(T)} = \phi_w^*$ for all $u \in \lim_{T\to\infty} K^{(T)}(w)$. Therefore for every $\epsilon_w > 0$ there exists some $T_w \geq 0$ such that for all $t \geq T_w$

$$|\phi_u^{(t)} - \phi_w^*| < \epsilon_w \qquad \forall u \in K^{(t)}(w) \tag{7.50}$$

Now we will prove finite time convergence via choosing the correct values for the ϵ's.

By Lemma 13 we know that if there exists a unique limiting point then it must be exactly approached in finite time. Suppose that there are $r \geq 2$ distinct limiting points $0 \leq \phi_1^* < \phi_2^* < \cdots < \phi_r^* \leq 1$. Now, fix $\epsilon > 0$. We know that for every $w \in U$ and $\epsilon_w = \epsilon$ there exists some finite $T_w \geq 0$ at which w reaches its limiting point within a distance of ϵ. Hence the maximum distance between two elements of $K^{(t)}(w)$ for $t \geq T_w$ is at most 2ϵ, by the triangle inequality, and the same applies for every pair of points that approach this limit. Let $W_1, \ldots, W_r \subseteq U$ be the subsets of U that approach their corresponding limits. From Theorem 5 these sets must contain consecutive agents. In order for finite convergence to occur we must impose a value of ϵ which splits the sets from each other. In this way, as we proved in Theorems 13 and 8, we will attain a finite convergence time.

First of all, let $T' = \max_{1\leq m\leq r} \max_{w\in W_m} T_w < \infty$ and let $D = \min_{i,j} \delta_{W_i W_j}^{(T')}$. A splitting will occur when the maximum distance between two points reaching the same limit, namely 2ϵ is less than the minimum distance D, hence $2\epsilon < D$. A good choice for ϵ will be the one which satisfies $2\epsilon + D < \min_{1\leq i\leq r-1}\{\phi_{i+1}^* - \phi_i^*\}$. Therefore, by these two conditions choosing $0 < \epsilon < \frac{1}{4}\min_{1\leq i\leq r-1}\{\phi_{i+1}^* - \phi_i^*\}$ isolates the sets $W_1, \ldots, W_r$, hence by Lemma 13 there exist $T_1, \ldots, T_r < \infty$ at which each W_i reaches its limit point. Now choose $T = T' + \max_{1\leq i\leq r} T_i + 1 < \infty$ and the proof is complete.

□

Determining a rigorous upper bound for the finishing time $T(n,k)$ can be obtained via a recursive (divide-and-conquer) proof of the above theorem. More specifically, when a cluster W_i becomes isolated then the sets $l(W_i) = \bigcup_{j<i} W_j$ and $r(W_i) = \bigcup_{j>i} W_j$ are also isolated by the metric properties of the minimum distance. The isolated cluster has at least k points and the rest has $n-k$ points. Let $n_L = |l(W_i)|, n_R = |r(W_i)|$ such that $n_L + n_R \leq n - k$ and assume that the difference between successive splits is $O(\tau)$. So, solving the problem for n agents is at most the running time for solving the problem with $n_L + n_R$ agents, since the solution treats the two sets as completely independent and does not need to recurse on each of the two sets. Therefore

$$T(n,k) \leq T(n_L + n_R, k) + O(\tau) \leq T(n-k,k) + O(\tau) \tag{7.51}$$

If $k = O(1)$ then convergence occurs in $O(n\tau/k)$ steps and if $k = (1-\varepsilon)n$ then convergence occurs in $O(\tau \log_{1/\varepsilon}(n))$ steps.

Lemma 13. *Suppose that the NNIM approaches (asymptotically) to a unique point ϕ^*, namely $\lim_{t\to\infty}\Phi(t) = \phi^*\mathbf{1}$. Then this point must be reached in finite time, i.e. there exists some (finite) $0 \leq T < \infty$ such that $\phi(T) = \phi^*\mathbf{1}$.*

Proof. At least one of the leftmost point or the rightmost point must have (in order for the one limit point to exist) a neighbor with different coordinate, to their right or to their left respectively. Since the points have continuous positions with preserved ordering there exists some finite time $0 \leq T < \infty$ at which they reach the same point ϕ^*.

□

Combining the above Lemmata we can state our main result for the NNIM model. Our proofs can extend to the multidimensional case by observing the convex-hulls of sets of agents for the isolation behaviour (instead of looking at the ordering as in 1D) and defining the same Lyapunov function, which now decomposes to each individual dimension.

Theorem 7. *The system of (7.21) converges in finite time under any consistent total ordering. Moreover, it suffices to perform $T = \lceil 2\log(1/D)/\log k\rceil$ iterations such that the total variation distance between the current state and the consensus state is strictly less than $d \cdot D$.*

Proof. We combine the finite time convergence result of Lemma 12 and the convergence rate of Lemma 4 to obtain a convergence rate of $o(k^{-t/2})$ a.a.s. for very large n. Moreover doing at least $T = \lceil 2\log(1/D)/\log k\rceil$ iterations guarantees a total variation distance (in 1D) strictly of at most D. In the case of the d-dimensional model the guarantee translates to a total variation distance of $d \cdot D$.

□

7.1.3 Complexity and Implementation

Table 7.2 gives an overview of the complexities for the various common data structures used for this problem, such as KD trees, Ball trees, Locality Sensitive Hashing (LSH), Dynamic Continuous Indexing (DCI) and Prioritized DCI (PDCI)[5]. Our implementation is developed in Python using Numpy, Sklearn, NetworkX and Annoy (for LSH) and experiments have been run on a Colaboratory Notebook. Figure 7.1 shows how NNIM scales with respect to the number of agents. The log-log plot for the specified hyperparameters ($k = \lceil \log n\rceil$, $D = 0.001$ and $d \in \{10, 100\}$ dimensions) scales as $n^{1.06\pm0.03}$; almost linearly.

7.2 Generalization Example: Multivariate Gaussian Opinions

The same approach that has been followed throughout the Thesis can be deduced if the agent's opinions follow Gaussian opinions. More precisely, as an introductory example

[5]We redirect the interested reader to Chapter 6 for more material.

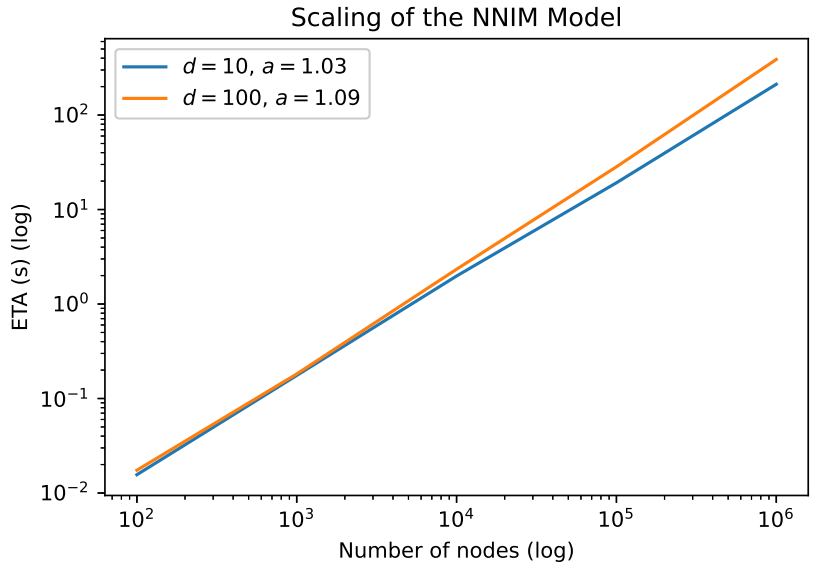


Figure 7.1: Log-log plot of the total time taken to perform inference to a network of up to 1M agents and $d \in \{10, 100\}$ with binary equiprobable artificial features, $D = 0.001$ and $k = \lceil \log n \rceil$. using LSH to obtain the nearest neighbors.

we will consider the case where the mean of each opinion vector is unknown and the covariance matrix is known and equal to Σ. After computing the parameters $\boldsymbol{\xi}_u^{(t+1)} = \frac{1}{k} \sum_{v \in \mathcal{K}^{(t)}(u)} \boldsymbol{X}_v^{(t)}$ — i.e. using the conventional Maximum Likelihood update rule — each agent draws an opinion from $\mathcal{N}\left(\boldsymbol{\xi}_u^{(t+1)}, \Sigma\right)$. Again — using the same procedure — one arrives at the objective function

$$\mathcal{L}_Q^{(t+1)} = \mathbb{E}_{Q^{(t)}} \left[\sum_{u \in U} \sum_{v \in \mathcal{K}^{(t)}(u)} \mathbf{1}\{v \in \mathcal{K}^{(t)}(u)\} \left[-\frac{1}{2} \log\left((2\pi)^d |\Sigma|\right) - \frac{1}{2} \left(\boldsymbol{X}_v^{(t)} - \boldsymbol{\xi}_u^{(t+1)}\right)^T \Sigma^{-1} \left(\boldsymbol{X}_v^{(t)} - \boldsymbol{\xi}_u^{(t+1)}\right) \right] \right]$$

