Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Ανίχνευση Κοινοτήτων σε Κοινωνικά Δίκτυα με Εφαρμογή σε Συστήματα Συστάσεων Συνεργατικής Διήθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΩΑΝΝΗΣ Β. ΚΑΖΑΚΟΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2017

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Ανίχνευση Κοινοτήτων σε Κοινωνικά Δίκτυα με Εφαρμογή σε Συστήματα Συστάσεων Συνεργατικής Διήθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΩΑΝΝΗΣ Β. ΚΑΖΑΚΟΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου 2017.

.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Παναγιώτης Τσανάκας Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2017

Ιωάννης Β. Καζάκος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της παρούσας διπλωματικής είναι η ενσωμάτωση της ανάλυσης κοινωνικών δικτύων και συγκεκριμένα της ανίχνευσης κοινοτήτων σε συστήματα συστάσεων συνεργατικής διήθησης προκειμένου να βελτιωθεί η ποιότητα των συστάσεών τους. Τα τελευταία χρόνια, τα συστήματα συστάσεων γίνονται ολοένα και πιο δημοφιλή στο διαδίκτυο με εφαρμογές σε διάφορες περιοχές όπως ταινίες, μουσική, ειδήσεις, βιβλία, επιστημονικά άρθρα και καταναλωτικά προϊόντα. Όταν οι χρήστες για τους οποίους γίνονται οι συστάσεις συνδέονται μεταξύ τους με κάποιο δεσμό, όπως σε ένα κοινωνικό δίκτυο, αυτή η πληροφορία μπορεί να χρησιμοποιηθεί ως μια επιπλέον παράμετρος του συστήματος ώστε να επιτευχθούν καλύτερες συστάσεις για τους χρήστες. Η παρούσα εργασία αποτελείται από δύο βασικά μέρη.

Το πρώτο από αυτά αφορά την ανίχνευση κοινοτήτων στο δίκτυο των χρηστών του υπό εξέταση συνόλου δεδομένων. Αρχικά, εντοπίζονται οι συνεκτικές συνιστώσες του δικτύου και εξαιρούνται από τη διαδικασία οι μικρότερες καθώς μπορούν να θεωρηθούν ως υφιστάμενες κοινότητες. Στη συνέχεια, εφαρμόζεται ο αλγόριθμος ανίχνευσης κοινοτήτων των Girvan-Newman στις μεγάλες συνεκτικές συνιστώσες του δικτύου. Ο αλγόριθμος αυτός εντοπίζει τις κοινότητες ενός δικτύου αφαιρώντας προοδευτικά ακμές οι οποίες συναντώνται συχνότερα σε διαδρομές συντομότερων μονοπατιών μεταξύ των κόμβων του δικτύου. Τα συνδεδεμένα μέρη του εναπομείναντος δικτύου αποτελούν τις τελικές κοινότητες.

Το δεύτερο μέρος περιλαμβάνει την εφαρμογή των αποτελεσμάτων της ανίχνευσης κοινοτήτων στα κλασικά συστήματα συστάσεων συνεργατικής διήθησης. Αυτό επιτυγχάνεται με τον περιορισμό της γειτονιάς του κάθε χρήστη-στόχου στους χρήστες οι οποίοι ανήκουν στην ίδια κοινότητα με αυτόν, αντί να λαμβάνεται υπόψη όλο το δίκτυο. Πιο συγκεκριμένα, σε πρώτη φάση αναπτύσσονται συστήματα συνεργατικής διήθησης βασισμένης στο χρήστη και στο αντικείμενο καθώς και απλά συστήματα κοινωνικής σύστασης προκειμένου να εντοπιστεί ο πιο αποδοτικός συνδυασμός συστημάτων και παραμέτρων. Κατόπιν, για τις βέλτιστες παραμέτρους αναπτύσσονται αντίστοιχα συστήματα στα οποία λαμβάνονται υπόψη οι κοινότητες του δικτύου που εντοπίστηκαν προηγουμένως κατά το βήμα εύρεσης της γειτονιάς του χρήστη-στόχου. Πρόκειται για το προτεινόμενο από την παρούσα εργασία κοινωνικό σύστημα συνεργατικής διήθησης.

Τέλος, γίνεται σύγκριση της ποιότητας των συστάσεων του προτεινόμενου συστήματος με τα υπόλοιπα συστήματα και αξιολόγηση των αποτελεσμάτων σε σχέση με τις προκλήσεις του προβλήματος ενώ δίνονται και μελλοντικές κατευθύνσεις έρευνας.

Λέξεις κλειδιά

Ανάλυση Κοινωνικών δικτύων, Ανίχνευση κοινοτήτων, Αλγόριθμος Girvan-Newman, Συστήματα συστάσεων, Συνεργατική διήθηση, Γειτονιά χρήστη.

Abstract

The purpose of this diploma thesis is the integration of social network analysis and specifically community detection in collaborative filtering recommender systems. In recent years, recommender systems are becoming increasingly popular on the web including various applications such as movies, music, news, books, scientific articles and consumer goods. In cases where the target users of the recommendations are connected with each other in some way, like in a social network, such information may be used as an extra parameter of the recommender system in order to improve the quality of its recommendations. The present work consists of two basic parts.

The first of them concerns the community detection in the network of users of the dataset under examination. As a first step, the network's connected components are located and the smaller of them are excluded from the procedure, being considered as existing communities. Subsequently, the Girvan-Newman community detection algorithm is applied to the bigger connected components of the graph. This algorithm locates the communities of a network by progressively removing edges which are encountered more frequently in shortest paths between pairs of nodes in the network.

The second part includes the application of the community detection results in classic collaborative filtering recommneder systems. This is achieved by restricting the neighborhood of each target user to the users who belong to the same community with him, instead of considering the whole network. Initially, classic user-based and item-based collaborative filtering recommender systems are developed, as well as simple social recommender systems, in order to find the optimal set of parameters. Having found the optimal configurations, another user-based collaborative filtering is developed, in which the communities located previously are taken into account at the step of computing a target user's neighborhood. This is the suggested social collaborative filtering recommender system of the present diploma thesis.

Finally, the suggested recommender system is compared with the rest of the systems and the results are evaluated, in reference with the challenges of the problem, while future research directions are also given.

Key words

Social network analysis, Community detection, Girvan-Newman algorithm, Recommender systems, Collaborative filtering, User neighborhood.

Ευχαριστίες

Η παρούσα διπλωματική εργασία, με την οποία ολοκληρώνεται η ακαδημαϊκή μου πορεία στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, είναι αποτέλεσμα της συμπόρευσης και συνεργασίας με διάφορους ανθρώπους οι οποίοι με βοήθησαν καθ' όλη τη διάρκεια αυτής.

Κατ' αρχάς, θα ήθελα να ευχαριστήσω τον κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π., ο οποίος μου έδωσε την ευκαιρία να εκπονήσω την παρούσα διπλωματική εργασία και να ασχοληθώ με ένα θέμα το οποίο με ενδιέφερε πολύ. Κατόπιν, θα ήθελα να ευχαριστήσω τους κ. Παναγιώτη Τσανάκα, Καθηγητή Ε.Μ.Π. και Γεώργιο Στάμου, Αναπληρωτή Καθηγητή Ε.Μ.Π., για την τιμή που μου έκαναν να είναι μέλη της επιτροπής εξέτασης της διπλωματικής μου εργασίας.

Ιδιαίτερα θερμές ευχαριστίες θα ήθελα να αποδώσω στον κ. Γεώργιο Αλεξανδρίδη, Διδάκτορα Ε.Μ.Π., ο οποίος μου παρείχε σημαντική βοήθεια σε όλα τα στάδια της διπλωματικής μου εργασίας. Η συνεργασία μας ήταν άκρως εποικοδομητική και η βοήθειά του καθοριστική, γι' αυτό και τον ευχαριστώ θερμά. Επίσης θα ήθελα να ευχαριστήσω και την κα. Ελένη Βάθη, Υποψήφια Διδάκτορα Ε.Μ.Π. για τη συνεισφορά της στην εκπόνηση αυτής της εργασίας.

Τέλος, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένεια μου, η οποία μου έδωσε τη δυνατότητα να φοιτήσω στη σχολή που διάλεξα και στάθηκε δίπλα μου σε όλες τις δυσκολίες, καθώς και στους φίλους μου, κάποιοι από τους οποίους και συμφοιτητές μου, οι οποίοι ήταν πάντα δίπλα μου σε όλη αυτή την πορεία.

Ιωάννης Β. Καζάκος,

Αθήνα, 18η Ιουλίου 2017

Η εργασία αυτή είναι επίσης διαθέσιμη ως Τεχνική Αναφορά , Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Τεχνολογίας Λογισμικού, Ιούλιος 2017.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Περιεχόμενα

Περίληψη 5

Abstract 7

Ευχαριστίες 9

Περιεχόμενα 11

Κατάλογος πινάκων 13

Κατάλογος σχημάτων 15

1. Εισαγωγή 17
1.1 Συστήματα Συστάσεων 17
1.2 Κοινωνικά Δίκτυα 20
1.3 Κυριότερες Προκλήσεις 21
1.4 Στόχοι-Δομή της Εργασίας 23

2. Συστήματα Συστάσεων 25
2.1 Συστήματα αναφοράς 25
2.1.1 Μέσος όρος χρήστη 25
2.1.2 Μέσος όρος αντικειμένου 26
2.2 Συνεργατική Διήθηση 26
2.2.1 Μέθοδοι βασισμένες στη μνήμη 26
2.2.2 Σύγκριση των βασισμένων στο χρήστη και βασισμένων στο αντικείμενο μεθόδων 31
2.2.3 Μοντέλα γράφων για συνεργατική διήθηση βασισμένη στη γειτονιά 32
2.2.4 Συνεργατική διήθηση βασισμένη σε μοντέλα 35
2.2.5 Πλεονεκτήματα και μειονεκτήματα των μεθόδων βασισμένων στη μνήμη και βασισμένων σε μοντέλα 40

3. Κοινωνικά Δίκτυα 43
3.1 Ανάλυση κοινωνικών δικτύων 43
3.1.1 Βασικές έννοιες θεωρίας γράφων 43
3.1.2 Μετρικές ανάλυσης κοινωνικών δικτύων 46
3.1.3 Εφαρμογές της ανάλυσης κοινωνικών δικτύων 47
3.2 Ανίχνευση κοινοτήτων 48
3.2.1 Στοιχεία Ανίχνευσης κοινοτήτων 48
3.2.2 Κλασσικές μέθοδοι ανίχνευσης κοινοτήτων 51
3.2.3 Ιεραρχική Συσταδοποίηση 53
3.2.4 Φασματική Συσταδοποίηση 54
3.2.5 Διαιρετικοί αλγόριθμοι 54
3.2.6 Αλγόριθμοι βασισμένοι στην τμηματικότητα 57

3.3 Τα κοινωνικά δίκτυα στα συστήματα συστάσεων 58
3.3.1 Ορισμοί της κοινωνικής σύστασης 58
3.3.2 Βασική ιδέα της κοινωνικής σύστασης 58
3.3.3 Υφιστάμενα συστήματα κοινωνικής σύστασης 59

4. Πειραματική Διαδικασία και Αποτελέσματα 65
4.1 Η συλλογή δεδομένων 65
4.2 Πειραματική Διαδικασία 66
4.2.1 Το δίκτυο των χρηστών 66
4.2.2 Η διαδικασία ανίχνευσης κοινοτήτων 67
4.2.3 Η διαδικασία της αξιολόγησης 70
4.3 Μετρικές 71
4.3.1 Ακρίβεια και Ανάκληση 72
4.3.2 Κάλυψη 74
4.4 Αποτελέσματα 75
4.4.1 Συστήματα αναφοράς 75
4.4.2 Συνεργατική διήθηση βασισμένη στο χρήστη (User-based CF) 76
4.4.3 Συνεργατική διήθηση βασισμένη στο αντικείμενο (Item-based CF) 79
4.4.4 Συστήματα κοινωνικής σύστασης που βασίζονται στην ανάλυση κοινωνικών δικτύων 79
4.4.5 Το προτεινόμενο σύστημα κοινωνικής σύστασης 81
4.4.6 Σύγκριση των συστημάτων που υλοποιήθηκαν 83

5. Συμπεράσματα και μελλοντικές κατευθύνσεις 87
5.1 Συμπεράσματα 87
5.2 Μελλοντικές Κατευθύνσεις 88

Βιβλιογραφία 91

Κατάλογος πινάκων

1.1 Παράδειγμα πίνακα βαθμολογιών αποτελούμενο από 4 χρήστες και 4 αντικείμενα . . 19

4.1 Η συλλογή δεδομένων `MovieTweetings` 65
4.2 Συνεκτικές συνιστώσες κοινωνικού δικτύου του `MovieTweetings` 68
4.3 Υλοποιούμενα συστήματα συστάσεων στην πειραματική διαδικασία 75
4.4 Παράμετροι συστημάτων συνεργατικής διήθησης βασιμένης στο χρήστη 76

Κατάλογος σχημάτων

1.1 Παράδειγμα συστήματος σύστασης ταινιών στον ιστότοπο IMDb 18
1.2 Λίστα με τα δημοφιλέστερα κοινωνικά δίκτυα με βάση τον αριθμό ενεργών χρηστών τους (πηγή https://www.slideshare.net/wearesocialsg/global-digital-statshot-q2-2017) 21
1.3 Το φαινόμενο long-tail 22

2.1 Πίνακας βαθμολογιών και ο αντίστοιχος γράφος χρήστη-αντικειμένου 34
2.2 Σύνοψη διαφορών μεταξύ συνεργατικής διήθησης και ταξινόμησης 37
2.3 Πρόβλεψη των βαθμολογιών ενός χρήστη με τη βοήθεια του λανθάνοντος διανύσματος 38
2.4 Παράδειγμα παραγοντοποίησης πίνακα 2ης τάξης 39

3.1 Γράφημα και ο αντίστοιχος πίνακας γειτνίασης 44
3.2 Διάσχιση γράφου με αναζήτηση κατά πλάτος και αναζήτηση κατά βάθος 45
3.3 Οι δύο κοινότητες του Karate Club του Zachary [Zach77] 48
3.4 Διαμερισμός γράφου σε δύο ομάδες με ελάχιστο μέγεθος τομής 52
3.5 Παράδειγμα εφαρμογής του αλγορίθμου Girvan-Newman σε μια συνεκτική συνιστώσα του γράφου με 17 κόμβους 57
3.6 Απλό δίκτυο εμπιστοσύνης 62