Using a similar bound for the concentration of the normal variable norm [98] and performing the algebraic operations as in the previous Section we arrive at the same result for the variational means $\boldsymbol{\phi}_u^{(t)}$ as in the Bernoulli case. Again the mean of opinions $\boldsymbol{\mu}^{(t)} = \frac{1}{n} \sum_{u \in U} \boldsymbol{\xi}_u^{(t)}$ can be calculated as the mean of the variational parameters and the covariance matrix $\boldsymbol{S}^{(t)} = \frac{1}{n} \sum_{u \in U} (\boldsymbol{\xi}_u^{(t)} - \boldsymbol{\mu}^{(t)})(\boldsymbol{\xi}_u^{(t)} - \boldsymbol{\mu}^{(t)})^T$ can be calculated as

$$\boldsymbol{S}^{(t)} = \frac{1}{n} \sum_{u \in U} \left(\boldsymbol{\phi}_u^{(t)} - \boldsymbol{\mu}^{(t)}\right) \left(\boldsymbol{\phi}_u^{(t)} - \boldsymbol{\mu}^{(t)}\right)^T \tag{7.52}$$

7.3 Experiments

7.3.1 Datasets

We use the following datasets for evaluating our method on

Table 7.2: Complexity of nnim with under various data structures (Brute-force, KD-tree, Metric Ball, LSH) for running the nnim model such that the total variation distance is at most $d \cdot D$ after execution. State-of-the-art is DCI and Prioritized DCI [63, 64]. The quantity d' is the intrinsic dimension [64, p.1]. The number m is the number of projection directions used in the DCI.

Data structure	Complexity	Notes
Brute-force	$O\left(nd(n+k)\log(1/D)\log^{-1}k\right)$	Efficient for very small n
KD/Ball tree [10, 85]	$O\left(nd(n^{1-1/d}+k)\log(1/D)\log^{-1}k\right)$	Efficient for $d \ll \log n$
LSH [41]	$O\left(n^{1+1/(1+\epsilon)^2}dk\log(1/D)\log^{-1}k\right)$	$(1+\epsilon)$-approximation
DCI/PDCI [63, 64]	$O\left(\left(mn\log\left(\frac{n}{k}\right)+\left(\frac{n}{k}\right)^{2-\frac{m}{d'}}\right)\frac{dk\log(1/D)}{\log k}\right)$	Efficient for large n and d'

facebook [60, 62]. Contains an ego-network of user 107 in the Facebook network. Friendships in Facebook are undirected. To avoid the obvious domination by the ego node, we have removed the outcoming links of the ego node and kept the incoming links.

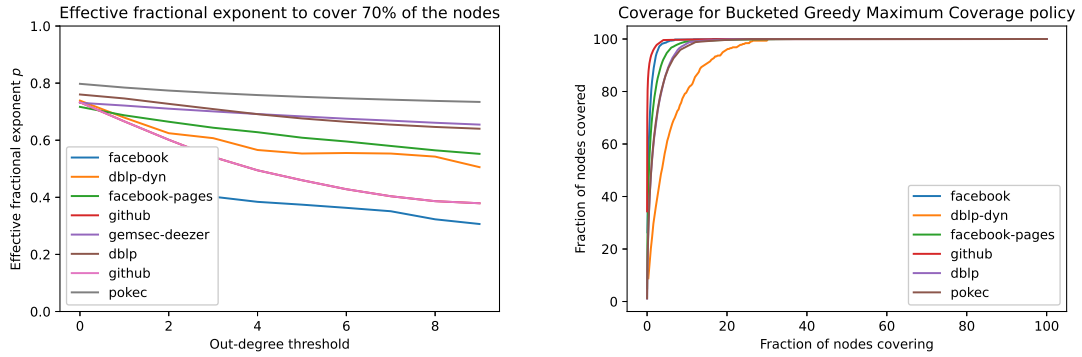
dblp-dyn [27]. Vertices of the graph are authors and an edge exists between them if the corresponding authors have written a paper together in a given period of time. Only authors who had at least 10 publications (in a selected set of 43 conferences/journals) from 1990 to 2010 are considered. There are in total 2,723 authors. Each vertex at each time is associated to a set of 43 attributes corresponding to the number of publications in each conference/journal during the related period. We have chosen to keep the period between 1994 and 1998.

facebook-pages [60, 94]. A page-page graph of verified Facebook pages, with nodes representing pages and links are mutual likes between them. The pages belong to four categories defined by facebook (oliticians, governmental organizations, television shows, companies).

github [60, 94]. Social network of GitHub developers as of June 2019 who have starred at least 10 repositories and edges are mutual follower relations between them. All users in this dataset have one label, whether the user is a web or a machine learning developer.

dblp [90]. This data set depicts a co-authorship graph built from the DBLP digital library. Each vertex represents an author who published at least one paper in one of the major conferences and journals of the Data Mining and Database communities between January 1990 and February 2011. Each edge links two authors who co-authored at least one paper (no matter the conference or journal). The labels are the number of publications in each of the 29 selected conferences or journals.

pokec [60, 103]. Pokec is a popular Slovakian social network, still active, despite the existence of considerably larger social networks, such as Facebook, containing 1.6 million users. Datasets contains anonymized data of the whole network. We extracted the labels



(a) (b)

Figure 7.2: Left: Engagement Threshold Effect. Right: Coverage curve for the BGMC policy for $\tau = 4$.

of Pokec manually using the information provided by the user profiles. More specifically, we kept the `hobbies` column and kept the 280 most common hobbies. We also removed the nodes that have not disclosed their profile information and connections and had their `public` attribute equal to 0.

7.3.2 Influencer Identification

We need a systematic way to identify the core set C of influencers of the OSN. Our problem is similar to the *Maximum Coverage* (MC) problem in combinatorial optimization, since we set a target amount and attempt to maximize the covered users with the number of influencers in question, which is NP-Hard [83, 34], and the greedy algorithm which proceeds in rounds and chooses the node with the maximum number of uncovered neighbors yields an optimal approximation ratio of $1-1/e$. Running the greedy algorithm ad-hoc has a very high computational cost as the number of nodes increases. For this reason, we reside on a fork of the original algorithm which we call Bucketed Greedy Bucketed MC (BGMC). In the BGMC setting, we have an upper bound K of nodes we want to use in our coverage. We sort the nodes according to their in-degree and put them into $\log(n/K)/\log\gamma$ non-uniform buckets $V_1, \ldots, V_r, \ldots$ of sizes $\lceil \gamma K \rceil, \ldots, \lceil \gamma^r K \rceil - \lceil \gamma^{r-1} K \rceil, \ldots$, for some $\gamma > 1$. We then start by constraining the neighborhoods of vertices to V_1 and run the greedy maximum coverage algorithm on it. If we either cover all the nodes or exshaust the K choices we return. Otherwise, we continue the same using the set V_2, and so on, via removing the already covered nodes at each iteration. Although it is evident that the BGMC algorithm does not in general yield a solution set that equals the conventional greedy solution and has a strictly lesser approximation ratio, the algorithm yields remarkably good results when run on OSN. More specifically, for a threshold value $\tau = 4$, a population of $n^{0.7}$ influencers dominate about $74.01 \pm 14.91\%$ of the networks in question (see Table 7.3)

7.3.3 Further Processing

To avoid dealing with high dimensionality prior to running the INFERENCE_NNIM procedure, we perform dimensionality reduction (PCA) keeping a 95% of the explained variance. After running the algorithm, we invert the transformation and clip the variables that fall outside $[0, 1]$.

7.3.4 Experimental Setting

To test the performance of NNIM, we reside in the multilabel classification task, where given a partially binary-labeled graph, we aim to predict the missing labels. We perform a same-input-same-output comparison, where our input consists of the bipartite influencer-user graph and the influencer labels and the desired output are the scores to be predicted. More specifically, for each node u and each label $1 \leq i \leq d$ we attribute a score in $\phi_{iu} \in [0, 1]$ that represents the probability that the user adopts that label (interest). We gather the influencers of the network using the BGMC heuristic and keep the bipartite graph between the influencers and the rest of the network. We use a thresholding value of $\tau = 4$ and an exponent of $p = 0.7$ as shown in Figure 7.2. We run experiments with $k \in \{\lceil\sqrt{n}\rceil, \lceil\log n\rceil\}$ neighbors, with and without Regularization (where we use the initial state as weighted extra opinion). In our experiments we have used LSH to infer the k nearest neighbors. Firstly, we compare our method with the Random HK model described in [36] which is the model that most closely resembles our work. Instead of looking at the k-nearest neighbors, Random HK picks a random subset of k neighbors within a radius ε of the user. Secondly, we train node2vec [43], GraphWave [29] and NodeSketch [111] embeddings on the same graph and then fit a multilabel logistic regression model. This kind of benchmark is almost standard, as we discuss in the Related Work Section, in graph mining. We chose node2vec as a classical random-walk-based approach, GraphWave as a transformation-based approach, and NodeSketch which is a new method based on recursive sketching. We report the AUC-ROC [32] between the ground-truth values and the predicted values for 100% of the labels, top-50% and the top-1 interests, the RMSE between the ground-truth interest distribution (sample means) and the methods' final interest distribution (means), the coverage percentage of the engaged network by the members of the core, the number of core members, and the runtime for the pokec experiment. The AUC-ROC metric quantifies the *quality of ranking* whereas the RMSE quantifies the results' *accuracy*.

7.3.5 Discussion

We report descent results in terms of AUC-ROC and RMSE in all of our experiments: In the facebook dataset we have the best performance in terms of RMSE and have AUC near the other methods; less than 1% for all labels, and similar results for top-50% and

top-1. In the dblp-dyn, fb-pages and github[6] dataset we outperform the other methods — with the exception of the AUC-ROC in top-50% in dblp-dyn where we have a 4% percent decrease. Moreover, in the fb-pages dataset, GraphWave achieves a very small RMSE however it yields a low AUC-ROC by far. Finally, in the pokec network, GraphWave and Random HK fail to run subject to our resources[7]. Moreover, the NNIM model runs two orders of magnitude faster with $k = \lceil \log n \rceil$ neighbors and one order of magnitude faster with $k = \lceil \sqrt{n} \rceil$ neighbors compared to node2vec and NodeSketch. The PCA step does not affect the runtime considerably needing only 1 sec since it fits only on the highly influential nodes that are $n^{0.7}$, which account for 1.92% of the network. We achieve an AUC-ROC of 91.84% and an RMSE of 0.025 where we surpass NodeSketch in terms of RMSE (6 times lower) and are surpassed in terms of AUC-ROC by 0.3%. Finally, node2vec has a higher AUC-ROC rate (by a small margin) compared to NNIM with $k = \lceil \sqrt{n} \rceil$ neighbors.

7.3.6 More Experiments

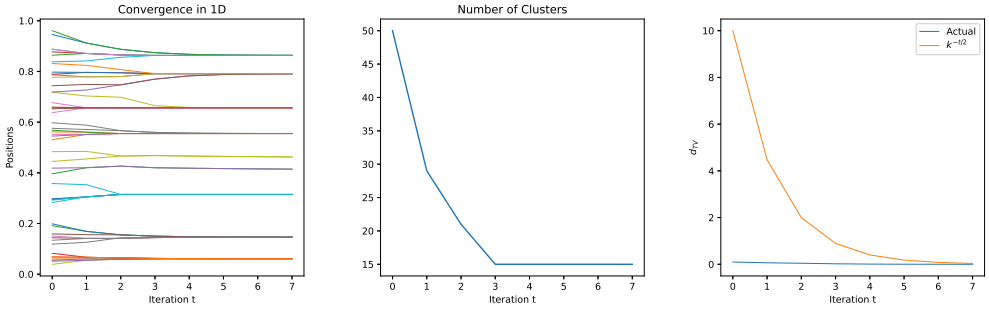
Mean-field equations properties

We perform experiments with synthetic data so as to obtain a better understanding of the convergence properties and the number of clusters of NNIM. We sketch the main conclusions of the experimental evaluation, which confirm and enhance our theoretical results. We initialize a set of $n = 100$ agents with d-dimensional opinions for values of d between 1 and 10 and number of neighbors k between 2 and 50. We plot the convergence time versus the number of neighbors k and the dimension d of the vectors as well as the final number of clusters that are formed upon convergence. The *microscopic properties* of the model, namely the agent positions, the number of clusters and the total variation distance compared to the strict upper bound $k^{-t/2}$ are presented in Figure 7.3. The *macroscopic properties* — namely how the convergence time and the number of clusters change with respect to either varying number of neighbors k — are presented in Figure 7.4 for $D = 10^{-7}$.

The behaviour of the NNIM model closely resembles the behaviour of the HK model [46, 81], the Random HK model and the Network HK model described in [36]. Remarkably, the convergence time decreases rapidly subject to increasing k in the form of a *power law* — namely the number of clusters is proportional to k^a. To observe this, we provide a log-log polynomial fit between the number of clusters and the number of nearest neighbors k for values of n in the range between 50 and 300 with a step of 50 agents. The observed values of the slope a are around -1.06 for all the values of the agents and the bias term increases as n increases which further validate the claim that the model has converged if and only if every cluster contains at least k agents, hence the total number of clusters is at most n/k.

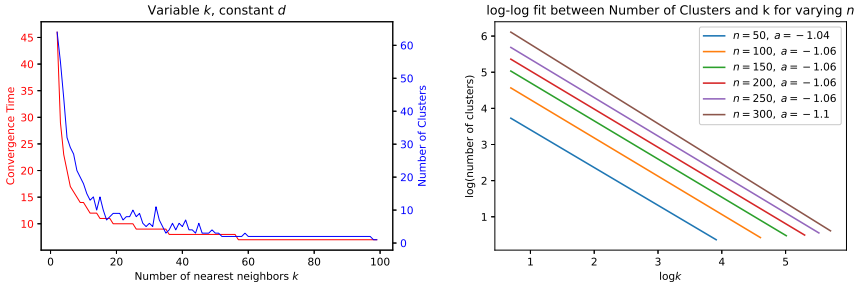
[6] The dataset contains one label hence AUC-ROC results remain the same.

[7] Denoted by the dagger (†) symbol. Experiments were run in a Google Colab Notebook.



(a) Agent positions (b) Number of clusters (c) Convergence Rate

Figure 7.3: Microscopic properties of the NNIM model for $n = 100$ agents, $D = 10^{-3}$, and $k = 3$ neighbors.



(a) Variable k, constant d (b) log-log plot between number of clusters and k

Figure 7.4: Macroscopic Properties of the NNIM model. We have run the with $D = 10^{-7}$.

Sentiment in the Linux Kernel Mailing List

Through this experiment, we evaluate our model's ability to predict sentiment in discussions. More specifically, we examine the Linux Kernel Mailing List (LKML) [8] archive between August 2017 and October 2017. The LKML dataset contains 18K email threads from Linux Kernel developers which are organized in threads and each email posseses had a timestamp and an author. Initially, in each thread we record the number of participants in the thread as well as the polarity of each email. We use the open-source Natural Language Processing Library TextBlob [66] to extract the polarity sentiment for each email — after we have removed the nested replies which start with the character >. For each participant in each thread we record his/her initial opinion upon his/her first message as well as his/her final opinion upon his final message. We give a value of 1 if the sentiment — which lies in $[-1, 1]$ — is positive and a value of 0 if the sentiment is negative. In order to avoid bias in our results, we filter out the threads which contain less than 5 participants and the threads for which no participant has changed his/her opinion. Moreover, we perform experiments by letting k run from 1 to the number of participants in the thread. We report the average *Mean Average Precision* (MAP) and compare our model to the Random HK model with radius $\varepsilon = 1$. We report a MAP of 88.07% using NNIM and a MAP of 81.36% using the Random HK model.

Hyperparameters Effect

We examine how the hyperparameters k and p affect the AUC-ROC and RMSE metrics. For the ego-facebook network we report the AUC-ROC and RMSE metrics for numbers of p between 0.4 and 0.9 with step 0.1 and k from 1 to 200 with step 1. Results are presented in Figure 7.5.

7.4 Further Related Work

Core-periphery structure of networks has mainly gathered attention from socio-economical [109, 58, 69] and network modeling perspectives [82, 115, 9]. Computer science literature is mostly concentrated in learning core-periphery models. From an algorithmic perspective, the closest work to ours is [6], where Avin et al. show how to speed up tasks in a distributed setting. However, they do not provide an algorithm for efficiently identifying the core in large networks, as we do in this work.

Opinion Dynamics models have been around for decades, with the best known being the DeGroot model [24], the Friedkin-Johnsen (FJ) model [38], and the HK model [46]. In [46, 11], the agents' opinions evolve as a discrete dynamical system and the opinions at the next timestep are the result of an aggregation of ones and her neighbors' opinions, where the neighborhood is built dynamically from the observations of the current timestep. [36] considers different models with the addition of local interactions between the agents,

[8] http://lkml.org

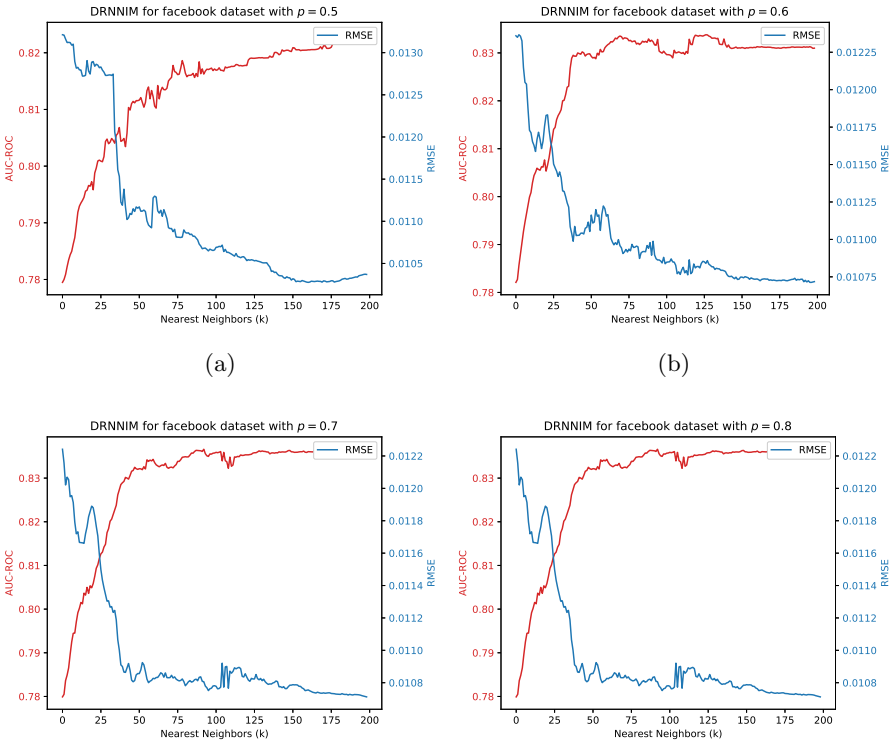


Figure 7.5: Effect of parameters and negative correlation between k and p on the ego-facebook dataset. The values of p range from 0.4 to 0.7 with 0.1 step and the values of k range from 1 to 200 with step size of 1.

with the Random HK model being conceptually closest to our model. In the Random HK model, each agent chooses uniformly at random k neighbors from a ball of radius ε centered at her opinion. Our work develops a *stochastic variant* of this family of models, thus generalizing existing deterministic ones.