4.1 Το φαινόμενο long-tail σε λογαριθμικούς άξονες 66
4.2 Η κατανομή βαθμών κορυφών του γράφου των χρηστών 67
4.3 Η κατανομή των κόμβων στις κοινότητες που δημιουργήθηκαν από τον αλγόριθμο Girvan-Newman 69
4.4 Αναπαράσταση k-fold cross validation με k=4. Από Fabian Flöck - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=51562781 71
4.5 Σχηματική απεικόνιση των μετρικών της ακρίβειας και ανάκλησης. Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283 73
4.6 Μέση Αντιπροσωπευτική Ακρίβεια για τα Συστήματα Αναφοράς 76
4.7 Κάλυψη ως προς τις τιμές κατωφλίου για διάφορες συναρτήσεις ομοιότητας στα User-based CF συστήματα 77
4.8 MAP of Good @5 συναρτήσει του μεγέθους γειτονιάς για τις διαφορετικές συναρτήσεις ομοιότητας και κατώφλι ομοιότητας ίσο με 0.75 77
4.9 MAP of Good @5 συναρτήσει του κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας και για μέγεθος γειτονιάς ίσο με 3 78
4.10 Κάλυψη συναρτήσει του κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας 79
4.11 MAP of Good @ 5 συναρτήσει του κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας και μέγεθος γειτονιάς ίσο με 10 80
4.12 MAP (of Good) @ 5 συναρτήσει του κατωφλίου ομοιότητας για συντελεστές Jaccard και Adamic-Adar 80
4.13 Κάλυψη συναρτήσει του κατωφλίου ομοιότητας για συντελεστές Jaccard και Adamic-Adar 81
4.14 Κάλυψη συναρτήσει του κατωφλίου ομοιότητας για το προτεινόμενο σύστημα . . . 82
4.15 MAP of Good @ 2 συναρτήσει του κατωφλίου ομοιότητας για το προτεινόμενο σύστημα 82

4.16 Μέση Αντιπροσωπευτική Ακρίβεια (καλών αντικειμένων) συναρτήσει του μεγέθους λίστας για το προτεινόμενο σύστημα 83
4.17 Σύγκριση των συστημάτων που εξετάστηκαν ως προς την Κάλυψη και την Μέση Αντιπροσωπευτική Ακρίβεια 84

Κεφάλαιο 1

Εισαγωγή

1.1 Συστήματα Συστάσεων

Η ραγδαία ανάπτυξη του Παγκόσμιου Ιστού έχει ως αποτέλεσμα οι χρήστες να δέχονται καθημερινά τεράστια ποσότητα πληροφορίας την οποία πιθανώς να μη μπορούν να διαχειριστούν αποτελεσματικά. Επιπλέον, το διαδίκτυο κατέχει όλο και σημαντικότερο ρόλο στη διαφήμιση και σε ηλεκτρονικές συναλλαγές και αγορές. Αυτό έχει ως αποτέλεσμα τα *συστήματα συστάσεων (recommender systems)*, τα οποία αφενός έχουν τη δυνατότητα να φιλτράρουν πληροφορίες οι οποίες πιθανόν να ενδιαφέρουν τους χρήστες και αφετέρου να συμβάλλουν στη διαφήμιση και στην αύξηση των διαδικτυακών αγορών, να γίνονται ολοένα και πιο δημοφιλή.

Τέτοια συστήματα έχουν υλοποιηθεί ευρέως σε διάφορους τομείς όπως είναι η σύσταση καταναλωτικών προϊόντων στο Amazon[1] και η σύσταση ταινιών στο IMDb[2] και στο Netflix[3]. Ένα τέτοιο παράδειγμα φαίνεται στο Σχήμα 1.1, το οποίο προέρχεται από τον ιστότοπο IMDb. Στο παράδειγμα αυτό, όταν ο χρήστης επιλέγει μια ταινία προκειμένου να τη βαθμολογήσει ή να δει τα χαρακτηριστικά της, το αυτόματο σύστημα σύστασης ταινιών προτείνει στο χρήστη άλλες ταινίες οι οποίες έλαβαν και αυτές καλή βαθμολογία από χρήστες που είχαν βαθμολογήσει θετικά την ταινία που εξετάζει ο χρήστης.

Ένας πολύ σημαντικός παράγοντας στην εξάπλωση των συστημάτων συστάσεων είναι σίγουρα η ευκολία με την οποία το διαδίκτυο επιτρέπει στους χρήστες του να αξιολογούν και να σχολιάζουν το περιεχόμενό του. Για παράδειγμα, σε ένα διαδικτυακό πάροχο περιεχομένου (content provider) όπως είναι το Netflix, οι χρήστες έχουν τη δυνατότητα να αξιολογούν το περιεχόμενο με ένα κλικ του ποντικιού. Μια τυπική μέθοδος σχολιασμού του περιεχόμενου είναι με τη μορφή βαθμολογιών στην οποία οι χρήστες επιλέγουν αριθμητικές τιμές μέσω ενός συγκεκριμένου συστήματος αξιολόγησης, π.χ. μια κλίμακα αξιολόγησης πέντε αστέρων (five-star rating system), δηλώνοντας έτσι την αρέσκεια ή δυσαρέσκειά τους. Υπάρχουν όμως και άλλες μορφές ανατροφοδότησης οι οποίες δεν είναι τόσο σαφείς αλλά εξίσου εύκολο να καταγραφούν μέσω της διαδικτυακής δραστηριότητας του χρήστη. Για παράδειγμα η προβολή ή αγορά ενός προϊόντος σε ένα διαδικτυακό κατάστημα μπορεί να θεωρηθεί ως επιδοκιμασία για το συγκεκριμένο αντικείμενο.

Η βασική ιδέα των συστημάτων συστάσεων είναι η χρήση των διάφορων αυτών πηγών δεδομένων προκειμένου να συμπεράνουν τα ενδιαφέροντα των χρηστών. Η οντότητα στην οποία παρέχεται η σύσταση αναφέρεται ως *χρήστης (user)* και το προϊόν το οποίο προτείνεται ως *αντικείμενο (item)*. Στα προβλήματα σύστασης χρησιμοποιείται πολύ συχνά η έννοια του *πίνακα βαθμολογιών (rating matrix)* ο οποίος είναι ένας $m \times n$ πίνακας, όπου m ο αριθμός των χρηστών και n ο αριθμός των αντικειμένων. Το στοιχείο r_{ij} του πίνακα βαθμολογιών R αντιστοιχεί στη βαθμολογία που έχει δώσει ο χρήστης i στο αντικείμενο j. Για παράδειγμα, στον Πίνακα 1.1 βλέπουμε μια απλή περίπτωση πίνακα βαθμολογιών με 4 χρήστες και 4 αντικείμενα, στον οποίο ο χρήστης U_3 έχει βαθμολογήσει το αντικείμενο I_4 με την τιμή 8. Οι γνωστές τιμές του πίνακα βαθμολογιών αναφέρονται συνήθως ως *προσδιορισμένες (specified) ή παρατηρούμενες (observed)* βαθμολογίες ενώ οι άγνωστες ως *απροσδιόριστες (unspecified) ή απαρατήρητες (unobserved)*.

[1] `https://www.amazon.com`

[2] `https://www.imdb.com`

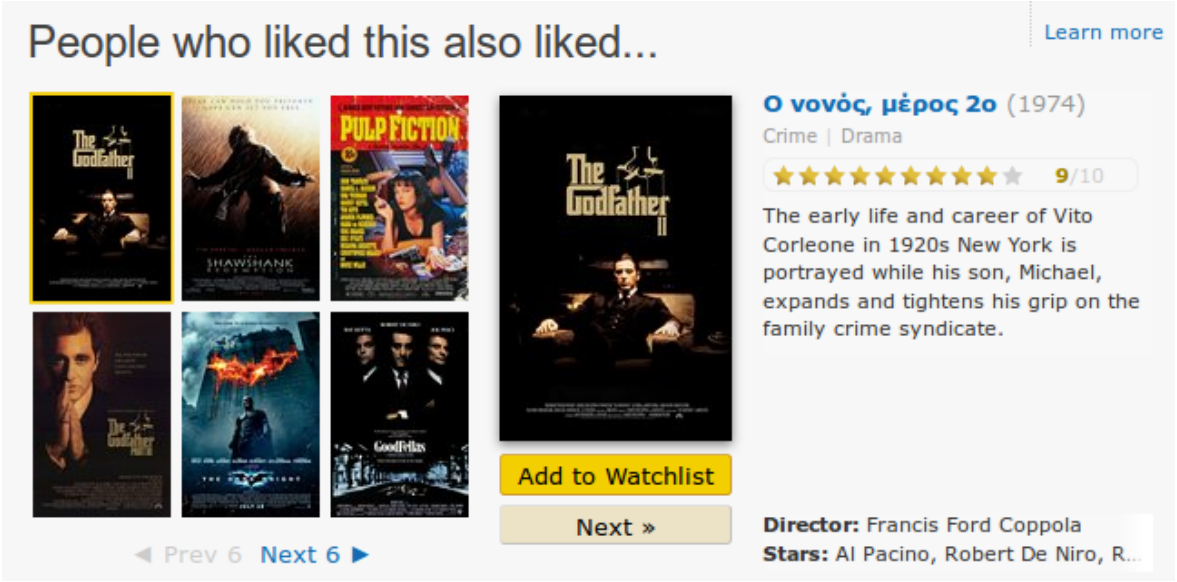
[3] `https://www.netflix.com`

Υπάρχουν διάφορες τεχνικές τις οποίες χρησιμοποιούν τα συστήματα συστάσεων, ωστόσο, η κύρια αρχή τους είναι ότι βασίζονται στις σημαντικές αλληλεπιδράσεις μεταξύ των χρηστών ή/και των αντικειμένων [Agga16]. Γενικότερα, τα συστήματα συστάσεων, ανάλογα με τον τύπο των δεδομένων που χρησιμοποιούν χωρίζονται σε συστήματα *συνεργατικής διήθησης (collaborative filterig)* και *βασισμένα στο περιεχόμενο (content-based)*. Τα μοντέλα συνεργατικής διήθησης χρησιμοποιούν τις συσχετίσεις που υπάρχουν συνήθως μεταξύ των προσδιορισμένων βαθμολογιών των χρηστών ή/και αντικειμένων προκειμένου να βρουν τις απροσδιόριστες βαθμολογίες. Για παράδειγμα, στην περίπτωση του Πίνακα 1.1, βλέποντας ότι οι χρήστες U_1 και U_4 έχουν δώσει παρόμοιες βαθμολογίες στα αντικείμενα I_1 και I_4 τα οποία έχουν βαθμολογήσει από κοινού, η βαθμολογία του χρήστη U_1 για το αντικείμενο I_3 η οποία είναι απροσδιόριστη θα μπορούσε να λάβει τιμή παραπλήσια με αυτή που έχει δώσει ο χρήστης U_4 στο συγκεκριμένο αντικείμενο, από τη στιγμή που οι δύο χρήστες βαθμολογούν παρόμοια.

Οι διάφοροι τύποι στους οποίους χωρίζονται τα συστήματα συστάσεων καθώς και οι τεχνικές που χρησιμοποιούν προκειμένου να παράγουν συστάσεις αναλύονται στο κεφάλαιο 2. Το *πρόβλημα της σύστασης (recommendation problem)* μπορεί να διατυπωθεί με τους παρακάτω δυο τρόπους:

Προβλεψη. *Η πρώτη προσέγγιση είναι να γίνει πρόβλεψη της βαθμολογίας για ένα συνδυασμό χρήστη-αντικειμένου. Υποτίθεται ότι υπάρχουν δεδομένα εκπαίδευσης, τα οποία υποδεικνύουν τις προτιμήσεις των χρηστών για αντικείμενα. Για m χρήστες και n αντικείμενα, αυτό αντιστοιχεί σε ένα $m \times n$ πίνακα, του οποίου οι γνωστές τιμές χρησιμοποιούνται για την εκπαίδευση του μοντέλου ενώ οι απροσδιόριστες προβλέπονται με τη χρήση του μοντέλου. Αυτό το πρόβλημα αναφέρεται και ως πρόβλημα συμπλήρωσης πίνακα (matrix completion problem).*

Κατάταξη. *Πολλές φορές δεν είναι απαραίτητο να προβλέψει κανείς τη βαθμολογία ενός χρήστη για ένα αντικείμενο προκειμένου να κάνει συστάσεις σε αυτόν, αλλά περισσότερο να προτείνει τα top-k αντικείμενα σε ένα χρήστη ή αντίστοιχα τους top-k χρήστες-στόχους για ένα αντικείμενο. Αυτό το πρόβλημα αναφέρεται ως πρόβλημα top-k σύστασης (top-k recommendation problem) και αποτελεί τη διατύπωση του προβλήματος της σύστασης με τη μορφή της κατάταξης.*



Σχήμα 1.1: Παράδειγμα συστήματος σύστασης ταινιών στον ιστότοπο IMDb

Έχοντας ορίσει με τους δύο παραπάνω τρόπους το πρόβλημα της σύστασης είναι πιο εύκολο να μιλήσει κάποιος για τους στόχους των συστημάτων συστάσεων. Ο πρωταρχικός στόχος των περισσότερων συστημάτων συστάσεων είναι να αυξήσει τις πωλήσεις προϊόντων και να μεγιστοποιήσει το κέρδος μέσω αυτών. Μέσω της προσεκτικής σύστασης επιλεγμένων αντικειμένων στους χρήστες,

	I_1	I_2	I_3	I_4
U_1	9	5		8
U_2	10		2	
U_3		6	1	8
U_4	9		4	10

Πίνακας 1.1: Παράδειγμα πίνακα βαθμολογιών αποτελούμενο από 4 χρήστες και 4 αντικείμενα

τα συστήματα συστάσεων φέρνουν σχετικά αντικείμενα στην προσοχή του χρήστη, αυξάνοντας τον όγκο των πωλήσεων και κατ' επέκταση τα κέρδη για την επιχείρηση [Ricc11].