Multilabel classification in graphs has a relatively long history. To begin with, the classical work on label propagation [91] infers community memberships in networks via propagating labels between the nodes until a consensus is reached. Besides, similar work in [62, 112] devises a random graph model to classify nodes with features within communities. Moreover, the upsurge of embedding methods, which use random walks, matrix factorization-based learning objectives, or signal processing transformations [43, 89, 29, 111, 33, 94, 44] has been used for multilabel classification. Multilabel classification with embeddings as a *standardized benchmark task* for evaluating embedding methods uses them as inputs to a supervised model, usually *logistic regression*. The input graph nodes typically have features in a high-dimensional space, whereas the target labels lie in a low-dimensional space. In contrast, in our work, inputs and outputs have the same dimensionality.

Our work is also related to **inference in probabilistic graphical models** with latent variables and with a likelihood that cannot be computed in a computationally efficient manner, because integration for the latent variables significantly affects the running time. Some characteristic examples are the MAG Model in OSN [55, 54] and training of HMMs [8] with the EM algorithm [26]. The EM algorithm maximizes the expectation of the joint likelihood of the data by imposing a distribution over the latent variables. We use the mean-field approximation in our paper [50, 101], a technique that is widely used in the statistical physics community.

7.5 Conclusion

In this Thesis, we benefit from the core-periphery structure of OSN and develop inference algorithms for interest prediction (equivalently multilabel classification) using partial information from influential users. Inspired by the strong homophilic properties of OSN, we introduce the NNIM model. This model considers a core in the steady-state and a periphery that exchanges opinions according to k nearest neighbors. We develop an algorithm for computationally efficient inference and establish a connection with traditional models, such as the HK. We prove that our algorithm converges in finite time and strictly bound the total variation distance from the consensus state. Our method is compared with others and in networks of various sizes and is capable of performing considerably faster with similar and most of the times better results.

7.6 Future Work

This Thesis' work can be extended to multiple interesting future directions. First and foremost, the utilization of the core-periphery structure to speedup algorithms can be extended to other problems as well. Example problems are all-pairs shortest paths (finding betweenness centrality measures in a network), ranking users in a network (e.g. PageRank) and speeding random-walk based algorithms. Moreover, concrete understanding of core-periphery structure through intuitive generative models is also an open line-of-work.

Furthermore, our works gives a statistical explaination for opinion dynamics, extending the existing game-theoretical understanding of the opinion formation processes [12, 11]. Interestingly, extending this line of work to account for more general settings (e.g. exponential families) and models could yield significant results in graph learning tasks and inference algorithms.

Table 7.3: Experimental results with $p = 0.7$, $\gamma = 2$, $D = 10^{-3}$, $\tau = 4$, and regularization with $\alpha = 1$.

	facebook	dblp-dyn	fb-pages	github	dblp	pokec	Time(s)
				AUC-ROC (all labels)			
node2vec	**86.35**	87.42	84.00	67.23	69.80	**96.93**	$\sim 10^3$
GraphWave	86.20	86.78	70.96	45.13	69.57	†	†
NodeSketch	80.90	81.90	68.68	49.96	58.88	92.14	$\sim 10^3$
Random HK ($k = \lceil \log n \rceil$)	85.75	86.30	71.90	50.34	68.83	†	†
NNIM ($k = \lceil \log n \rceil$)	84.24	88.05	**91.86**	68.07	78.64	85.60	$\sim$ **10^1**
NNIM ($k = \lceil \sqrt{n} \rceil$)	85.82	**91.16**	91.62	67.86	**81.65**	91.84	$\sim 10^2$
NNIM w/ Reg ($k = \lceil \log n \rceil$)	84.17	87.39	91.78	**72.31**	78.86	85.05	$\sim$ **10^1**
				AUC-ROC (top 50% of labels)			
node2vec	54.98	**94.92**	78.69	67.23	68.53	**96.94**	$\sim 10^3$
GraphWave	53.97	92.91	40.11	45.13	65.70	†	†
NodeSketch	55.91	92.37	46.50	49.96	58.13	92.14	$\sim 10^3$
Random HK ($k = \lceil \log n \rceil$)	52.82	93.10	56.14	50.34	64.49	†	†
NNIM ($k = \lceil \log n \rceil$)	59.08	79.32	**89.00**	68.27	78.69	85.80	$\sim$ **10^1**
NNIM ($k = \lceil \sqrt{n} \rceil$)	58.30	90.59	88.04	67.86	**80.85**	91.84	$\sim 10^2$
NNIM w/ Reg ($k = \lceil \log n \rceil$)	**59.20**	81.11	88.65	**72.31**	79.10	85.05	$\sim$ **10^1**
				AUC-ROC (top-1 label)			
node2vec	52.56	62.82	80.17	67.23	60.28	**55.87**	$\sim 10^3$
GraphWave	**57.19**	67.00	61.37	45.13	52.89	†	†
NodeSketch	53.02	63.06	59.07	49.96	49.22	50.78	$\sim 10^3$
Random HK ($k = \lceil \log n \rceil$)	50.17	48.40	49.48	50.34	49.96	†	†
NNIM ($k = \lceil \log n \rceil$)	53.29	82.89	90.18	68.27	70.31	54.64	$\sim$ **10^1**
NNIM ($k = \lceil \sqrt{n} \rceil$)	53.62	**84.16**	**90.38**	67.86	**71.27**	55.34	$\sim 10^2$
NNIM w/ Reg ($k = \lceil \log n \rceil$)	51.52	80.47	90.35	**72.31**	70.71	54.59	$\sim$ **10^1**
				RMSE (all labels)			
node2vec	0.012	0.059	0.093	0.438	0.166	**0.022**	$\sim 10^3$
GraphWave	**0.010**	0.052	7e-6	0.400	**0.082**	†	†
NodeSketch	0.096	0.123	0.098	0.440	0.316	0.128	$\sim 10^3$
Random HK ($k = \lceil \log n \rceil$)	**0.010**	0.056	**4e-17**	0.412	0.096	†	†
NNIM ($k = \lceil \log n \rceil$)	0.011	0.062	**4e-17**	0.389	0.143	0.026	$\sim$ **10^1**
NNIM ($k = \lceil \sqrt{n} \rceil$)	**0.010**	**0.050**	**4e-17**	**0.388**	0.128	0.025	$\sim 10^2$
NNIM w/ Reg ($k = \lceil \log n \rceil$)	0.012	0.066	4e-16	**0.388**	0.145	0.025	$\sim$ **10^1**
Coverage (%)	88.36	97.16	72.20	68.61	66.04	51.70	—
Influencers (Core size) (%)	12.47	11.83	4.94	4.23	4.12	1.92	—

Appendix A

Concentration Bounds

"Years ago a statistician might have claimed that statistics deals with the processing of data. Today's statisticians will be more likely to say that statistics are concerned with decision making in the face of uncertainty."
— Herman Chernoff

A.1 Motivation

The study and analysis of algorithms usually requires the study of random variables that are composed from sums of usually independent and identically distributed random variables (i.i.d.). The exact distributional properties of the sum variables are usually difficult to identify analytically. Moreover, the analysis of algorithms usually has to do with properties referring to the *average performance* of a quantity. Luckily — for all of us who analyze algorithms frequently — randomness can be adequately limited, that is that the random variables usually do not deviate from their mean. The more characteristic examples are the Weak Law of Large Numbers which states that the sample average of inifinitely many samples of i.i.d. random variables from a distribution $\mathcal{D}$ converges to the expectation $\mathbb{E}_{\mathcal{D}}[X]$. Intuitively, if we let $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $X_i \sim \mathcal{D}$ are i.i.d. with probability density function $p_S(s)$, then we will observe that $p_S(s)$ contains a very large amount of probability mass near $\mathbb{E}_{\mathcal{D}}[X]$. Even though we can spend multiple pages describing similar bounds, we will reside in explaining the idea behind them as well as introduce the main tool we will use in our analysis — namely Talagrand's Inequality — since further describing them is out of the scope of this thesis. For a more detailed investigation of the Chernoff-type bounds family we redirect the interested reader to [98].

The most intuitive way to study these properties is to study the tail probability, namely how small is the deviation outside this region. The tail probability of a random variable is the probability $\Pr[|X - \mu| > \lambda]$ for some $\mu \in \mathbb{R}$ and $\lambda \geq 0$. The first tool that can be used to study the tail probabilities is *Markov's Inequality*, stated below:

Theorem 8 (Markov's Inequality)**.** *Let $X \geq 0$ be a random variable and λ be a positive*

real number. Then

$$\Pr[X \geq \lambda] \leq \frac{\mathbb{E}[X]}{\lambda} \tag{A.1}$$

Proof. From basic proability theory we know that

$$\mathbb{E}[X] = \int_0^\infty x p_X(x) dx \geq \lambda \int_\lambda^\infty p_X(x) dx = \lambda \Pr[X \geq \lambda]$$

Rearranging terms we arive at

$$\Pr[X \geq \lambda] \leq \frac{\mathbb{E}[X]}{\lambda}$$

□

This inequality, albeit seeming simple in principle, posseses some very powerful properties, most importantly its generalization upon a monotone function $f(x|t)$ of $x \geq 0$ which is parametrized by $t \in \mathcal{T}$. More specifically, using the monotonicity of f inside the probability we can refer to the more general bound of

$$\Pr[f(X|t) \geq f(\lambda|t)] \leq \sup_{t \in \mathcal{T}} \frac{\mathbb{E}[f(X|t)]}{f(\lambda|t)} \tag{A.2}$$

A characteristic example is when f is taken to be the squared distance from the mean, i.e. $f(X) = (X - \mu)^2$ where $\mu = \mathbb{E}[X]$ is the expectation of the random variable X. Using $\lambda = k^2 V(X) > 0$, we arrive at the well known *Chebysev Inequality* that is

Theorem 9 (Chebysev's Inequality)**.** *Let $X > 0$ be a random variable with expectation μ and standard deviation $\sigma = \sqrt{V(X)}$ and some $k > 0$. Then*

$$\Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2} \tag{A.3}$$

Proof. We study the random variable $Y = (X - \mu)^2 \geq 0$ with $\mathbb{E}[Y] = V(X)$ and pick $\lambda = k^2 V(X)$. □

Now, we begin to grasp the reason for which this inequality is so strong. More specifically, the study of the moment generating exponential $f(X|t) = \exp(tX)$ for some $t > 0$ leverages Markov's Inequality power. More precisely, in a similar way, we are going to introduce the Chernoff-Hoeffdig Bound

Theorem 10 (Chernoff-Hoeffdigg (CH) Bound)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $\mathbb{E}[X_i] = p$ for $1 \leq i \leq n$ and $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ and some $\varepsilon > 0$. Then*

$$\Pr[|\hat{p} - p| > \varepsilon] \leq 2\exp(-2n\varepsilon^2) \tag{A.4}$$

Proof. We have that

$$\mathbb{P}[\hat{p} > p + \varepsilon] = \mathbb{P}\left[e^{t\hat{p}} > e^{p+\varepsilon}\right] \leq \frac{\mathbb{E}\left[e^{tX_1}\right]^n}{e^{n(p+\varepsilon)}} \leq \min_{t>0} \frac{\mathbb{E}\left[e^{tX_1}\right]^n}{e^{n(p+\varepsilon)}}$$

where $\mathbb{E}\left[e^{tX_1}\right] = p(e^t - 1) + 1$. The function at the right hand side has a minimum at

$$t_0 = \ln \frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}$$

With value

$$e^{-D(p+\varepsilon||p)n} \leq e^{-2n\varepsilon^2}$$

The second inequality is derived from the variable sequence $Z_i = 1 - X_i$ which are Bernoulli independent variables with parameter $1 - p$. Combining both inequalities we have that

$$\Pr[|\hat{p} - p| > \varepsilon] = \Pr[\hat{p} > p + \varepsilon \vee \hat{p} < p - \varepsilon] \leq \Pr[\hat{p} > p + \varepsilon] + \Pr[\hat{p} < p - \varepsilon] \leq 2\exp(-2n\varepsilon^2)$$

□

It is not hard to show that the above tail bound generalizes for i.i.d. random variables that lie in $[0, 1]$ with the same proof technique as well as in $[\ell, u]$, for which we need to study the normalized variables $Y_i = \frac{X_i - \ell}{u - \ell} \in [0, 1]$. Other useful bounds include the bound for the χ^2 variables which is used to prove the infamous Johnson-Lindenstrauss Random Projection Lemma which is widely used for dimensionality reduction in database systems, introduced in reference [3], Bennet's Inequality, Azuma's Inequality, Bernstein's Inequality and many more [98].

Example: Sample Complexity of the Bernoulli MLE. Suppose that we are running an election poll with two parties — red and blue — which are represented with the values 0 and 1 respectively. We are interested in calculating the probability that the blue party wins the election race. For that reason we ask n people and gather their preferences $X_1, \ldots, X_n$ and calculate the sample mean $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The basic question that rises here is how many people should we ask beyond which we are confident with probability $1 - \delta$ that $\hat{p}$ will be within $\varepsilon > 0$ distance from the actual value p. The answer comes via directly employing the CH Bound and restricting the error probability to be less than or equal to δ. Rearranging terms we arive that $n \geq \left\lceil \frac{\log(2/\delta)}{2\varepsilon^2} \right\rceil$ samples[1] are needed to achieve the desired result.

Finally, we give the tail bound for the χ^2 variables so that we are able to study the behaviour of random normal vectors. A variable Z is said to follow the χ^2 distribution if it is composed out of a sum of $\mathcal{N}(0, 1)$ variables.

Theorem 11 (Concentration of χ^2 Variables)**.** *Let $Z \sim \chi^2(n)$. Then for all $\varepsilon > 0$ we have*

$$\Pr[Z \leq (1 - \varepsilon)n] \leq \exp(-\varepsilon^2 n/6) \tag{A.5}$$

For all $\varepsilon \in (0, 3)$ we have that

[1] The ceiling of x, that is $\lceil x \rceil$, is defined as the value of x rounded up to the nearest integer.

$$\Pr[Z \geq (1+\varepsilon)n] \leq \exp(-\varepsilon^2 n/6) \tag{A.6}$$

and

$$\Pr[|Z-n| \geq n\varepsilon] \leq 2\exp(-\varepsilon^2 n/6) \tag{A.7}$$

Proof. Let $Z = \sum_{i=1}^{n} X_i^2$ where X_i are i.i.d. $\mathcal{N}(0,1)$-distributed random variables. From basic calculus we know that $\exp(-a) \leq 1 - a + a^2/2$ for all $a \geq 0$. Therefore

$$\mathbb{E}\left[\exp(-tX_1^2)\right] = 1 - t\mathbb{E}\left[X_1^2\right] + \frac{t^2}{2}\mathbb{E}\left[X_1^4\right] = 1 - t + \frac{3}{2}t^2 \leq \exp\left(-t + \frac{3}{2}t^2\right) \tag{A.8}$$

Therefore

$$\mathbb{E}\left[\exp(-tZ)\right] \leq \exp\left(-nt + \frac{3}{2}nt^2\right) \tag{A.9}$$

For $t = \varepsilon/3$ we get the first inequality. For the sencond, inequality we reside in

$$\mathbb{E}\left[\exp(tZ)\right] \leq (1-2t)^{-n/2} \tag{A.10}$$

for all $t < 1/2$. Hence $\Pr[Z \geq (1+\varepsilon)n] \leq \exp(-\varepsilon n t)$. For $t = \varepsilon/6 < 1/2$. The third inequality is obtained by applying a Union Bound to the above two inequalities.

□

A.2 Talagrand's Inequality

Next, we introduce the main inequality upon which we base some of our contributed theorems. The inequality we are going to study is due to Talagrand [104] and is of special importance to Measure Theory and subsequently Probability Theory. In layman's words, it states that the image of a bounded random variable through a Lipschitz function F is concentrated around its mean. For completeness purposes, we give the definition of a Lipschitz function below

Definition 9 (Lipschitz Function)**.** *Given two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ where $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the metrics for each of the spaces, a function $f : \mathcal{X} \to \mathcal{Y}$ is called L-Lipschitz if there exists a real constant $L \geq 0$ such that for all $x_1, x_2 \in \mathcal{X}$ the following condition is true*

$$d_{\mathcal{Y}}\left(f(x_1), f(x_2)\right) \leq L d_{\mathcal{X}}(x_1, x_2) \tag{A.11}$$

Intuitively, the images of the points x_1 and x_2 cannot be arbitrarily far apart. When $L \in [0,1)$ the mapping f is called *Lipschitz contraction* and when $L = 1$ the function is called a *short map*. Here, it is evident that we are interested in the best (tightest) possible value of the constant L.

Example 1: The function $f(t) = \sin t$. A simple example to demonstrate Lipschitzness in 1 Dimension with respect to the absolute value $d_{\mathcal{X}}(x, y) = d_{\mathcal{Y}}(x, y) = |x - y|$ is the function $f(t) = \sin t$. To elaborate, the function's derivative is $f'(t) = \cos t$ which is absolutely (and tightly) bounded by 1. By applying the Mean Value Theorem in the interval $[x, y]$ we have that there exists some $\xi \in (x, y)$ such that $f'(\xi) = \cos \xi = \frac{\sin y - \sin x}{y - x}$. Since $|f'(\xi)| \leq 1$ we obtain that $|\sin y - \sin x| \leq |y - x|$ for all $x, y \in \mathbb{R}$ and therefore $L = 1$.