Αν εξετάσει κανείς τα πράγματα από τη σκοπιά του χρήστη, είναι εύκολα κατανοητό ότι οι καλές συστάσεις μπορούν να επιδράσουν θετικά στη συνολική ικανοποίηση ενός χρήστη για έναν ιστότοπο. Δηλαδή, ένας χρήστης ο οποίος χρησιμοποιεί μια ιστοσελίδα για αγορές προϊόντων και δέχεται επανειλημμένα συστάσεις σχετικών αντικειμένων της αρεσκείας του είναι πιο πιθανό να χρησιμοποιήσει ξανά την ιστοσελίδα και επομένως να αυξήσει τις αγορές προϊόντων. Επομένως, τα συστήματα συστάσεων ενισχύουν την *καταναλωτική εμπιστοσύνη (consumer loyalty)* με την έννοια ότι οι πελάτες τείνουν να επιστρέφουν στους ιστότοπους οι οποίοι εξυπηρετούν καλύτερα τις ανάγκες τους [Scha01].

Προκειμένου όμως να επιτευχθεί ο ευρύτερος στόχος της αύξησης του κέρδους για μια επιχείρηση, μπορούν να διατυπωθούν οι παρακάτω κοινοί λειτουργικοί και τεχνικοί στόχοι των συστημάτων συστάσεων:

1. *Συνάφεια (Relevance)*: Ο πλέον προφανής και ταυτόχρονα ο πιο σημαντικός λειτουργικός στόχος ενός συστήματος συστάσεων είναι να προτείνει αντικείμενα τα οποία να είναι σχετικά με τον εκάστοτε χρήστη, εφόσον οι χρήστες είναι πιθανότερο να καταναλώσουν αντικείμενα του ενδιαφέροντός τους. Παρότι πρωταρχικός, ο στόχος της συνάφειας δεν πρέπει να απομονώνεται και αυτό φαίνεται από τη σημασία των δευτερευόντων στόχων που αναφέρονται παρακάτω.

2. *Καινοτομία (Novelty)*: Τα συστήματα συστάσεων μπορούν να γίνουν εξαιρετικά χρήσιμα όταν τα αντικείμενα τα οποία προτείνουν αποτελούν κάτι νέο για το χρήστη. Για παράδειγμα, δημοφιλείς ταινίες ενός προτιμώμενου είδους σπάνια θα ήταν καινοφανείς για το χρήστη. Πέρα από αυτό, η επαναλαμβανόμενη σύσταση δημοφιλών αντικειμένων μπορεί να οδηγήσει σε μείωση της ποικιλίας των πωλήσεων [Fled07].

3. *Έκπληξη (Serendipity)*: Παρότι σχετίζεται με την έννοια της καινοτομίας, η έκπληξη έχει περισσότερο τη σημασία της τυχαίας ανακάλυψης ενός αντικειμένου, σε αντίθεση με τις προφανείς συστάσεις [Good99]. Είναι πιθανό δηλαδή ο χρήστης να καταναλώνει προϊόντα ενός συγκεκριμένου τύπου, παρότι μπορεί να υφίσταται κάποιο κρυφό ενδιαφέρον για αντικείμενα άλλων τύπων το οποίο μπορεί ο ίδιος ο χρήστης να βρίσκει απρόσμενο. Αυτός ο στόχος είναι οι συστάσεις τέτοιου είδους αντικειμένων.

4. *Ποικιλία (Diversity)*: Από τη στιγμή που ένα τυπικό σύστημα συστάσεων μπορεί να προτείνει μια λίστα με top-k αντικείμενα, όταν αυτά είναι πολύ παρόμοια μεταξύ τους αυξάνεται ο κίνδυνος ο χρήστης να μην αρέσκεται σε κανένα από αυτά. Αντίθετα, αν η προτεινόμενη λίστα αποτελείται από αντικείμενα διαφορετικού τύπου, υπάρχει μεγαλύτερη πιθανότητα έστω ένα από αυτά να αρέσει στο χρήστη. Επομένως, η ποικιλία εξασφαλίζει ότι χρήστης δεν πρόκειται να βαρεθεί από την επαναλαμβανόμενη πρόταση ίδιων αντικειμένων.

Ο τρόπος και ο βαθμός στον οποίο οι διάφοροι τύποι συστημάτων συστάσεων πετυχαίνουν τους παραπάνω στόχους περιγράφεται πιο αναλυτικά στο Κεφάλαιο 2.

1.2 Κοινωνικά Δίκτυα

Ένα *κοινωνικό δίκτυο (social network)* είναι μια κοινωνική δομή η οποία αποτελείται από ένα σύνολο κοινωνικών "δραστών" (συνήθως πρόσωπα, ομάδες ή επιχειρήσεις) οι οποίοι ονομάζονται *κόμβοι (nodes)* και συνδέονται μεταξύ τους με έναν ή περισσότερους τύπους αλληλεξάρτησης όπως αξίες, εμπιστοσύνη, φιλία, συγγένεια, γνωριμία, επικοινωνία, οικονομικές συναλλαγές, αντιπάθεια κ.α. Οι συνδέσεις αυτές ονομάζονται *ακμές (edges)* ή *δεσμοί (ties)* ή *σύνδεσμοι (links)* [Bras98][Wass94]. Παραδείγματα κοινωνικών δικτύων αποτελούν οι ιστότοποι κοινωνικής δικτύωσης (όπως είναι το Twitter[4], το Facebook[5], το Instagram[6] κ.α.), τα δίκτυα φιλίας, τα δίκτυα συνεργασίας (όπως τα δίκτυα συν-συγγραφής επιστημονικών άρθρων), τα δίκτυα αποστολής μηνυμάτων ηλεκτρονικού ταχυδρομείου, τα δίκτυα αεροπορικών γραμμών και τα δίκτυα εξάπλωσης επιδημικών ασθενειών.

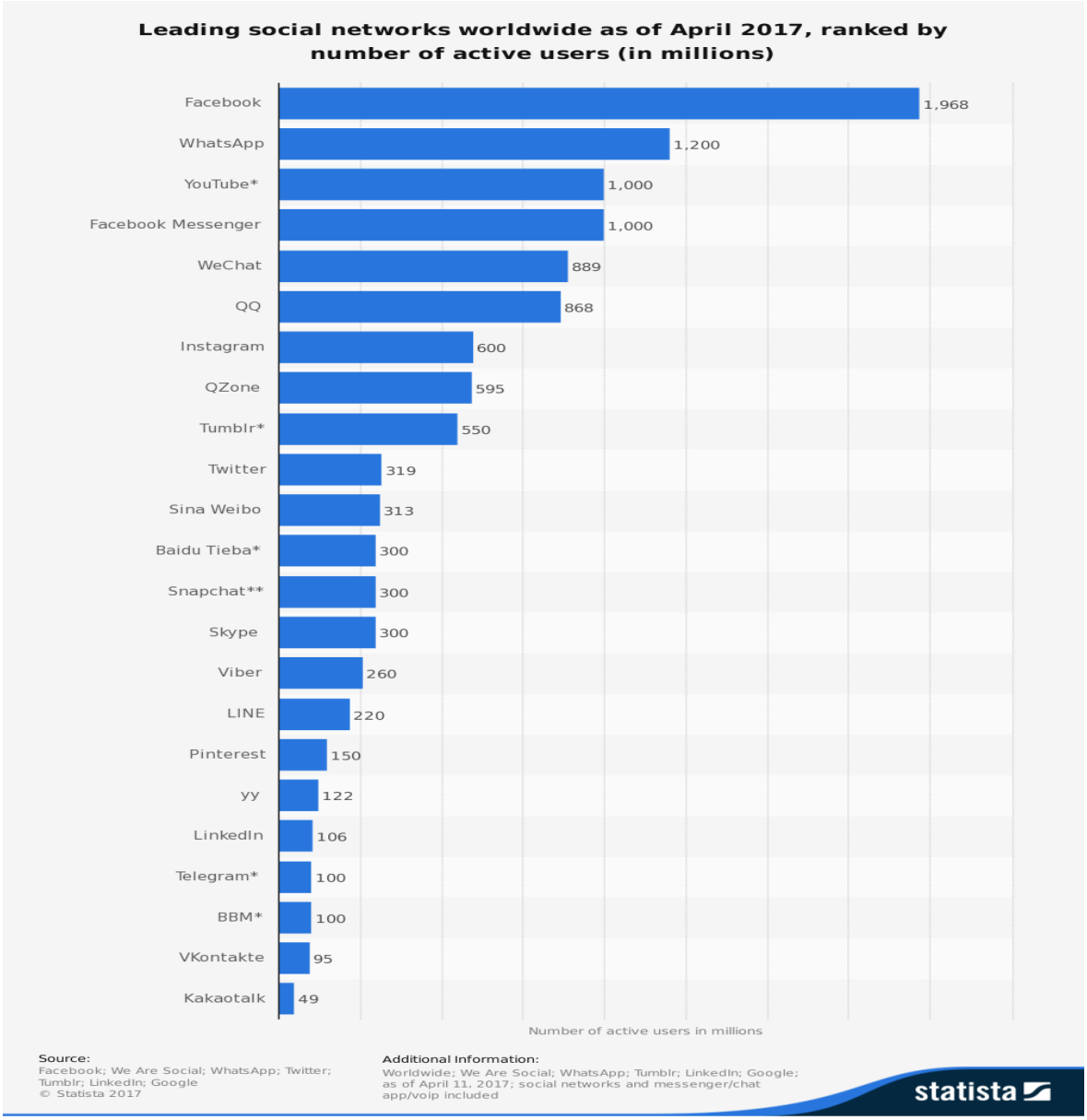
Τα κοινωνικά δίκτυα αποτελούν υποκατηγορία των πολύπλοκων δικτύων (complex networks), που ορίζονται ως δίκτυα με μη τετριμμένα τοπολογικά χαρακτηριστικά, των οποίων τα πρότυπα των συνδέσεων μεταξύ των στοιχείων τους δεν είναι ούτε αμιγώς κανονικά ούτε αμιγώς τυχαία. Τα πολύπλοκα δίκτυα περιλαμβάνουν μεταξύ άλλων τα πληροφοριακά δίκτυα (όπως είναι ο Παγκόσμιος Ιστός), τα τεχνολογικά δίκτυα (όπως είναι τα δίκτυα τηλεπικοινωνιών, τα ηλεκτρικά δίκτυα, τα οδικά δίκτυα και τα σιδηροδρομικά δίκτυα) και τα βιολογικά δίκτυα (όπως τα δίκτυα αλληλεπίδρασης πρωτεϊνών, τα τροφικά πλέγματα, τα μονοπάτια μετάδοσης σήματος και τα νευρωνικά δίκτυα).

Ο όρος *διαδικτυακά μέσα κοινωνικής δικτύωσης (online social media)* αναφέρεται στα μέσα αλληλεπίδρασης ομάδων ανθρώπων μέσω διαδικτυακών κοινοτήτων. Τα διαδικτυακά μέσα κοινωνικής δικτύωσης εμφανίζονται σε διάφορες μορφές όπως για παράδειγμα ιστολόγια, ιστοσελίδες, φόρουμ κ.α. Τα τελευταία χρόνια οι ιστότοποι κοινωνικής δικτύωσης αποτελούν τα μεγαλύτερα και πιο δημοφιλή κοινωνικά δίκτυα χάρη στα καινοτόμα χαρακτηριστικά τους. Στο Σχήμα 1.2 φαίνεται μια λίστα με τα δημοφιλέστερα online κοινωνικά δίκτυα με βάση τον αριθμό ενεργών χρηστών τους.

Τα online κοινωνικά δίκτυα ορίζονται ως διαδικτυακές υπηρεσίες οι οποίες επιτρέπουν στους χρήστες να δημιουργήσουν ένα δημόσιο ή ημι-δημόσιο προφίλ μέσα σε ένα οριοθετημένο σύστημα, να επικοινωνήσουν με μια λίστα από άλλους χρήστες με τους οποίους μοιράζονται μια μορφή σύνδεσης και να δουν και να διανείμουν την δικιά τους λίστα των συνδέσεων και αυτών που φτιάχτηκαν από άλλους μέσα στο σύστημα [boyd07]. Τα δίκτυα αυτά διαθέτουν τα παρακάτω χαρακτηριστικά στα οποία οφείλουν σε μεγάλο βαθμό τη δημοτικότητά τους:

- Υποστηρίζουν ποικιλία των μορφών περιεχομένου, όπως κείμενο, βίντεο , φωτογραφίες , ήχο κ.τ.λ. Πολλά από αυτά κάνουν χρήση παραπάνω από μιας από αυτές τις επιλογές
- Χαρακτηρίζονται από διαφορετικά επίπεδα εμπλοκής του χρήστη ο οποίος μπορεί να δημιουργεί, να σχολιάζει ή απλά να παρακολουθεί
- Βελτιώνουν την ταχύτητα και το εύρος της διάδοσης των πληροφοριών
- Προσφέρουν ενός- προς-ένα, ενός-προς-πολλούς και πολλών προς-πολλούς επικοινωνία
- Επιτρέπουν η επικοινωνία αυτή να πραγματοποιείται είτε σε πραγματικό χρόνο ή ασύγχρονα με την πάροδο του χρόνου
- Είναι ανεξάρτητα της συσκευής. Ο χρήστης μπορεί να χρησιμοποιήσει για τη δικτύωση έναν υπολογιστή ή κινητές συσκευές όπως smartphones και tablets

[4] `https://twitter.com`
[5] `https://www.facebook.com`
[6] `https://www.instagram.com`



Σχήμα 1.2: Λίστα με τα δημοφιλέστερα κοινωνικά δίκτυα με βάση τον αριθμό ενεργών χρηστών τους (πηγή https://www.slideshare.net/wearesocialsg/global-digital-statshot-q2-2017)