Example 2: Contractive Mappings. Contractive mapping demonstrate remarkable properties. The most notable one is that contraction mappings have fixed points. A fixed point $\boldsymbol{x} \in A$ of a function $f : A \to A$ is a point such that $f(\boldsymbol{x}) = \boldsymbol{x}$. It has been shown by Banach that contractions have exactly one fixed point $\boldsymbol{x}^*$. We state *Banach's Fixed Point Theorem* below:

Theorem 12 (Banach's Fixed Point Theorem)**.** *Let $(\mathcal{X}, d)$ be a non-empty complete metric space and $f : \mathcal{X} \to \mathcal{X}$ be a contraction. Then f has a unique fixed point $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) = \boldsymbol{x}^*$. Moreover, the fixed point can be found by starting from an initial position $\boldsymbol{x}_0 \in \mathcal{X}$ and then use the update rule $\boldsymbol{x}_{n+1} = f(\boldsymbol{x}_n)$ such that $\lim_{n\to\infty} \boldsymbol{x}_n = \boldsymbol{x}^*$.*

Recently, a converse to Banach's Fixed Point Theorem was established by Daskalakis, Tzamos and Zampetakis in reference [23] which states that

Theorem 13 (Due to [23])**.** *Let $(\mathcal{X}, d)$ be a proper metric space and $f : \mathcal{X} \to \mathcal{X}$ be continuous with respect to d and the following hold*

1. *f has a fixed point $\boldsymbol{x}^* \in \mathcal{X}$*
2. *for every $\boldsymbol{x}_0 \in \mathcal{X}$ the sequence $\boldsymbol{x}_{n+1} = f(\boldsymbol{x}_n)$ converges to $\boldsymbol{x}^*$ with respect to d and there exists an open neighborhood U of $\boldsymbol{x}^*$ such that $f^{[n]}(U) \to \{\boldsymbol{x}^*\}$*

Then, for every $c \in (0, 1)$ and $\varepsilon > 0$ there exists a function $d_{c,\varepsilon}$ that is topologically equivalent to d such that

1. *f is a contraction with respect to $d_{c,\varepsilon}$*
2. *$d_{c,\varepsilon}(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon \implies \min\{d(\boldsymbol{x}, \boldsymbol{x}^*), d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{x}^*)\} \leq 2\varepsilon$*

We will refer to this theorem later in this Thesis, when we will study the *stability of dynamical systems*. It is now evident, that contractive mappings are tightly related with the convergence of iterative processes; algorithms in our framework.

A.2.1 Statement of Talagrand's Inequality

The importance of Lipschitz functions comes to the forefront in Probability Theory as well. The powerful inequality states that the image of a contractive mapping is well-concentrated around its mean. We state the inequality here

Theorem 14 (Talagrand's Inequality (due to [104, 105]))**.** *Let $X_1, \dots X_n$ be independent complex random variables — not necessarily identically distributed — and some $K > 0$ such that $|X_i| \leq K$ for all $1 \leq i \leq n$. Let $F : \mathbb{C}^n \to \mathbb{R}$ be a 1-Lipschitz function. Then for any $\lambda > 0$ one has*

$$\Pr[|F(X_1, \dots X_n) - \mathbb{E}\left[F(X_1, \dots, X_n)\right]| \geq \lambda K] \leq C \exp(-c\lambda^2) \tag{A.12}$$

For some constants $c, C > 0$.

For the interested reader, the lengthy proof can be found at [105, pp. 86-91], since the tools and lemmas used to prove it lie beyond the scope of this Thesis. A direct Corollary of the inequality is the following

Corollary 3. *Let $X_1, \dots X_n$ be independent complex random variables — not necessarily identically distributed — and some $K > 0$ such that $|X_i| \leq K$ for all $1 \leq i \leq n$. Let $F : \mathbb{C}^n \to \mathbb{R}$ be a L-Lipschitz function for some $L > 0$. Then for any $\lambda > 0$ one has*

$$\Pr[|F(X_1, \dots X_n) - \mathbb{E}\left[F(X_1, \dots, X_n)\right]| \geq \lambda K] \leq C \exp(-c\lambda^2/L^2) \tag{A.13}$$

For some constants $c, C > 0$.

Proof. The proof is straightforward. We apply Talagrand's Inequality to the function $G(X_1, \dots, X_n) = \frac{1}{L} F(X_1, \dots, X_n)$ and rearrange the terms. The constant λ scales to λ/L hence the L^2 in the denominator. □

Example: Application of Talagrand's Inequality. An example application for the inequality can be that X_i are i.i.d. Bernoulli with probability p and $F(X_1, \dots X_n) = \sum_{i=1}^{n} X_i^2$, where $|X_i| \leq 1$. We know that $Y_i = X_i^2$ follow the same distribution — Bernoulli with parameter p — and the function F is $2\sqrt{n}$-Lipschitz with respect to the Euclidean Norm in $[0, 1]^n$. Moreover, $\mathbb{E}\left[F(X_1, \dots, X_n)\right] = np$. Consequently, for every $\lambda > 0$ we have that

$$\Pr\left[\left|\sum_{i=1}^{n} X_i^2 - np\right| \geq \lambda\right] \leq C \exp\left(-\frac{c\lambda^2}{4n}\right) \tag{A.14}$$

Of course, Talagrands's Inequality can be used to derive more rigorous bounds. More precisely, in this work, we prove that the distance between two Bernoulli vectors with independent components is near to the distance of their parameter vectors with high probability. This will be later used to show that the k nearest neighbors of a Bernoulli Vector can be approximated by the k nearest parameter vectors in the parameter space for sufficiently small k and sufficiently large number of points.

A.3 McDiarmid's Inequality

For the analysis of some theorems in this thesis, we reside in McDiarmid's Inequality. More, specifically, McDiarmid's Inequality states the following

Theorem 15 (McDiarmid's Inequality)**.** *Let $X_1, \ldots, X_n$ be independed random variables where $X_i \in \mathbb{X}_i$. Let also $f : \mathbb{X}_1 \times \mathbb{X}_2 \cdots \times \mathbb{X}_n \to \mathbb{R}$ be a function of the random variables such that for all $1 \leq i \leq n$*

$$\sup_{X_1,\ldots,X_i,X_i',\ldots,X_n} |f(X_1,\ldots,X_i,\ldots,X_n) - f(X_1,\ldots,X_i',\ldots,X_n)| \leq c_i$$

for some constant $c_i \geq 0$. Then the following hold for every $\varepsilon > 0$

$$\Pr[f(X_1,\ldots,X_n) - \mathbb{E}\left[f(X_1,\ldots,X_n\right] \geq \varepsilon] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{A.15}$$

$$\Pr[f(X_1,\ldots,X_n) - \mathbb{E}\left[f(X_1,\ldots,X_n\right] \leq -\varepsilon] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{A.16}$$

$$\Pr[|f(X_1,\ldots,X_n) - \mathbb{E}\left[f(X_1,\ldots,X_n\right]| \geq \varepsilon] \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{A.17}$$

Example: Global Clustering Coefficient of $\mathcal{G}(n,p)$. We give an example by proving that the Global Clustering Coefficient of $\mathcal{G}(n,p)$ is p a.a.s. using McDriamid's Inequality. The Global Clustering Coefficient is the ratio of the number of triangles to the number of triplets (open and closed) in a network. Suppose $G \sim \mathcal{G}(n,p)$, then the expected number of closed triangles is $3\binom{n}{3}p^3$ and the number of connected triplets is $3\binom{n}{3}p^2$. It is clear that the ratio of the two expectations is p. Now we are going to prove that indeed the Global Clustering Coefficient is p for large enough n. Suppose that we take an edge of the graph and change it. Then the number of triangles can change by at most $c_i = 3n$. Also $\sum_{i=1}^n (3n)^2 = 9n^3$. Plugging everything to McDriamid's Inequality we get that the number of closed triangles T_C satisfy

$$\Pr[|T_C - \mathbb{E}\left[T_C\right]| \geq \varepsilon] \leq 2\exp\left(-\frac{4n}{9}\right) \tag{A.18}$$

And the number of triplets T_R

$$\Pr[|T_R - \mathbb{E}\left[T_R\right]| \geq \varepsilon] \leq 2\exp\left(-\frac{4n}{9}\right) \tag{A.19}$$

Therefore $T_C/T_R = p + O(n^{-2.5})$ with probability of at most $1 - 4\exp\left(-\frac{4n}{9}\right)$. Therefore $T_C/T_R \to p$ with probability 1 as $n \to \infty$.