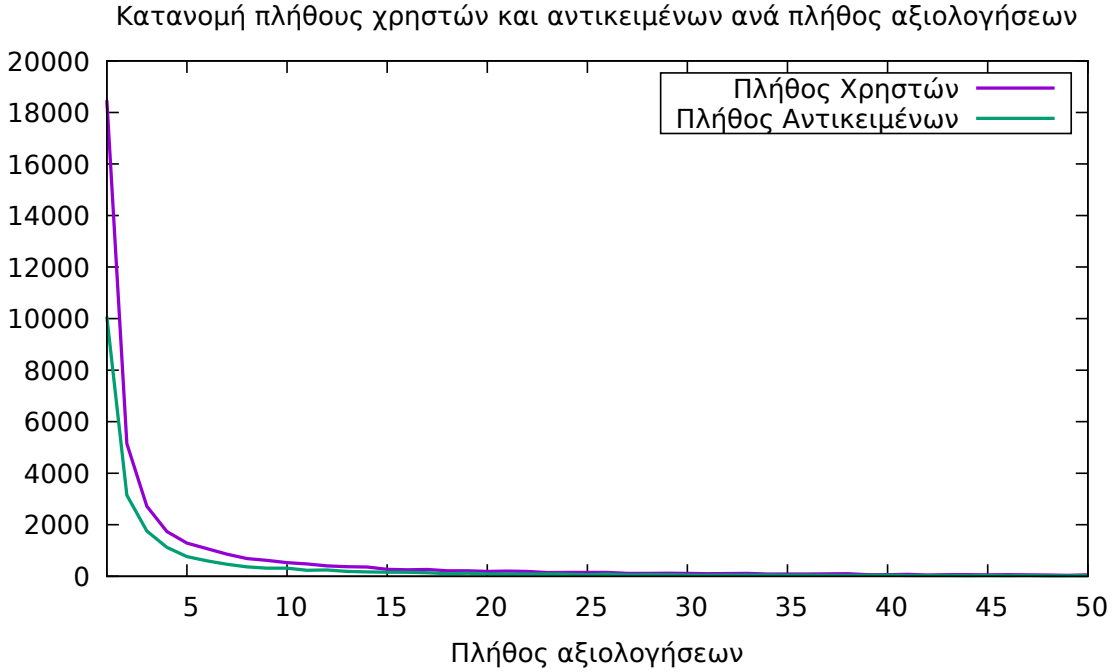
1.3 Κυριότερες Προκλήσεις

Τόσο τα συστήματα συστάσεων όσο και τα κοινωνικά δίκτυα έχουν αποτελέσει περιοχές συστηματικής έρευνας τα τελευταία χρόνια αντιμετωπίζοντας μια σειρά από προκλήσεις. Όσον αφορά τα συστήματα συστάσεων οι κυριότερες προκλήσεις που συναντώνται είναι συνοπτικά οι παρακάτω:

- *Φαινόμενο long tail*: Όταν έχουμε να κάνουμε με πραγματικά δεδομένα όπως είναι και το dataset που χρησιμοποιήθηκε στην παρούσα εργασία, είναι πολύ συχνό φαινόμενο η κατανομή των βαθμολογιών αναμεταξύ των αντικειμένων να ικανοποιεί μια ιδιότητα η οποία αναφέρεται ως *long tail*. Σύμφωνα με την ιδιότητα αυτή, μόνο ένα μικρό ποσοστό αντικειμένων βαθμολογούνται συχνά, που αποτελείται από τα πιο δημοφιλή αντικείμενα. Αντίθετα, η μεγάλη πλειοψηφία των αντικειμένων βαθμολογείται σπάνια οδηγώντας έτσι σε μια κατανομή η οποία ακολουθεί *νόμο δύναμης (power-law)* ή *νόμο Zipf (Zipf law)*. Ένας νόμος δύναμης στη στατιστική είναι μια συναρτησιακή σχέση μεταξύ δύο ποσοτήτων όπου η μια ποσότητα μεταβάλλεται ως δύναμη της

άλλης [Newm05]. Όπως φαίνεται στο Σχήμα 1.3, όπου απεικονίζεται η κατανομή του πλήθους χρηστών και αντικειμένων ανά πλήθος αξιολογήσεων, η μεγάλη πλειοψηφία των αντικειμένων/χρηστών έχουν ελάχιστα βαθμολογήσει/βαθμολογηθεί ενώ ελάχιστα αντικείμενα/χρήστες αφορούν μεγάλο αριθμό αξιολογήσεων δημιουργώντας μια μεγάλη *ουρά (tail)* στην κατανομή η οποία πλησιάζει ασυμπτωτικά το x-άξονα.

- *Αραιότητα πίνακα βαθμολογιών*: Το πρόβλημα της σύστασης όπως αναφέρθηκε στην Ενότητα 1.1 μπορεί να αναχθεί σε πρόβλημα συμπλήρωσης του $m \times n$ πίνακα βαθμολογιών, όπως αυτός του Πίνακα 1.1. Στην πλειοψηφία των συστημάτων συστάσεων, οι μη μηδενικές τιμές του πίνακα βαθμολογιών αποτελούν ποσοστό μικρότερο του 1%. Σε αυτές τις περιπτώσεις ο πίνακας βαθμολογιών χαρακτηρίζεται *αραιός (sparse)*. Ένας αραιός πίνακας βαθμολογιών δυσκολεύει σε μεγάλο βαθμό τη διαδικασία της σύστασης αφού είναι λιγότερες οι αλληλεπιδράσεις μεταξύ χρηστών και αντικειμένων στις οποίες βασίζονται όλα τα συστήματα συστάσεων.



Σχήμα 1.3: Το φαινόμενο long-tail

Από την άλλη μεριά, παρόμοιες προκλήσεις και δυσκολίες συναντώνται και στο κομμάτι της ανάλυσης κοινωνικών δικτύων όσον αφορά την ανίχνευση κοινοτήτων:

- *Έλλειψη κοινοτικής δομής*: Οι αλγόριθμοι ανίχνευσης κοινοτήτων έχουν σαν στόχο να αποκαλύπτουν την λανθάνουσα *κοινοτική δομή (community structure)* των κοινωνικών δικτύων. Ωστόσο, είναι πιθανό σε ένα δίκτυο να μην υφίσταται κοινοτική δομή με αποτέλεσμα οι αλγόριθμοι ανίχνευσης κοινοτήτων να μην αποδίδουν καλά και ουσιαστικά να "εφευρίσκουν" κοινότητες.

- *Πολυπλοκότητα αλγορίθμων ανίχνευσης κοινοτήτων*: Ένα πολύ σημαντικό πρόβλημα που αντιμετωπίζουν οι περισσότεροι αλγόριθμοι ανίχνευσης κοινοτήτων είναι η αδυναμία τους να εφαρμοστούν αποδοτικά σε πολύ μεγάλα δίκτυα. Παρότι υπάρχουν διάφορες μέθοδοι για την ανίχνευση κοινοτήτων σε κοινωνικά δίκτυα, όπως φαίνεται στο Κεφάλαιο 3, η πλειοψηφία αυτών είναι αρκετά αργοί.

1.4 Στόχοι-Δομή της Εργασίας

Η παρούσα διπλωματική εργασία έχει σαν στόχο τη μελέτη της επιρροής των κοινοτήτων που μπορεί να υφίστανται σε ένα κοινωνικό δίκτυο στη βελτίωση της ποιότητας των συστάσεων προς τα μέλη του δικτύου. Η μελέτη αυτή πραγματοποιείται με τη χρήση του *MovieTweetings*[7], ενός πραγματικού συνόλου δεδομένων από το Twitter το οποίο περιλαμβάνει τις βαθμολογίες ταινιών διάφορων χρηστών του κοινωνικού δικτύου. Το δίκτυο των χρηστών προσομοιώνεται σε μια γραφοθεωρητική βάση δεδομένων και εφαρμόζεται ένας αλγόριθμος ανίχνευσης κοινοτήτων σε αυτό.

Πιο συγκεκριμένα, στο *Κεφάλαιο 2* παρουσιάζονται οι κυριότεροι τύποι συστημάτων συστάσεων και αναλύονται λεπτομερώς τα συστήματα συνεργατικής διήθησης, τα οποία χρησιμοποιήθηκαν για το πειραματικό μέρος της εργασίας. Επίσης πραγματοποιείται σύγκριση των διαφορετικών μοντέλων συνεργατικής διήθησης με βάση τα πλεονεκτήματα και μειονεκτήματα του καθενός.

Στο *Κεφάλαιο 3* γίνεται εμβάθυνση στην ανάλυση κοινωνικών δικτύων και συγκεκριμένα στην ανίχνευση κοινοτήτων σε κοινωνικά δίκτυα. Παρουσιάζονται οι σημαντικότερες μέθοδοι ανίχνευσης κοινοτήτων και περιγράφονται αναλυτικά αλγόριθμοι όπως ο αλγόριθμος των Girvan-Newman που χρησιμοποιήθηκε στο πειραματικό κομμάτι της εργασίας. Επίσης, στην τελευταία ενότητα του κεφαλαίου γίνεται η σύνδεση κοινωνικών δικτύων και συστημάτων συστάσεων.

Το *Κεφάλαιο 4* περιλαμβάνει εκτενή περιγραφή της πειραματικής διαδικασίας και παράθεση των αποτελεσμάτων αυτής. Ειδικότερα, παρουσιάζεται και σχολιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε, γίνεται επεξήγηση των μετρικών που χρησιμοποιήθηκαν και παρουσίαση των αποτελεσμάτων με διάφορες γραφικές παραστάσεις και πίνακες.

Τέλος, στο *Κεφάλαιο 5* συνοψίζονται τα συμπεράσματα που προκύπτουν από τα πειραματικά αποτελέσματα που παρουσιάστηκαν στο Κεφάλαιο 4 και προτείνονται μελλοντικές κατευθύνσεις έρευνας στην περιοχή που μελετήθηκε.

[7] `https://github.com/sidooms/MovieTweetings`

Κεφάλαιο 2

Συστήματα Συστάσεων

Τα συστήματα συστάσεων χωρίζονται σε δύο βασικές κατηγορίες ανάλογα με τον τύπο των δεδομένων που χρησιμοποιούν για την δημιουργία συστάσεων. Πρόκειται για τα συστήματα συστάσεων *βασισμένα στη συνεργατική διήθηση (collaborative filtering recommdender systems)*, τα οποία χρησιμοποιούν ως δεδομένα τις αλληλεπιδράσεις χρηστών-αντικειμένων όπως π.χ. οι βαθμολογίες ή η αγοραστική συμπεριφορά, και για τα συστήματα συστάσεων *βασισμένα στο περιεχόμενο (content-based recommender systems)*, τα οποία χρησιμοποιούν τα χαρακτηριστικά των χρηστών και των αντικειμένων για την παραγωγή συστάσεων.

Εκτός από τις δύο βασικές αυτές κατηγορίες, υπάρχουν κι άλλα μοντέλα συστημάτων συστάσεων τα οποία συνήθως χρησιμοποιούνται σε συνδυασμό με τα παραπάνω. Ένα από αυτά είναι τα λεγόμενα *συστήματα αναφοράς (baseline)* στα οποία για τη σύσταση χρησιμοποιείται η μέση τιμή της βαθμολογίας ενός χρήστη ή ενός αντικείμενου. Άλλες κατηγορίες αποτελούν τα συστήματα *βασισμένα στη γνώση (knowledge-based)* και τα *δημογραφικά (demographic)* συστήματα. Τα πρώτα από αυτά βασίζονται σε υπάρχουσα γνώση σχετικά με τις προτιμήσεις του χρήστη, την ποικιλία των αντικείμενων και τα κριτήρια της σύστασης, ενώ τα δεύτερα βασίζονται στο δημογραφικό προφίλ του χρήστη. Επίσης, πολλές φορές συνδυάζονται τα χαρακτηριστικά των παραπάνω συστημάτων παράγοντας έτσι τα λεγόμενα *υβριδικά (hybrid)* συστήματα συστάσεων. Στην παρούσα εργασία δίνεται έμφαση στα συστήματα συνεργατικής διήθησης, τα οποία έχουν γνωρίσει και τη μεγαλύτερη διάδοση, τόσο στο ερευνητικό όσο και στο πρακτικό επίπεδο υλοποίησης.

Για την περιγραφή των διαφόρων συστημάτων συστάσεων γίνεται συστηματική χρήση του πίνακα βαθμολογιών, η έννοια του οποίου παρουσιάστηκε στην Ενότητα 1.1. Ο πίνακας βαθμολογιών συμβολίζεται με $R = [r_{ij}]$ και έχει μέγεθος $m \times n$, όπου m ο αριθμός των χρηστών και n ο αριθμός των αντικειμένων. Ο πίνακας R είναι τυπικά *ελλιπής (incomplete)* αφού μόνο ένα υποσύνολο των εγγραφών του είναι προσδιορισμένο. Η (i, j)-οστή εγγραφή του πίνακα R αναπαριστά τη βαθμολογία του χρήστη i για το αντικείμενο j και συμβολίζεται με r_{ij} όταν είναι προσδιορισμένη. Όταν η εγγραφή (i, j) αποτελεί πρόβλεψη κάποιου αλγορίθμου σύστασης (αντί να έχει καθοριστεί από έναν χρήστη) συμβολίζεται με $\widehat{r_{ij}}$.

2.1 Συστήματα αναφοράς

Τα συστήματα αναφοράς (baseline systems) είναι επί της ουσίας απλά συστήματα που εφαρμόζουν τετριμμένους αλγορίθμους παραγωγής συστάσεων. Χρησιμοποιούνται συνήθως σαν συστήματα βάσης στην περίπτωση που το κανονικό σύστημα δεν είναι σε θέση να παράγει κάποια πρόβλεψη, π.χ. όταν πρόκειται για έναν καινούργιο χρήστη ο οποίος δεν έχει βαθμολογήσει αλλά αντικείμενα. Τα δύο πιο γνωστά απλά συστήματα είναι ο *μέσος όρος χρήστη (user mean)* και ο *μέσος όρος αντικειμένου (item mean)*.

2.1.1 Μέσος όρος χρήστη

Στο σύστημα αυτό, για ένα χρήστη u, ένα αντικείμενο j προτείνεται στο χρήστη με βαθμολογία ίση με το μέσο όρο των βαθμολογιών που έχει δώσει ο χρήστης σε άλλα αντικείμενα. Επομένως, για

έναν $m \times n$ πίνακα βαθμολογιών $R = [r_{ij}]$, η συνάρτηση πρόβλεψης για το μέσο όρο χρήστη έχει την παρακάτω μορφή:

$$\hat{r}_{uk} \equiv \bar{r}_u = \frac{\sum_{k \in I_u} r_{uk}}{|I_u|} \tag{2.1}$$

όπου $|I_u|$ το πλήθος του συνόλου των αντικειμένων που έχει βαθμολογήσει ο χρήστης u.

2.1.2 Μέσος όρος αντικειμένου

Στο μέσο όρο αντικειμένου, για ένα χρήστη u, ένα αντικείμενο j προτείνεται στο χρήστη με βαθμολογία ίση με το μέσο όρο των βαθμολογιών που έχει λάβει το συγκεκριμένο αντικείμενο από άλλους χρήστες. Η συνάρτηση πρόβλεψης επομένως έχει την παρακάτω μορφή:

$$\hat{r}_{uk} \equiv \bar{r}_k = \frac{\sum_{u \in U_j} r_{uk}}{|U_j|} \tag{2.2}$$

όπου $|U_j|$ το πλήθος του συνόλου των χρηστών οι οποίοι έχουν βαθμολογήσει το αντικείμενο j.

2.2 Συνεργατική Διήθηση

Η συνεργατική διήθηση αποτελεί μια από τις πιο δημοφιλείς τεχνικές για τη δημιουργία συστημάτων συστάσεων δεδομένου ότι παρουσιάζει πολύ καλά αποτελέσματα σε περιπτώσεις όπου ο πίνακας βαθμολογιών είναι αραιός, δηλαδή οι απροσδιόριστες βαθμολογίες είναι πολύ περισσότερες από αυτές που έχουν καθοριστεί [Agga16]. Η βασική ιδέα της συνεργατικής διήθησης είναι ότι οι απροσδιόριστες αξιολογήσεις μπορούν να προσδιοριστούν καθώς οι υπάρχουσες αξιολογήσεις είναι πολύ συχνά υψηλά συσχετισμένες ανάμεσα σε συγκεκριμένους χρήστες ή αντικείμενα. Με απλά λόγια η συνεργατική διήθηση υποστηρίζει ότι εάν οι αξιολογήσεις δύο χρηστών συμφώνησαν μεταξύ τους στο παρελθόν τότε είναι πιθανότερο οι χρήστες αυτοί να συμφωνήσουν μεταξύ τους στην αξιολόγηση ενός νέου αντικειμένου στο μέλλον παρά με έναν τυχαίο χρήστη. Οι μέθοδοι που χρησιμοποιούν τα συστήματα συστάσεων συνεργατικής διήθησης χωρίζονται σε *βασισμένες στη μνήμη (memory-based)* και σε *βασισμένες στο μοντέλο (model-based)*.

2.2.1 Μέθοδοι βασισμένες στη μνήμη

Οι βασισμένες στη μνήμη μέθοδοι, οι οποίες συχνά αναφέρονται και ως *βασισμένες στη γειτονιά (neighborhood-based)*, ήταν ανάμεσα στους πρώτους αλγορίθμους συνεργατικής διήθησης των οποίων οι βαθμολογίες χρηστών-αντικειμένων προβλέπονται με βάση τη γειτονιά αυτών [Agga16]. Οι γειτονιές αυτές μπορούν να οριστούν με δύο τρόπους :

- *Συνεργατική διήθηση βασισμένη στο χρήστη (User-based collaborative filtering)*: Σε αυτή την περίπτωση, οι βαθμολογίες χρηστών παρόμοιων με έναν χρήστη-στόχο U χρησιμοποιούνται για να παράγουν συστάσεις για τον U. Οι βαθμολογίες του U υπολογίζονται ως οι σταθμισμένοι μέσοι όροι αυτών των γειτονικών χρηστών για κάθε αντικείμενο.
- *Συνεργατική διήθηση βασισμένη στο αντικείμενο (Item-based collaborative filtering)*: Προκειμένου να γίνουν συστάσεις σε έναν χρήστη U για ένα αντικείμενο-στόχο I πρέπει να οριστεί ένα σύνολο S από αντικείμενα που έχουν λάβει παρόμοια βαθμολογία με το I. Η προβλεπόμενη βαθμολογία του χρήστη U για το I προκύπτει από το σταθμισμένο μέσο όρο των βαθμολογιών των αντικειμένων που ανήκουν στο S και ο έχουν λάβει βαθμολογία από τον U.

Μια σημαντική διαφορά μεταξύ της συνεργατικής διήθησης βασισμένης στο χρήστη και της βασισμένης στο αντικείμενο είναι ότι στην πρώτη περίπτωση, οι βαθμολογίες υπολογίζονται με βάση αυτές των *γειτονικών χρηστών* ενώ στη δεύτερη χρησιμοποιούνται οι βαθμολογίες του *ίδιου χρήστη* σε *γειτονικά αντικείμενα*.

Συνεργατική διήθηση βασισμένη στο χρήστη

Σε αυτή την προσέγγιση, ορίζονται γειτονιές βασισμένες στο χρήστη προκειμένου να βρεθούν χρήστες παρόμοιοι με το χρήστη-στόχο, για τον οποίο υπολογίζονται οι προβλεπόμενες βαθμολογίες. Προκειμένου να προσδιοριστεί η γειτονιά του χρήστη-στόχου είναι απαραίτητο να υπολογιστεί η ομοιότητά του με όλους τους υπόλοιπους χρήστες. Αυτό γίνεται με τη βοήθεια μιας *συνάρτησης ομοιότητας (similarity function)* μεταξύ των βαθμολογιών των διάφορων χρηστών.

Για τον $m \times n$ πίνακα βαθμολογιών $R = [r_{ij}]$ με m χρήστες και n αντικείμενα, έστω I_u το σύνολο των δεικτών των αντικείμενων για τα οποία έχει προσδιοριστεί βαθμολογία από το χρήστη u. Για παράδειγμα, εάν έχουν καθοριστεί από το συγκεκριμένο χρήστη οι βαθμολογίες για το πρώτο, δεύτερο και πέμπτο αντικείμενο ενώ τα υπόλοιπα είναι άγνωστα, τότε $I_u = \{1, 2, 5\}$. Επομένως, το σύνολο των αντικειμένων που έχουν από κοινού βαθμολογηθεί από τους χρήστες u και v δίνεται από το $I_u \cap I_v$. Παραδείγματος χάρη, αν ο χρήστης v έχει βαθμολογήσει τα τέσσερα πρώτα αντικείμενα, τότε $I_v = \{1, 2, 3, 4\}$ και $I_u \cap I_v = \{1, 2, 5\} \cap \{1, 2, 3, 4\} = \{1, 2\}$. Ωστόσο, είναι πιθανό (και μάλιστα αρκετά συνηθισμένο) το σύνολο $I_u \cap I_v$ να είναι ίσο με το κενό σύνολο δεδομένου ότι οι πίνακες βαθμολογιών είναι γενικά αραιοί. Το σύνολο $I_u \cap I_v$ ορίζει τις κοινές βαθμολογίες των χρηστών u και v, οι οποίες χρησιμοποιούνται για τον υπολογισμό της ομοιότητας τους και επομένως την εύρεση της γειτονιάς του καθενός.

Για τον υπολογισμό της ομοιότητας μεταξύ των διανυσμάτων των βαθμολογιών δυο χρηστών u και v χρησιμοποιείται μια συνάρτηση ομοιότητας. Οι πιο γνωστές συναρτήσεις ομοιότητας που χρησιμοποιούνται είναι ο *συντελεστής συσχέτισης του Pearson (Pearson correlation coefficient)*, η *ομοιότητα διανύσματος συνημιτόνου (cosine vector similarity)*, ο *συντελεστής συσχέτισης του Spearman (Spearman's rank correlation coefficient)*, η *απόσταση Manhattan*, και η *προσαρμοσμένη ομοιότητα συνημιτόνου (adjusted cosine similarity)*.

Συναρτήσεις Ομοιότητας

Ο συντελεστής συσχέτισης του Pearson αποτελεί μια από τις πιο δημοφιλείς συναρτήσεις ομοιότητας όσον αφορά εφαρμογές συνεργατικής διήθησης βασισμένης στη γειτονιά. Για τον υπολογισμό του, το πρώτο βήμα είναι να βρεθεί η μέση βαθμολογία κάθε χρήστη u χρησιμοποιώντας τις προσδιορισμένες βαθμολογίες του:

$$\mu_u = \frac{\sum_{k \in I_u} r_{uk}}{|I_u|} \quad \forall u \in \{1...m\} \tag{2.3}$$

Στη συνέχεια, ο συντελεστής ομοιότητας του Pearson μεταξύ των χρηστών u και v ορίζεται ως εξής:

$$\text{Pearson}(u,v) = \frac{\sum_{k \in I_u \cap I_v} (r_{uk} - \mu_u) \cdot (r_{vk} - \mu_v)}{\sqrt{\sum_{k \in I_u \cap I_v} (r_{uk} - \mu_u)^2} \cdot \sqrt{\sum_{k \in I_u \cap I_v} (r_{vk} - \mu_v)^2}} \tag{2.4}$$

Όπως φαίνεται στον τύπο 2.4, ο συντελεστής συσχέτισης του Pearson εφαρμόζεται πάνω στις από κοινού προσδιορισμένες βαθμολογίες δύο χρηστών και είναι ουσιαστικά η διασπορά των διανυσμάτων βαθμολογιών προς την τυπική απόκλιση τους. Το πεδίο τιμών του είναι το κλειστό διάστημα $[-1, 1]$ με την τιμή 1 να υποδηλώνει απόλυτα θετική γραμμική συσχέτιση, την τιμή 0 όχι γραμμική συσχέτιση και την τιμή -1 απόλυτα αρνητική γραμμική συσχέτιση. Επομένως, αν ο συντελεστής συσχέτισης Pearson δύο χρηστών έχει τιμή κοντά στη μονάδα, οι χρήστες αυτοί θεωρούνται "όμοιοι" ως προς τις βαθμολογήσεις τους.

Ο συντελεστής συσχέτισης του Spearman μεταξύ δύο μεταβλητών ισούται με το συντελεστή συσχέτισης Pearson μεταξύ των τιμών κατάταξης των μεταβλητών αυτών. Σε αντίθεση με αυτόν του Pearson ο οποίος εκτιμά γραμμικές σχέσεις, ο συντελεστής συσχέτισης του Spearman εκτιμά πόσο καλά μπορεί να περιγραφεί η σχέση δυο μεταβλητών χρησιμοποιώντας μια μονότονη συνάρτηση (γραμμική η όχι). Επομένως ο τύπος για τον υπολογισμό του είναι ο ίδιος με τον παραπάνω τύπο με

τη διαφορά ότι αντί για τις τιμές βαθμολογιών r_{uk} και r_{vk} χρησιμοποιούνται οι τιμές κατάταξης των βαθμολογιών αυτών R_{uk} και R_{vk}.

$$\text{Spearman}(u,v) = \frac{\sum_{k \in I_u \cap I_v} (R_{uk} - \mu_u) \cdot (R_{vk} - \mu_v)}{\sqrt{\sum_{k \in I_u \cap I_v} (R_{uk} - \mu_u)^2} \cdot \sqrt{\sum_{k \in I_u \cap I_v} (R_{vk} - \mu_v)^2}} \tag{2.5}$$

Ο συντελεστής συσχέτισης Spearman έχει και αυτός πεδίο τιμών το [-1,1] με τις τιμές του να εκφράζουν ό,τι και αυτές του Pearson, δηλαδή θετική και αρνητική συσχέτιση όσο πιο κοντά στο 1 και το -1 βρίσκονται αντίστοιχα και μη γραμμική συσχέτιση όταν είναι κοντά στο μηδέν.

Μια άλλη ευρέως διαδεδομένη συνάρτηση ομοιότητας είναι η ομοιότητα συνημιτόνου η οποία δε χρησιμοποιεί τις κεντραρισμένες στο μέσο όρο τιμές (mean-centered values), όπως οι προηγούμενες, αλλά τις ακατέργαστες (raw values) και ορίζεται ως:

$$\text{Cosine}(u,v) = \frac{\sum_{k \in I_u \cap I_v} r_{uk} \cdot r_{vk}}{\sqrt{\sum_{k \in I_u \cap I_v} r_{uk}^2} \cdot \sqrt{\sum_{k \in I_u \cap I_v} r_{vk}^2}} \tag{2.6}$$

Από τη στιγμή που έχουμε να κάνουμε με θετικές βαθμολογίες το πεδίο τιμών της ομοιότητας συνημιτόνου μεταξύ δύο διανυσμάτων είναι το διάστημα [0,1], όπου όσο μεγαλύτερη είναι η τιμή τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ των δύο διανυσμάτων.

Επίσης, σε αρκετές εφαρμογές της ομοιότητας συνημιτόνου οι παράγοντες κανονικοποίησης του παρανομαστή βασίζονται σε όλες τις καθορισμένες βαθμολογίες και όχι σε αυτές που είναι κοινές στους δύο χρήστες, μεταβάλλοντας τον τύπο ως εξής:

$$\text{Cosine}(u,v) = \frac{\sum_{k \in I_u \cap I_v} r_{uk} \cdot r_{vk}}{\sqrt{\sum_{k \in I_u} r_{uk}^2} \cdot \sqrt{\sum_{k \in I_v} r_{vk}^2}} \tag{2.7}$$

Γενικά, ο συντελεστής συσχέτισης του Pearson και του Spearman είναι προτιμότεροι από τον τύπο του συνημιτόνου λόγω της μείωσης της πόλωσης που προσφέρει το κεντράρισμα των τιμών στο μέσο όρο. Αυτή η πόλωση οφείλεται στο γεγονός ότι διαφορετικοί χρήστες χρησιμοποιούν διαφορετικά πρότυπα βαθμολόγησης. Για παράδειγμα, μπορεί ένας χρήστης να βαθμολογεί γενικά με υψηλές τιμές της κλίμακας βαθμολόγησης ενώ ένας άλλος να είναι πιο αυστηρός χρησιμοποιώντας μόνο χαμηλές τιμές.

Η απόσταση Manhattan, γνωστή και ως *απόσταση City Block*, αποτελεί ένα άλλο μέτρο ομοιότητας που χρησιμοποιείται στα συστήματα συνεργατικής διήθησης. Η απόσταση Manhattan μεταξύ δυο σημείων ορίζεται ως το άθροισμα των απόλυτων διαφορών των καρτεσιανών συντεταγμένων τους. Στην περίπτωση που θέλουμε να βρούμε την ομοιότητα μεταξύ των χρηστών u και v η έχουμε:

$$\text{Manhattan}(u,v) = \frac{1}{(R_{max} - R_{min})|I_u \cap I_v|} \sum_{k \in I_u \cap I_v} |r_{uk} - u_{uk}| \tag{2.8}$$

Ο υπολογισμός, όπως και στις προηγούμενες συναρτήσεις, γίνεται πάνω στις κοινές βαθμολογίες των δύο χρηστών, ενώ συνήθως το άθροισμα διαιρείται με το μέγεθος του συνόλου των κοινών αντικειμένων που έχουν βαθμολογήσει οι δύο χρήστες πολλαπλασιασμένο επί τη διαφορά μέγιστης και ελάχιστης δυνατής βαθμολογίας, για λόγους κανονικοποίησης [Kaus15]. Με τη χρήση του όρου αυτού η απόσταση Manhattan λαμβάνει τιμές στο διάστημα [0,1] και όσο υψηλότερη είναι η τιμή της τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ των δύο χρηστών. Βεβαίως, είναι δυνατή η παράλειψη του παρανομαστή ωστόσο τότε η συνάρτηση παύει να λαμβάνει τιμές στο διάστημα [0,1] και επομένως δεν είναι εύχρηστη στην ανάλυση με τη χρήση κατωφλίου ομοιότητας.