A.4 Nomenclature for the Asymtotic Behaviour of Randomized Algorithms

Frequently, when someone analyzes algorithms and is interested in the asymptotic behaviour of some random process, one needs nomenclature for explaining the behaviour — mainly the asymptotic one — of such processes. For instance, when estimating the parameter p of a Bernoulli Random Variable with the sample mean $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$, we

say that $\hat{p} \to p$ asymptotically almost surely (a.a.s.) or $\hat{p}$ approaches p under the "probability limit" when $n \to \infty$. In this section, we give such definitions which are of little mathematical importance, however they serve as a robust communication basis for the analysis of randomized algorithms between computer scientists.

Definition 10 (Probability Limit)**.** *We say that* $\text{plim}_{n\to\infty} X_n = X$*, or equivalently that the sequence* $\{X_n\}$ *approaches* X *under the probability limit iff*

$$\lim_{n\to\infty} \Pr[|X_n - X| > \varepsilon] = 0 \tag{A.20}$$

for every $\varepsilon > 0$*. Equivalently,* $\text{plim}_{n\to\infty} X_n = X$*.*

Definition 11 (With high Probability)**.** *We say that an event* $A(n)$ *happens with high probability (w.h.p.) iff* $\Pr[A(n)] \geq 1 - O(1/n)$*.*

Definition 12 (Asymptotically Almost Surely)**.** *In asymptotic analysis, a property is said to hold asymptotically almost surely (a.a.s.), if over a sequence of sets, the probability converges to 1.*

A desideratum for randomized algorithms[2] is that the error probability goes to zero for a large value of some parameter, usually the sample size n. That means that if E is the erroneous event then $\lim_{n\to\infty} \Pr[E] = 0$. A usual practice is that for some finite n we bound $\Pr[E]$ by $O(1/n)$ and then let $n \to \infty$. Such requirement will soon be evident when analyzing our contributed algorithm in the forthcoming sections.

Example: The Random Graph $G(n, p_n)$**.** In random graph theory [14], the statement $G(n, p_n)$ is connected happens a.a.s. when for some $\varepsilon > 0$ the probability of connection satisfies $p_n > \frac{(1+\varepsilon)\log n}{n}$.

[2]We refer to *Monte Carlo* algorithms with the term randomized algorithms. In Monte Carlo algorithms the output of the algorithm is stochastic and the running time of the algorithm is deterministic. The other large category of randomized algorithms are the *Las Vegas* algorithms where the output is deterministic and the runtime is stochastic. For a more detailed introduction, we redirect the interested reader to the classical textbook of Mitzenmacher and Upfal [77].

List of Figures

2.1 Power law degree distributions in software projects. Log-log plots of the Linux Kernel 20.3M-long codebase. 36

2.2 Barabási-Albert Model for $T = 3$ iterations. Source: Wikipedia. 38

2.3 Behaviour of the Watts-Strogatz model on a graph with $n = 12$ vertices. In the leftmost figure the value of β is 0 indicating complete order. In the middle figure the value of β has increased near $1/2$ and small-world phenomena start to appear. Finally, in the rightmost figure consists of the case of $\beta = 1$ where the $k/2$ rightmost edges do rewire uniformly at random, each with probability $\frac{1}{n-1}$. 39

2.4 Core-periphery structure in an airline network. Source: [48]. 40

3.1 Expectation-Maximization for a Mixture of $k = 2$ Gaussians. 49

5.1 Lyapunov Function $V(x_1, x_2)$ of the example system. The trajectory $\gamma = \{(x_1, x_2) \mid \dot{x}_1 = x_2 - x_1^3 - x_1 x_2^2,\ \dot{x}_2 = -x_1 - x_1^2 x_2 - x_2^3,\ x_{10} = x_{20} = 0.4,\ t \geq 0\}$ is also provided as a scatter plot. Since $\dot{V} < 0$ for all $(x_1, x_2)^T \neq (0, 0)^T$ the system moves towards the base of the paraboloid $V = x_1^2 + x_2^2$ 64

5.2 Worst-Case Convergence Behaviour of a Markov Chain with a k-regular transition Matrix for various values of $k \geq 3$. 66

6.1 Nearest neighbor classifier example with 3 classes. 71

7.1 Log-log plot of the total time taken to perform inference to a network of up to 1M agents and $d \in \{10, 100\}$ with binary equiprobable artificial features, $D = 0.001$ and $k = \lceil \log n \rceil$. using LSH to obtain the nearest neighbors. . . . 94

7.2 Left: Engagement Threshold Effect. Right: Coverage curve for the BGMC policy for $\tau = 4$. 96

7.3 Microscopic properties of the NNIM model for $n = 100$ agents, $D = 10^{-3}$, and $k = 3$ neighbors. 99

7.4 Macroscopic Properties of the NNIM model. We have run the with $D = 10^{-7}$. 99

7.5 Effect of parameters and negative correlation between k and p on the ego-facebook dataset. The values of p range from 0.4 to 0.7 with 0.1 step and the values of k range from 1 to 200 with step size of 1. 101

List of Tables

7.1 Dataset Statistics and Homophilic Index are reported. We count directed edges where the network is undirected. The Homophilic Index is calculated after dimensionality reduction with PCA so that 95% of the original variance is explained after the transformation. 79

7.2 Complexity of nnim with under various data structures (Brute-force, KD-tree, Metric Ball, LSH) for running the nnim model such that the total variation distance is at most $d \cdot D$ after execution. State-of-the-art is DCI and Prioritized DCI [63, 64]. The quantity d' is the intrinsic dimension [64, p.1]. The number m is the number of projection directions used in the DCI. 95

7.3 Experimental results with $p = 0.7$, $\gamma = 2$, $D = 10^{-3}$, $\tau = 4$, and regularization with $\alpha = 1$.
. 104

Bibliography

[1] Six degrees of wikipedia. https://www.sixdegreesofwikipedia.com/. Accessed: 2019-08-01.

[2] Rediet Abebe, Jon Kleinberg, David Parkes, and Charalampos E Tsourakakis. Opinion dynamics with varying susceptibility to persuasion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1089–1098, 2018.

[3] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

[4] Laurence Aitchison, Nicola Corradi, and Peter E Latham. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS computational biology*, 12(12):e1005110, 2016.

[5] Phipps Arabie and Scott A Boorman. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10(2):148–203, 1973.

[6] Chen Avin, Michael Borokhovich, Zvi Lotker, and David Peleg. Distributed computing on core–periphery networks: Axiom-based design. *Journal of Parallel and Distributed Computing*, 99:51–67, 2017.

[7] EA Barbashin and Nikolai Nikolaevich Krasovskii. On stability of motion in the large. Technical report, TRW SPACE TECHNOLOGY LABS LOS ANGELES CALIF, 1961.

[8] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[9] Austin Benson and Jon Kleinberg. Link prediction in networks with core-fringe data. In *The World Wide Web Conference*, pages 94–104, 2019.

[10] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[11] Kshipra Bhawalkar, Sreenivas Gollapudi, and Kamesh Munagala. Coevolutionary opinion formation games. In *Proc. of the 45th ACM Symposium on Theory of Computing Conference (STOC 2013)*, pages 41–50. ACM, 2013.

[12] David Bindel, Jon Kleinberg, and Sigal Oren. How bad is forming your own opinion? *Games and Economic Behavior*, 92:248–265, 2015.

[13] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[14] Béla Bollobás. *Random graphs*. Number 73. Cambridge university press, 2001.

[15] Anthony Bonato, David F Gleich, Myunghwan Kim, Dieter Mitsche, Paweł Prałat, Yanhua Tian, and Stephen J Young. Dimensionality of social networks using motifs and eigenvalues. *PloS one*, 9(9):e106052, 2014.

[16] Anthony Bonato, Jeannette Janssen, and Pawel Prałat. The geometric protean model for on-line social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 110–121. Springer, 2010.

[17] Anthony Bonato, Jeannette Janssen, and Paweł Prałat. Geometric protean graphs. *Internet Mathematics*, 8(1-2):2–28, 2012.

[18] Anthony Bonato, Marc Lozier, Dieter Mitsche, Xavier Pérez-Giménez, and Paweł Prałat. The domination number of on-line social networks and random geometric graphs. In *International Conference on Theory and Applications of Models of Computation*, pages 150–163. Springer, 2015.

[19] Timothy Caulfield and Declan Fahy. Science, celebrities, and public engagement. *Issues in Science and Technology*, 32(4):24, 2016.

[20] PHILIP SM CHIN. Generalized integral method to derive lyapunov functions for nonlinear systems. *International Journal of Control*, 46(3):933–943, 1987.

[21] Fan Chung and Linyuan Lu. The diameter of sparse random graphs. *Advances in Applied Mathematics*, 26(4):257–279, 2001.

[22] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. *arXiv preprint arXiv:1609.00368*, 2016.

[23] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. A converse to banach's fixed point theorem and its cls-completeness. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 44–50, 2018.

[24] Judith IM De Groot and Linda Steg. Morality and prosocial behavior: The role of awareness, responsibility, and norms in the norm activation model. *The Journal of social psychology*, 149(4):425–449, 2009.

[25] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[26] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[27] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. Cohesive co-evolution patterns in dynamic attributed graphs. In *International Conference on Discovery Science*, pages 110–124. Springer, 2012.

[28] Lee DeVille. Optimizing gershgorin for symmetric matrices. *Linear Algebra and its Applications*, 577:360–383, 2019.

[29] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. Learning structural node embeddings via diffusion wavelets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1320–1329, 2018.