Προσδιορισμός Γειτονιάς

Για τον προσδιορισμό της γειτονιάς ενός χρήστη-στόχου υπάρχουν γενικά δύο τρόποι. Ο πρώτος από αυτούς είναι η χρήση των top-N κοντινότερων χρηστών, όπου το N ουσιαστικά ισούται με το

μέγεθος της γειτονιάς. Ο δεύτερος τρόπος είναι με τη χρήση του κατωφλίου ομοιότητας, έστω t όπου γειτονιά του χρήστη-στόχου θεωρούνται οι χρήστες οι οποίο εμφανίζουν τιμή ομοιότητας μεγαλύτερη ή ίση με t. Ωστόσο, επειδή μόνο με τον ένα ή τον άλλο τρόπο μπορεί από τη μία οι top-N γείτονες να έχουν μικρή τιμή ομοιότητας με το χρήστη-στόχο και από την άλλη οι χρήστες των οποίων η τιμή ομοιότητας ξεπερνάει το κατώφλι t να είναι πάρα πολλοί, συνήθως χρησιμοποιείται ο συνδυασμός των δύο μεθόδων. Αυτό σημαίνει ότι η τελική γειτονιά ενός χρήστη u αποτελείται από τους top-N χρήστες, οι οποίοι έχουν τιμή ομοιότητας με το χρήστη u μεγαλύτερη ή ίση με t. Εδώ σημειώνεται ότι δεν είναι πάντα σίγουρο ότι θα υπάρχουν N όμοιοι χρήστες με το χρήστη u οι οποίοι να έχουν προσδιορίσει βαθμολογία για ένα αντικείμενο j. Αυτό το σενάριο είναι ιδιαίτερα συνηθισμένο σε αραιούς πίνακες βαθμολογιών.

Ένας τρόπος εύρεσης της γειτονιάς του χρήστη-στόχου θα ήταν αυτή να αποτελείται από το σύνολο των πιο όμοιων χρηστών, που βρέθηκαν με τη βοήθεια των συναρτήσεων ομοιότητας που περιγράφηκαν παραπάνω, για το σύνολο των αντικειμένων που έχει βαθμολογήσει ο χρήστη-στόχος. Ωστόσο, δεδομένου ότι ο αριθμός των καθορισμένων βαθμολογιών μεταξύ των όμοιων χρηστών μπορεί να ποικίλει για κάθε διαφορετικό αντικείμενο, οι N κοντινότεροι χρήστες στο χρήστη-στόχο υπολογίζονται εκ νέου για κάθε ξεχωριστό αντικείμενο έτσι ώστε καθένας από αυτούς να έχει ορίσει βαθμολογία για το εκάστοτε αντικείμενο [Agga16].

Συναρτήσεις πρόβλεψης

Από τη στιγμή που έχει υπολογιστεί η γειτονιά του χρήστη-στόχου, δηλαδή το σύνολο των πιο όμοιων χρηστών (peer group), οι προβλεπόμενες βαθμολογίες υπολογίζονται χρησιμοποιώντας μια *συνάρτηση πρόβλεψης (prediction function)* η οποία εμφανίζεται σε αρκετές παραλλαγές. Η πιο απλή περίπτωση είναι η προβλεπόμενη βαθμολογία να δίνεται ως ο σταθμισμένος μέσος όρος των βαθμολογίων των k κοντινότερων χρηστών του χρήστη-στόχου, με το βάρος κάθε βαθμολογίας να ισούται με την τιμή της ομοιότητας του εκάστοτε χρήστη ως προς το χρήστη-στόχο. Επομένως εάν $P_u(j)$ είναι το σύνολο των πιο όμοιων χρηστών για το χρήστη u οι οποίοι έχουν ορίσει βαθμολογίες για το αντικείμενο j, τότε η προβλεπόμενη βαθμολογία ορίζεται ως:

$$\hat{r}_{uj} = \frac{\sum_{u \in P_u(j)} Sim(u,v) \cdot r_{uj}}{\sum_{u \in P_u(j)} |Sim(u,v)|} \tag{2.9}$$

όπου το βάρος $Sim(u,v)$ είναι η τιμή της ομοιότητας του χρήστη v με το χρήστη-στόχο u.

Ένα σημαντικό πρόβλημα που παρουσιάζει η παραπάνω προσέγγιση είναι το γεγονός ότι διαφορετικοί χρήστες βαθμολογούν σε διαφορετικές κλίμακες, δηλαδή κάποιοι χρήστες συνηθίζουν να δίνουν υψηλές βαθμολογίες γενικότερα στα αντικείμενα ενώ άλλοι χαμηλές. Προκειμένου να εξαλειφθεί αυτό το πρόβλημα, οι ακατέργαστες βαθμολογίες πρέπει να κεντραριστούν στο μέσο όρο των βαθμολογιών του χρήστη. Η κεντραρισμένη στο μέσο όρο βαθμολογία s_{uj} ενός χρήστη u για ένα αντικείμενο j ορίζεται αφαιρώντας τη μέση βαθμολογία του από την ακατέργαστη όπως παρακάτω:

$$\hat{r}_{uj} = r_{uj} - \mu_u \quad \forall u \in \{1...m\} \tag{2.10}$$

Ο σταθμισμένος μέσος όρος της κεντραρισμένης βαθμολογίας μας δίνει μια πρόβλεψη κεντραρισμένη στη μέση τιμή (mean-centered prediction). Επομένως, προκειμένου να δοθεί μια ακατέργαστη προβλεπόμενη βαθμολογία (raw prediction), η μέση τιμή των βαθμολογιών του χρήστη μ_u επαναπροστίθεται στην πρόβλεψη αυτή δίνοντας την παρακάτω συνάρτηση πρόβλεψης, γνωστή και ως *τύπος του Resnick (Resnick's formula)* [Resn94]:

$$\hat{r}_{uj} = \mu_u + \frac{\sum_{u \in P_u(j)} Sim(u,v) \cdot s_{uj}}{\sum_{u \in P_u(j)} |Sim(u,v)|} = \mu_u + \frac{\sum_{u \in P_u(j)} Sim(u,v) \cdot (r_{uj} - \mu_u)}{\sum_{u \in P_u(j)} |Sim(u,v)|} \tag{2.11}$$

Υπάρχουν αρκετές παραλλαγές της ανωτέρω συνάρτησης πρόβλεψης. Για παράδειγμα, αντί να χρησιμοποιείται η κεντραρισμένη στη μέση τιμή βαθμολογία s_{uj}, μια άλλη προσέγγιση χρησιμοποιεί

το Z-score z_{uj} το οποίο προκύπτει από τη διαίρεση του s_{uj} με την τυπική απόκλιση σ_u των καθορισμένων βαθμολογιών του χρήστη u, η οποία ορίζεται ως εξής:

$$\sigma_u = \sqrt{\frac{\sum_{j \in I_u} (r_{uj} - \mu_u)^2}{|I_u| - 1}} \quad \forall u \in \{1...m\} \tag{2.12}$$

Έτσι, η βαθμολογία υπολογίζεται σύμφωνα με τον τύπο:

$$z_{uj} = \frac{r_{uj} - \mu_u}{\sigma_u} = \frac{s_{uj}}{\sigma_u} \tag{2.13}$$

και η συνάρτηση πρόβλεψης παιρνει την παρακατω μορφη :

$$\hat{r}_{uj} = \mu_u + \sigma_u \frac{\sum_{u \in P_u(j)} Sim(u,v) \cdot z_{uj}}{\sum_{u \in P_u(j)} |Sim(u,v)|} \tag{2.14}$$

Όπως φαίνεται στον παραπάνω τύπο, ο όρος του σταθμισμένου μέσου όρου πολλαπλασιάζεται με την τυπική απόκλιση σ_u. Γενικότερα, όταν μια συνάρτηση $g(\cdot)$ εφαρμόζεται στην κανονικοποίηση των βαθμολογιών, τότε η αντίστροφη αυτής πρέπει να εφαρμοστεί στην τελική διαδικασία της πρόβλεψης [Agga16]. Παρότι είναι γενικά αποδεκτό ότι η κανονικοποίηση βελτιώνει την ακρίβεια της πρόβλεψης, φαίνεται να υπάρχουν αντικρουόμενα συμπεράσματα σε διάφορες μελέτες σχετικά με το εάν το κεντράρισμα στη μέση τιμή ή το Z-score οδηγεί σε καλύτερα αποτελέσματα [Herl99],[Howe08].

Συνεργατική διήθηση βασισμένη στο αντικείμενο

Στη συνεργατική διήθηση βασισμένη στο αντικείμενο η διαδικασία είναι αντίστοιχη με αυτή της προηγούμενης υποενότητας, με τη διαφορά ότι οι γειτονιές φτιάχνονται με αντικείμενα και όχι χρήστες. Για αυτό το λόγο, οι ομοιότητες πρέπει να υπολογιστούν μεταξύ των αντικειμένων, δηλαδή των στηλών του πίνακα βαθμολογιών. Έστω U_i το σύνολο των χρηστών οι οποίοι έχουν βαθμολογήσει το αντικείμενο i. Για παράδειγμα, εάν ο πρώτος,ο δεύτερος και ο πέμπτος χρήστης έχουν δώσει βαθμολογίες για το αντικείμενο i τότε έχουμε $U_i = \{1, 2, 5\}$.

Όπως στην περίπτωση της συνεργατικής διήθησης βασισμένης στο χρήστη, ο μέσος όρος των βαθμολογιών του αντικειμένου αφαιρείται από κάθε βαθμολογία δίνοντας έτσι μια κεντραρισμένη στη μέση τιμή βαθμολογία.Για τον υπολογισμό της γειτονιάς των όμοιων αντικειμένων μπορεί να χρησιμοποιηθεί οποιαδήποτε από τις συναρτήσεις ομοιότητας που περιγράφηκαν προηγουμένως, ωστόσο, η *προσαρμοσμένη ομοιότητα συνημιτόνου (adjusted cosine similarity)* φαίνεται να δίνει γενικά καλύτερα αποτελέσματα από τις υπόλοιπες, στην περίπτωση των αντικειμένων [Agga16].

Ο τύπος της προσαρμοσμένης ομοιότητας συνημιτόνου μεταξύ των αντικειμένων i και j ορίζεται ως εξής:

$$\text{AdjustedCosine}(i,j) = \frac{\sum_{u \in U_i \cap U_j} s_{ui} \cdot s_{uj}}{\sqrt{\sum_{u \in U_i \cap U_j} s_{ui}^2} \cdot \sqrt{\sum_{u \in U_i \cap U_j} s_{uj}^2}} \tag{2.15}$$

Έστω η περίπτωση όπου ζητείται να βρεθεί η βαθμολογία του αντικειμένου-στόχου t για το χρήστη u. Αρχικά είναι απαραίτητο να καθοριστούν τα παρόμοια αντικείμενα με το αντικείμενο t με βάση τη συνάρτηση ομοιότητας που έχει επιλεγεί. Έστω $Q_t(u)$ το σύνολο των k γειτονικών αντικειμένων για τα οποία ο χρήστης u έχει προσδιορίσει βαθμολογία. Ο σταθμισμένος μέσος όρος των βαθμολογιών αυτών αποτελεί την προβλεπόμενη βαθμολογία του αντικειμένου-στόχου t ενώ το βάρος κάθε αντικειμένου j σε αυτό το μέσο όρο είναι η τιμή της προσαρμοσμένης ομοιότητας συνημιτόνου μεταξύ του j και του t. Επομένως, η τελική συνάρτηση πρόβλεψης της βαθμολογίας $\widehat{r}_{ut}$ του χρήστη u για το αντικείμενο-στόχο t είναι η παρακάτω:

$$\widehat{r}_{ut} = \frac{\sum_{j \in Q_t(u)} AdjustedCosine(j,t) \cdot r_{uj}}{\sum_{j \in Q_t(u)} |AdjustedCosine(j,t)|} \tag{2.16}$$

Το βασικό πλεονέκτημα της μεθόδου αυτής είναι ότι χρησιμοποιούνται οι βαθμολογίες του ίδιου χρήστη σε παρόμοια αντικείμενα για την τελική πρόβλεψη της βαθμολογίας. Για παράδειγμα, σε ένα σύστημα συστάσεων για κινηματογραφικές ταινίες, το σύνολο $Q_t(u)$ των γειτονικών αντικειμένων ενός αντικειμένου t θα είναι συνήθως ταινίες του ίδιου είδους. Σε μια τέτοια περίπτωση, οι παλαιότερες βαθμολογίες του ίδιου χρήστη σε αυτές τις ταινίες αποτελούν μια αρκετά αξιόπιστη πρόβλεψη για τη ζητούμενη βαθμολογία.

Όπως και στην περίπτωση της βασισμένης στο χρήστη συνεργατικής διήθησης, πολλές φορές χρησιμοποιείται ο τύπος του Resnick προσαρμοσμένος στο μοντέλο με βάση τα αντικείμενα. Στον τύπο αυτό αφαιρείται από την ακατέργαστη βαθμολογία ο μέσος όρος μ_j των βαθμολογιών του αντικειμένου j και επίσης προστίθεται στην τελική συνάρτηση πρόβλεψης ο μέσος όρος μ_t των βαθμολογιών του αντικειμένου-στόχου t. Ο τύπος του Resnick για συνεργατική διήθηση βασισμένη στα αντικείμενα φαίνεται παρακάτω [Resn94]:

$$\widehat{r}_{ut} = \mu_t + \frac{\sum_{j \in Q_t(u)} Sim(j,t) \cdot s_{uj}}{\sum_{j \in Q_t(u)} |Sim(j,t)|} = \mu_t + \frac{\sum_{j \in Q_t(u)} Sim(j,t) \cdot (r_{uj} - \mu_j)}{\sum_{j \in Q_t(u)} |Sim(j,t)|} \tag{2.17}$$

Γενικά, επειδή τα συστήματα που είναι βασισμένα στο αντικείμενο είναι παρόμοια με τα αντίστοιχα που έχουν ως βάση το χρήστη, οι διάφορες παραλλαγές των συναρτήσεων ομοιότητας και πρόβλεψης που περιγράφηκαν παραπάνω μπορούν να τροποποιηθούν ώστε να εφαρμοστούν και στην περίπτωση των αντικειμένων.