[30] Joseph L Doob. Regularity properties of certain families of chance variables. *Transactions of the American Mathematical Society*, 47(3):455–486, 1940.

[31] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[32] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[33] Alessandro Epasto and Bryan Perozzi. Is a single embedding enough? learning node representations that capture multiple social contexts. In *The World Wide Web Conference*, pages 394–404. ACM, 2019.

[34] Uriel Feige. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[35] Dimitris Fotakis, Vardis Kandiros, Vasilis Kontonis, and Stratis Skoulakis. Opinion dynamics with limited information. In *International Conference on Web and Internet Economics*, pages 282–296. Springer, 2018.

[36] Dimitris Fotakis, Dimitris Palyvos-Giannas, and Stratis Skoulakis. Opinion dynamics with local interactions. In *IJCAI*, pages 279–285, 2016.

[37] Karen Freberg, Kristin Graham, Karen McGaughey, and Laura A Freberg. Who are the social media influencers? a study of public perceptions of personality. *Public Relations Review*, 37(1):90–92, 2011.

[38] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

[39] Joel Friedman. *A proof of Alon's second eigenvalue conjecture and related problems*. American Mathematical Soc., 2008.

[40] Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. Bine: Bipartite network embedding. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 715–724. ACM, 2018.

[41] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999.

[42] Gail Gong and Francisco J Samaniego. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, pages 861–869, 1981.

[43] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

[44] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[45] Teresa W Haynes, Stephen Hedetniemi, and Peter Slater. *Fundamentals of domination in graphs*. CRC press, 2013.

[46] Rainer Hegselmann, Ulrich Krause, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*, 5(3), 2002.

[47] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[48] Junteng Jia and Austin R Benson. Random spatial network models for core-periphery structure. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 366–374. ACM, 2019.

[49] Seunga Venus Jin. "celebrity 2.0 and beyond!" effects of facebook profile sources on social networking advertising. *Computers in Human Behavior*, 79:154–168, 2018.

[50] Leo P Kadanoff. More is the same; phase transitions and mean field theories. *Journal of Statistical Physics*, 137(5-6):777, 2009.

[51] Vardis Kandiros. Opinion dynamics with limited information. *Barchelor's Thesis*, 2018.

[52] Susie Khamis, Lawrence Ang, and Raymond Welling. Self-branding,'micro-celebrity'and the rise of social media influencers. *Celebrity studies*, 8(2):191–208, 2017.

[53] Myunghwan Kim. *Modeling Networks with Auxiliary Information.* PhD thesis, Stanford University, 2014.

[54] Myunghwan Kim and Jure Leskovec. Modeling social networks with node attributes using the multiplicative attribute graph model. *arXiv preprint arXiv:1106.5053*, 2011.

[55] Myunghwan Kim and Jure Leskovec. Multiplicative attribute graph model of real-world networks. *Internet mathematics*, 8(1-2):113–160, 2012.

[56] Jon M Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[57] David Krackhardt. A plunge into networks, 2009.

[58] Paul Krugman. Increasing returns and economic geography. *Journal of political economy*, 99(3):483–499, 1991.

[59] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

[60] Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection, 2014.

[61] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[62] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[63] Ke Li and Jitendra Malik. Fast k-nearest neighbour search via dynamic continuous indexing. In *International Conference on Machine Learning*, pages 671–679, 2016.

[64] Ke Li and Jitendra Malik. Fast k-nearest neighbour search via prioritized dci. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2081–2090. JMLR. org, 2017.

[65] Torgny Lindvall et al. On strassen's theorem on stochastic domination. *Electronic communications in probability*, 4:51–59, 1999.

[66] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 3, 2014.

[67] Panagiotis Louridas, Diomidis Spinellis, and Vasileios Vlachos. Power laws in software. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 18(1):1–26, 2008.

[68] Aleksandr Mikhailovich Lyapunov. The general problem of the stability of motion. *International journal of control*, 55(3):531–534, 1892.

[69] Alan MacCormack. The architecture of complex systems: Do" core-periphery" structures dominate? In *Academy of Management Proceedings*, volume 2010, pages 1–6. Academy of Management Briarcliff Manor, NY 10510, 2010.

[70] Tondani Makhuvha, Geoffrey Pegram, Ross Sparks, and Walter Zucchini. Patching rainfall data using regression methods.: 1. best subset selection, em and pseudo-em methods: theory. *Journal of Hydrology*, 198(1-4):289–307, 1997.

[71] Fragkiskos D Malliaros, Maria-Evgenia G Rossi, and Michalis Vazirgiannis. Locating influential nodes in complex networks. *Scientific reports*, 6:19307, 2016.

[72] J Miller McPherson and James R Ranger-Moore. Evolution on a dancing landscape: organizations and networks in dynamic blau space. *Social Forces*, 70(1):19–42, 1991.

[73] Miller McPherson. An ecology of affiliation. *American Sociological Review*, pages 519–532, 1983.

[74] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[75] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[76] Kurt Miller, Michael I Jordan, and Thomas L Griffiths. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, pages 1276–1284, 2009.

[77] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[78] F Molnár, Sameet Sreenivasan, Boleslaw K Szymanski, and Gyorgy Korniss. Minimum dominating sets in scale-free network ensembles. *Scientific reports*, 3:1736, 2013.

[79] F Molnár Jr, Noemi Derzsy, Éva Czabarka, L Székely, Boleslaw K Szymanski, and Gyorgy Korniss. Dominating scale-free networks using generalized probabilistic methods. *Scientific reports*, 4:6308, 2014.

[80] Jose C Nacher and Tatsuya Akutsu. Dominating scale-free networks with variable scaling exponent: heterogeneous networks are not difficult to control. *New Journal of Physics*, 14(7):073005, 2012.

[81] Angelia Nedić and Behrouz Touri. Multi-dimensional hegselmann-krause dynamics. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 68–73. IEEE, 2012.

[82] Roger J Nemeth and David A Smith. International trade and world-system structure: A multiple network analysis. *Review (Fernand Braudel Center)*, 8(4):517–560, 1985.

[83] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

[84] Christine Leigh Myers Nickel. *Random dot product graphs a model for social networks*. PhD thesis, Johns Hopkins University, 2008.

[85] Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

[86] Marios Papachristou. Software clusterings with vector semantics and the call graph. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 1184–1186, 2019.

[87] Marios Papachristou and Dimitris Fotakis. Stochastic opinion dynamics for interest prediction in social networks. In *Under Review*, 2020.

[88] Vilfredo Pareto. *The mind and society*, volume 1. Harcourt, Brace and Howe, 1935.

[89] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[90] Adriana Prado, Marc Plantevit, Céline Robardet, and Jean-Francois Boulicaut. Mining graph topological patterns: Finding covariations among vertex descriptors. *IEEE Transactions on Knowledge and Data Engineering*, 25(9):2090–2104, 2012.

[91] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[92] Tim Roughgarden and Florian Schoppmann. Local smoothness and the price of anarchy in splittable congestion games. *Journal of Economic Theory*, 156:317–342, 2015.

[93] Tim Roughgarden and Gregory Valiant. Cs168: The modern algorithmic toolbox lecture# 4: Dimensionality reduction. 2015.

[94] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding, 2019.

[95] David J Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Physical review letters*, 113(6):068102, 2014.

[96] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[97] Mateen Shaikh, Paul D McNicholas, and Anthony F Desmond. A pseudo-em algorithm for clustering incomplete longitudinal data. *The international journal of biostatistics*, 6(1), 2010.

[98] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[99] David Snyder and Edward L Kick. Structural position in the world system and economic growth, 1955-1970: A multiple-network analysis of transnational interactions. *American journal of Sociology*, 84(5):1096–1126, 1979.

[100] Frederic Stahl, Mohamed Medhat Gaber, and Mariam Adedoyin-Olowe. A survey of data mining techniques for social media analysis. *Journal of Data Mining & Digital Humanities*, 2014, 2014.

[101] HE Stanley. Mean field theory of magnetic phase transitions. *Introduction to Phase Transitions and Critical Phenomena*, 1971.

[102] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[103] Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

[104] Michel Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

[105] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[106] Hermann Thorisson. *Regeneration*. Springer, 2000.

[107] Behrouz Touri. Averaging dynamics in general state spaces. In *Product of Random Stochastic Matrices and Distributed Averaging*, pages 113–126. Springer, 2012.

[108] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. In *Social Networks*, pages 179–197. Elsevier, 1977.

[109] Immanuel Wallerstein. World-systems analysis, 1987.

[110] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[111] Dingqi Yang, Paolo Rosso, Bin Li, and Philippe Cudre-Mauroux. Nodesketch: Highly-efficient graph embeddings via recursive sketching. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1162–1172, 2019.

[112] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596, 2013.

[113] Stephen J Young and Edward Scheinerman. Directed random dot product graphs. *Internet Mathematics*, 5(1-2):91–111, 2008.

[114] Xiao Zhang, Travis Martin, and Mark EJ Newman. Identification of core-periphery structure in networks. *Physical Review E*, 91(3):032803, 2015.

[115] Xiao Zhang, Travis Martin, and Mark EJ Newman. Identification of core-periphery structure in networks. *Physical Review E*, 91(3):032803, 2015.