2.2.2 Σύγκριση των βασισμένων στο χρήστη και βασισμένων στο αντικείμενο μεθόδων

Οι μέθοδοι που είναι βασισμένες στο αντικείμενο συχνά παρέχουν πιο ακριβείς προβλέψεις χάρη στο πλεονέκτημα που αναφέρθηκε προηγουμένως, δηλαδή ότι βασίζονται στις βαθμολογίες του *ίδιου* του χρήστη για τον οποίο γίνεται η πρόβλεψη για την παραγωγή συστάσεων. Στις μεθόδους αυτές, επιλέγονται παρόμοια αντικείμενα με το αντικείμενο-στόχο και χρησιμοποιούνται οι βαθμολογίες του ίδιου του χρήστη σε αυτά τα αντικείμενα για να προβλεφθεί η βαθμολογία του αντικειμένου-στόχου. Έτσι, ο μέσος όρος των βαθμολογιών του ίδιου του χρήστη στο σύνολο των παρόμοιων αντικειμένων είναι πολλές φορές αρκετά ενδεικτικός της προτίμησης που μπορεί να έχει ο χρήστης για το αντικείμενο-στόχο. Αντίθετα, αυτό δε συμβαίνει στην περίπτωση των μεθόδων που είναι βασισμένες στο χρήστη, όπου οι προβλέψεις εξάγονται από τις βαθμολογίες διαφορετικών χρηστών οι οποίοι μπορεί να έχουν επικαλυπτόμενα αλλά διαφορετικά ενδιαφέροντα. Ως αποτέλεσμα, οι μέθοδοι που είναι βασισμένες στο αντικείμενο αποδίδουν συχνά μεγαλύτερη ακρίβεια [Agga16].

Ένα άλλο πλεονέκτημα των μεθόδων που βασίζονται στο αντικείμενο είναι ότι μπορούν να παρέχουν στο χρήστη ακριβή αιτιολόγηση για τη σύσταση ενός αντικειμένου. Για παράδειγμα, εάν πρόκειται για ταινίες κινηματογράφου, μια πολύ συχνή πρόταση ενός συστήματος συστάσεων είναι το παρακάτω :

Επειδή παρακολουθήσατε την ταινία Α, σας προτείνουμε τις ταινίες : < Ταινία Β, Ταινία Γ ...>

Αυτή η αιτιολόγηση δεν είναι εύκολο να προσδιοριστεί στην περίπτωση των βασισμένων στο χρήστη μεθόδων όπου η γειτονία είναι ένα σύνολο ανώνυμων χρηστών των οποίων τα στοιχεία δεν είναι διαθέσιμα κατά τη διαδικασία της σύστασης. Αυτό αναγκάζει τα συστήματα αυτά να παρέχουν άλλου τύπου επεξηγήσεις για τις συστάσεις τους όπως για παράδειγμα ένα ιστόγραμμα που να δείχνει πώς αυτές οι ταινίες έχουν βαθμολογηθεί από χρήστες με παρόμοιες προτιμήσεις. Αυτού του είδους η αιτιολόγηση ωστόσο, μειώνει την ισχύ της σύστασης αφού ο χρήστης δεν μπορεί να καταλάβει πως οι προτεινόμενες ταινίες σχετίζονται με τα δικά του ενδιαφέροντα ή με ανθρώπων που ξέρει και εμπιστεύεται.

Τέλος, οι μέθοδοι βασισμένες στο αντικείμενο παρουσιάζουν μεγαλύτερη σταθερότητα ένεκα αλλαγών στις βαθμολογίες και αυτό οφείλεται σε δύο λόγους. Ο πρώτος είναι ότι γενικά ο αριθμός των χρηστών είναι μεγαλύτερος από αυτόν των αντικειμένων. Για το λόγο αυτό, δύο χρήστες μπορεί να έχουν έναν πολύ μικρό αριθμό αντικειμένων που έχουν από κοινού βαθμολογήσει ενώ αντίθετα δύο αντικείμενα είναι πολύ πιθανό να έχουν ένα μεγάλο αριθμό χρηστών οι οποίοι τα έχουν βαθμολογήσει και τα δύο. Έτσι, στην περίπτωση των μεθόδων που βασίζονται στο χρήστη, η προσθήκη ορισμένων καινούργιων βαθμολογιών μπορεί να αλλάξει δραστικά τις τιμές των ομοιοτήτων ενώ δε συμβαίνει το ίδιο και στην περίπτωση των αντικειμένων. Ο δεύτερος λόγος είναι ότι σε ένα σύστημα συστάσεων είναι πολύ πιθανότερο να εισάγονται νέοι χρήστες παρά νέα αντικείμενα. Έτσι, ενώ ο υπολογισμός των γειτονιών των αντικειμένων μπορεί να γίνεται περιστασιακά καθώς αυτές δεν αναμένεται να αλλάξουν δραστικά με την είσοδο νέων χρηστών, ο υπολογισμός των γειτονικών χρηστών οφείλει να γίνεται αρκετά συχνότερα.

Από την άλλη μεριά, τα βασισμένα στο χρήστη μοντέλα, παρότι συχνά υστερούν σε ακρίβεια, πολλές φορές υπερτερούν στις έννοιες της ποικιλίας (diversity) και της έκπληξης (serendipity) (Ενότητα 1.1) έναντι των βασισμένων στο αντικείμενο μοντέλων. Οι δύο αυτές έννοιες είναι πολύ σημαντικές διότι, όταν σε μια λίστα από προτεινόμενα αντικείμενα δεν υπάρχει ποικιλία, τότε στην περίπτωση που στο χρήστη δεν αρέσει το πρώτο αντικείμενο είναι πιθανό να μην αρέσει και κανένα από τα υπόλοιπα. Οι μέθοδοι που βασίζονται στο αντικείμενο τείνουν πολλές φορές να προτείνουν προφανή ή αναμενόμενα αντικείμενα, γεγονός το οποίο μπορεί να μην είναι ευχάριστο για το χρήστη.

2.2.3 Μοντέλα γράφων για συνεργατική διήθηση βασισμένη στη γειτονιά

Η αραιότητα των πινάκων βαθμολογιών αποτελεί σημαντικό πρόβλημα για τον υπολογισμό της ομοιότητας. Έναν τρόπο αντιμετώπισης αυτού του προβλήματος αποτελούν τα μοντέλα γράφων. Οι γράφοι παρέχουν μια δομική αναπαράσταση των σχέσεων μεταξύ χρηστών και αντικειμένων, επιτρέποντας την εφαρμογή πολλών αλγοριθμικών εργαλείων από την ερευνητική περιοχή της γραφοθεωρίας. Οι γράφοι μπορούν να κατασκευαστούν με βάση τους χρήστες, τα αντικείμενα ή το συνδυασμό τους. Σε αυτούς τους διαφορετικούς τύπους γράφων εφαρμόζεται πληθώρα αλγορίθμων οι οποίοι χρησιμοποιούν μεθόδους *συντομότερης διαδρομής (shortest-path)* ή *τυχαίων περιπάτων (random-walks)* για τη διαδικασία της σύστασης. Παρακάτω παρουσιάζονται οι τρεις βασικοί τύποι γράφων μαζί με τους αλγορίθμους που εφαρμόζονται σε αυτούς.

Γράφοι χρήστη-αντικειμένου

Ένας *γράφος χρήστη-αντικειμένου (user-item graph)* ορίζεται ως ένας διμερής (bipartite), μη κατευθυντικός γράφος $G = (N_u \cup N_i, A)$, όπου N_u είναι το σύνολο των κόμβων που αντιπροσωπεύουν χρήστες και N_i το σύνολο αυτών που αντιπροσωπεύουν αντικείμενα. Το A είναι το σύνολο των ακμών του γράφου οι οποίες συνδέουν μόνο χρήστες με αντικείμενα. Μια μη κατευθυντική ακμή μεταξύ του χρήστη u και του αντικειμένου i ανήκει στο A μόνο αν ο χρήστης u έχει βαθμολογήσει το αντικείμενο i. Γι' αυτό το λόγο, ο αριθμός των ακμών ισούται με τον αριθμό των προσδιορισμένων βαθμολογιών του πίνακα.

Ένα παράδειγμα γράφου χρήστη-αντικειμένου φαίνεται στο Σχήμα 2.1. Το σημαντικότερο πλεονέκτημα των συστημάτων που βασίζονται σε γράφους είναι ότι δύο χρήστες δε χρειάζεται να έχουν βαθμολογήσει από κοινού πολλά αντικείμενα για να θεωρηθούν γείτονες, από τη στιγμή που θα υπάρχουν ήδη μικρά μονοπάτια μεταξύ τους. Βέβαια, σε περίπτωση που οι δύο χρήστες έχουν βαθμολογήσει από κοινού πολλά αντικείμενα μπορούν να θεωρηθούν ως κοντινοί γείτονες. Έτσι, αυτός ο ορισμός επιτρέπει την δημιουργία γειτονιών με την έννοια της *έμμεσης συνεκτικότητας (indirect connectivity)* μεταξύ των κόμβων. Αυτή η προσέγγιση παρέχει ένα διαφορετικό τρόπο ορισμού της γειτονιάς, ο οποίος μπορεί να είναι χρήσιμος σε περιπτώσεις αραιών πινάκων βαθμολογιών.

Δύο συνηθισμένες μέθοδοι προσδιορισμού της έννοιας της έμμεσης συνδεσιμότητας είναι η χρήση μετρικών τυχαίων περιπάτων και της μετρικής του *Katz*, οι οποίες παρουσιάζονται παρακάτω.

Προσδιορισμός της γειτονιάς με τη χρήση τυχαίων περιπάτων

Με τη χρήση των τυχαίων περιπάτων, η γειτονιά ενός χρήστη u μπορεί να προσδιοριστεί ως το σύνολο των χρηστών οι οποίοι συναντώνται σε έναν τυχαίο περίπατο, ο οποίος ξεκινάει από το χρήστη u. Επομένως, χρησιμοποιώντας αλγορίθμους τυχαίων περιπάτων όπως οι *PageRank* και *SimRank* μπορούν να προσδιοριστούν οι k κοντινότεροι χρήστες ενός χρήστη για την εφαρμογή συνεργατικής διήθησης βασισμένης στο χρήστη. Με αντίστοιχο τρόπο, μπορούν να καθοριστούν τα k πιο παρόμοια αντικείμενα ενός δεδομένου αντικειμένου i, προκειμένου να εφαρμοστεί συνεργατική διήθηση βασισμένη στο αντικείμενο, ξεκινώντας έναν τυχαίο περίπατο από το i.

Αυτή η προσέγγιση παρουσιάζει ένα σημαντικό πλεονέκτημα όσον αφορά τον προσδιορισμό της γειτονιάς σε αραιούς πίνακες βαθμολογιών, έναντι των μεθόδων που παρουσιάστηκαν προηγουμένως, όπως για παράδειγμα ο συντελεστής ομοιότητας του Pearson. Στην περίπτωση της ομοιότητας του Pearson, δύο χρήστες πρέπει να συνδέονται άμεσα μέσω κάποιου αντικειμένου που έχουν βαθμολογήσει και οι δύο προκειμένου να μπορέσει να οριστεί γειτονιά. Αυτή η σύνδεση, ωστόσο, στις περιπτώσεις αραιών πινάκων βαθμολογιών μπορεί να μην υφίσταται για αρκετούς χρήστες. Από την άλλη, μια μέθοδος τυχαίου περιπάτου εκμεταλλεύεται και τις έμμεσες συνδέσεις, αφού ένας περίπατος από ένα χρήστη σε έναν άλλο μπορεί να περιλαμβάνει οποιοδήποτε αριθμό βημάτων. Επομένως, αν μεγάλα τμήματα του γράφου χρηστών-αντικειμένων είναι συνδεδεμένα, μπορεί πάντα να οριστεί γειτονιά μεταξύ των χρηστών η των αντικειμένων.

Προσδιορισμός της γειτονιάς με τη χρήση της μετρικής του Katz

Η *κεντρικότητα κατά Katz (Katz's centrality)* ενός κόμβου, στη θεωρία γράφων, αποτελεί μια μετρική κεντρικότητας σε δίκτυα. Παρουσιάστηκε από τον Leo Katz το 1953 και χρησιμοποιείται για τη μέτρηση του βαθμού επιρροής ενός κόμβου μέσα σε ένα δίκτυο [Katz53]. Σε αντίθεση με άλλες τυπικές μετρικές κεντρικότητας, οι οποίες βασίζονται μόνο στις συντομότερες διαδρομές μεταξύ των κόμβων ενός γραφου, η μετρική του Katz μετράει την επιρροή ενός κόμβου λαμβάνοντας υπόψη το συνολικό αριθμό περιπάτων μεταξύ ενός ζεύγους κόμβων.

Έστω $n_{ij}^{(t)}$ ο αριθμός των περιπάτων μήκους t μεταξύ των κόμβων i και j. Τότε, για ένα βάρος $\beta < 1$, η μετρική του Katz μεταξύ των κόμβων αυτών ορίζεται ως εξής [Agga11]:

$$\textit{Katz(i,j)} = \sum_{t=1}^{\infty} \beta^t \cdot n_{ij}^{(t)} \tag{2.18}$$

Το βάρος β είναι ένας συντελεστής απόσβεσης ο οποίος μειώνει την επιρροή περιπάτων μεγάλου μήκους στον τελικό υπολογισμό του συντελεστή. Το β λαμβάνει αρκετά μικρές τιμές ώστε το παραπάνω άθροισμα να συγκλίνει.

Η μετρική του Katz μπορεί να βρεθεί συνολικά για όλα τα ζεύγη κόμβων χρησιμοποιώντας τον πίνακα γειτνίασης. Εάν A είναι ο συμμετρικός πίνακας γειτνίασης ενός μη-κατευθυντικού γράφου, όπως αυτός του Σχήματος 3.1b, τότε μπορεί να υπολογιστεί ο $m \times m$ πίνακας των συντελεστών Katz μεταξύ των ζευγών των χρηστών (ή των αντικειμένων) όπως φαίνεται παρακάτω:

$$K = \sum_{t=1}^{\infty} (\beta A)^i = (I - \beta A)^{-1} - I \tag{2.19}$$

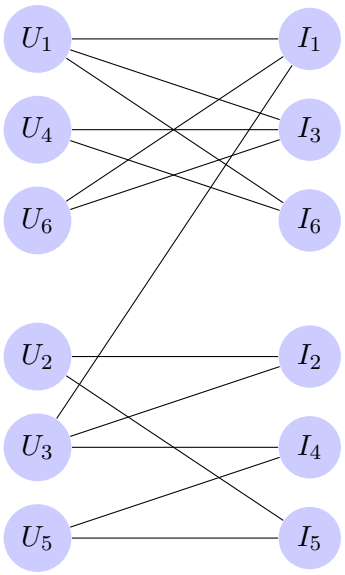
Οι top $-\ k$ κόμβοι με το μεγαλύτερο συντελεστή Katz ως προς ένα κόμβο-στόχο t μπορούν να θεωρηθούν ως η γειτονιά του t.

Γράφοι χρήστη-χρήστη

Στους γράφους χρήστη-αντικειμένου που περιγράφηκαν παραπάνω, η σύνδεση μεταξύ χρηστών ορίζεται από έναν άρτιο αριθμό βημάτων (hops), αφού δύο χρήστες συνδέονται μέσω ενός ή παραπάνω αντικειμένων. Επομένως υπάρχει η δυνατότητα, αντί για γράφο χρήστη αντικειμένου, κά-

	I_1	I_2	I_3	I_4	I_5	I_6
U_1	3		10			4
U_2		3			8	
U_3	10	6		2		
U_4			7			8
U_5				6	10	
U_6	9		7			

(a) Πίνακας βαθμολογιών



(b) Γράφος χρήστη-αντικειμένου

Σχήμα 2.1: Πίνακας βαθμολογιών και ο αντίστοιχος γράφος χρήστη-αντικειμένου

ποιος να κατασκευάσει απευθείας ένα *γράφο χρήστη-χρήστη (user-user graph)*, βασισμένο στη *συνεκτικότητα 2-βημάτων (2-hop connectivity)*. Το πλεονέκτημα των γράφων χρήστη-χρήστη έναντι των γράφων χρήστη-αντικειμένου είναι ότι στους πρώτους οι ακμές του γράφου περιέχουν περισσότερη πληροφορία. Αυτό συμβαίνει διότι η συνεκτικότητα 2-βημάτων λαμβάνει υπόψη τον αριθμό και την ομοιότητα των κοινών αντικειμένων μεταξύ δύο χρηστών κατά τη δημιουργία των ακμών. Οι γράφοι χρήστη-χρήστη χρησιμοποιούν την έννοια της *κάλυψης* προκειμένου να ποσοτικοποιήσουν τον αριθμό των από κοινού προσδιορισμένων βαθμολογιών μεταξύ δύο χρηστών (κόμβων) και την έννοια της *προβλεψιμότητας (predictability)* προκειμένου να ποσοτικοποιήσουν το επίπεδο ομοιότητας μεταξύ των κοινών αυτών βαθμολογιών.

Έστω I_u το σύνολο των αντικειμένων για τα οποία ο χρήστης u έχει προσδιορίσει βαθμολογίες και I_v το σύνολο των αντικειμένων για τα οποία έχουν προσδιοριστεί βαθμολογίες από το χρήστη v. Οι ακμές του γράφου ορίζονται με την έννοια της κάλυψης. Η κάλυψη είναι μια ασύμμετρη σχέση μεταξύ των χρηστών η οποία ορίζεται με βάση τα από κοινού βαθμολογημένα αντικείμενα.

Ορισμός 1 (Κάλυψη). Ένας χρήστης u λέγεται ότι καλύπτει ένα χρήστη v στο επίπεδο (F, G), εάν κάποιο από τα ακόλουθα είναι αληθές:

$$|I_u \cap I_v| \geq F$$

$$|I_u \cap I_v|/|I_u| \geq G$$

Εδώ τα F,G είναι παράμετροι του αλγορίθμου. Σημειώνεται ότι αρκεί μια από τις δυο παραπάνω συνθήκες να αληθεύει προκειμένου ο χρήστης u να καλύπτει το χρήστη v. Η έννοια της κάλυψης χρησιμοποιείται για να οριστεί μετέπειτα η έννοια της προβλεψιμότητας.

Ορισμός 2 (Προβλεψιμότητα). Ο χρήστης v λέγεται ότι προβλέπει το χρήστη u, εάν ο u καλύπτει τον v και υπάρχει μια γραμμική συνάρτηση μετασχηματισμού $f(\cdot)$ τέτοια ώστε η ακόλουθη ανισότητα να ισχύει:

$$\frac{\sum_{k \in I_u \cap I_v} |r_{uk} - f(r_{uk})|}{|I_u \cap I_v|} \leq U$$

όπου το U είναι άλλη μια παράμετρος του αλγορίθμου. Αξίζει να σημειωθεί ότι η απόσταση $\frac{\sum_{k \in I_u \cap I_v} |r_{uk} - f(r_{vk})|}{|I_u \cap I_v|}$ μεταξύ των βαθμολογιών του χρήστη u και των μετασχηματισμένων βαθμολογιών του χρήστη v αποτελεί μια παραλλαγή της απόστασης Manhattan, (Εξίσωση 2.8) πάνω στις

από κοινού προσδιορισμένες βαθμολογίες των δύο χρηστών. Τελικά, κατασκευάζεται ένας κατευθυνόμενος γράφος G, στον οποίο μια ακμή υφίσταται από τον κόμβο u στον v, όταν ο v προβλέπει τον u. Ο γράφος αυτός αναφέρεται επίσης και ως *γράφος προβλεψιμότητας χρήστη-χρήστη (user-user predictability graph)*. Κάθε ακμή του γραφου αντιστοιχεί σε ένα γραμμικό μετασχηματισμό, όπως περιγράφηκε στον παραπάνω ορισμό. Ο γραμμικός αυτός μετασχηματισμός ορίζει μια πρόβλεψη, στην οποία η βαθμολογία στην κεφαλή της ακμής μπορεί να χρησιμοποιηθεί για να προβλέψει τη βαθμολογία στην ουρά της ακμής.

Γράφοι αντικειμένου-αντικειμένου

Για την παραγωγή συστάσεων είναι επίσης δυνατή η κατασκευή ενός *γράφου αντικειμένου - αντικειμένου (item-item graph)*. Ένας τέτοιος γράφος ονομάζεται επίσης *γράφος συσχέτισης (correlation graph)* [Gori07]. Στην περίπτωση αυτή, κατασκευάζεται ένας σταθμισμένος και κατευθυνόμενος γράφος $G = (N, A)$ στον οποίο κάθε κόμβος του N αντιστοιχεί σε ένα αντικείμενο και κάθε ακμή του A αντιστοιχεί σε μια σχέση μεταξύ των αντικειμένων. Επίσης, ένα βάρος w_{ij} αντιστοιχίζεται σε κάθε ακμή (i, j). Εάν τα αντικείμενα i,j έχουν βαθμολογηθεί από τουλάχιστον ένα κοινό χρήστη, τότε υπάρχουν οι κατευθυντικές ακμές (i, j) και (j, i). Αλλιώς, δεν υφίσταται ακμή μεταξύ των κόμβων i και j. Ωστόσο, το κατευθυντικό αυτό δίκτυο είναι ασύμμετρο καθώς το βάρος της ακμής (i, j) δεν είναι απαραίτητα ίδιο με αυτό της ακμής (j, i). Το βάρος κάθε ακμής υπολογίζεται με τον τρόπο που περιγράφεται παρακάτω.

Έστω U_i το σύνολο των χρηστών οι οποίοι έχουν βαθμολογήσει το αντικείμενο i και U_j οι χρήστες οι οποίοι έχουν βαθμολογήσει το αντικείμενο j. Αρχικά, το βάρος w_{ij} κάθε ακμής (i, j) αρχικοποιείται στην τιμή $|U_i \cap U_j|$. Σε αυτό το σημείο τα βάρη είναι συμμετρικά καθώς ισχύει $w_{ij} = w_{ji}$. Στη συνέχεια, τα βάρη των ακμών κανονικοποιούνται έτσι ώστε το άθροισμα των βαρών των εξερχόμενων ακμών ενός κόμβου να ισούται με τη μονάδα. Αυτό πραγματοποιείται διαιρώντας το βάρος w_{ij} με το άθροισμα των ακμών που ξεκινούν από τον κόμβο i. Με τον τρόπο αυτό καταλήγει κάποιος σε ένα γράφο όπου τα βάρη των ακμών αντιστοιχούν σε πιθανότητες τυχαίου περιπάτου. Εδώ αξίζει να σημειωθεί ότι οι τιμές των βαθμολογιών δε χρησιμοποιούνται κατά την κατασκευή του γράφου συσχέτισης. Αυτό δεν είναι πάντα επιθυμητό και γι' αυτό το λόγο πολλές φορές ο γράφος συσχέτισης μπορεί να οριστεί με διαφορετικό τρόπο όπως για παράδειγμα χρησιμοποιώντας τη συνάρτηση συνημιτόνου μεταξύ των διανυσμάτων βαθμολογιών δύο αντικειμένων.

2.2.4 Συνεργατική διήθηση βασισμένη σε μοντέλα

Οι βασισμένες στη μνήμη (ή στη γειτονιά) μέθοδοι συνεργατικής διήθησης, οι οποίες παρουσιάστηκαν στην Ενότητα 2.2.1, μπορούν να θεωρηθούν σαν γενικεύσεις των ταξινομητών k-κοντινότερων γειτόνων, οι οποίοι χρησιμοποιούνται συχνά στη μηχανική μάθηση. Αυτές οι μέθοδοι βασίζονται στα εκάστοτε στιγμιότυπα ενώ δε δημιουργείται κάποιο μοντέλο σε προηγούμενο στάδιο για την πρόβλεψη. Γενικότερα, οι βασισμένες στη γειτονιά μέθοδοι αποτελούν γενικεύσεις *αλγορίθμων μηχανικής μάθησης βασισμένων στα στιγμιότυπα (instance-based learning algorithms)* ή *αλγορίθμων νωχελικής μάθησης (lazy learning algorithms)* λόγω του γεγονότος ότι η προσέγγιση της πρόβλεψης είναι εξειδικευμένη στο στιγμιότυπο για το οποίο γίνεται πρόβλεψη [Agga16]. Για παράδειγμα, στις μεθόδους που βασίζονται στο χρήστη, για *κάθε* χρήστη-στόχο υπολογίζονται οι κοντινοί του χρήστες προκειμένου να πραγματοποιηθεί η πρόβλεψη.

Αντίθετα, στη συνεργατική διήθηση βασισμένη σε μοντέλα (model-based collaborative filtering) δημιουργείται ένα μοντέλο σε πρώτο στάδιο, όπως συμβαίνει στους αλγορίθμους επιβλεπόμενης ή μη επιβλεπόμενης μηχανικής μάθησης. Επομένως, στις μεθόδους αυτές υπάρχει ξεκάθαρος διαχωρισμός μεταξύ των φάσεων εκπαίδευσης (ή κατασκευής του μοντέλου) και πρόβλεψης. Για τη δημιουργία του μοντέλου πρόβλεψης γίνεται χρήση αλγορίθμων μηχανικής μάθησης και εξόρυξης δεδομένων. Σε περιπτώσεις όπου το μοντέλο είναι παραμετροποιημένο, οι παράμετροι του μοντέλου μαθαίνονται στα πλαίσια ενός προβλήματος βελτιστοποίησης. Μερικά παραδείγματα μεθόδων βασισμένων σε μοντέλα περιλαμβάνουν *δέντρα απόφασης (decision trees)*, *κανόνες ταξινόμησης (classification)*,

μπεϋσιανές μεθόδους (bayesian methods), *μοντέλα λανθανουσών μεταβλητών (latent factor models)*, *μηχανές διανυσμάτων υποστήριξης (support vector machines ή SVMs)* και *νευρωνικά δίκτυα (neural networks)* [Agga15]. Όλες αυτές οι μέθοδοι μπορούν να γενικευθούν στο σενάριο της συνεργατικής διήθησης όπως ακριβώς οι ταξινομητές k-κοντινότερων γειτόνων μπορούν να γενικευθούν σε συστήματα συνεργατικής διήθησης βασισμένα στη γειτονιά. Αυτό οφείλεται στο γεγονός ότι τα παραδοσιακά προβλήματα ταξινόμησης και παλινδρόμησης αποτελούν ειδικές περιπτώσεις του προβλήματος συμπλήρωσης πίνακα (ή συνεργατικής διήθησης).

Σε ένα πρόβλημα ταξινόμησης δεδομένων, έστω ένας $m \times n$ πίνακας σαν αυτόν του Σχήματος 2.2a, στον οποίο οι πρώτες (n-1) στήλες αποτελούν τις χαρακτηριστικές μεταβλητές (ανεξάρτητες) ενώ η n-οστή στήλη τη μεταβλητή τάξης (εξαρτημένη). Οι εγγραφές των πρώτων (n-1) στηλών είναι πλήρως προσδιορισμένες ενώ μόνο ένα υποσύνολο των εγγραφών της τελευταίας στήλης είναι απροσδιόριστο. Επομένως, ένα υποσύνολο των γραμμών του πίνακα είναι πλήρως προσδιορισμένο, αποτελώντας τα *δεδομένα εκπαίδευσης (training data)*, ενώ οι υπόλοιπες στις οποίες εμφανίζονται οι απροσδιόριστες εγγραφές αποτελούν τα *δεδομένα επαλήθευσης (test data)*.

Σε αντίθεση με το πρόβλημα της ταξινόμησης, στη συνεργατική διήθηση (ή συμπλήρωση πίνακα) οποιαδήποτε εγγραφή μπορεί να είναι απροσδιόριστη, όπως φαίνεται στο Σχήμα 2.2b. Επομένως είναι ξεκάθαρο ότι το πρόβλημα της ταξινόμησης αποτελεί υποκατηγορία του προβλήματος συμπλήρωσης πίνακα. Οι βασικές διαφορές μεταξύ των δύο αυτών προβλημάτων συνοψίζονται παρακάτω:

1. Στο πρόβλημα της ταξινόμησης, υπάρχει σαφής διαχωρισμός μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών. Αυτό αντίθετα δεν ισχύει στο πρόβλημα συμπλήρωσης πίνακα, όπου κάθε στήλη του πίνακα είναι ταυτόχρονα ανεξάρτητη και εξαρτημένη μεταβλητή, ανάλογα με το ποιες εγγραφές εξετάζονται για την πρόβλεψη κάθε φορά.

2. Στην ταξινόμηση υπάρχει διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης (training) και σε δοκιμαστικά (test). Στη συμπλήρωση πίνακα δεν υφίσταται αυτός ο διαχωρισμός μεταξύ των γραμμών του πίνακα, παρά μόνο η θεώρηση ότι οι προσδιορισμένες εγγραφές αποτελούν τα δεδομένα εκπαίδευσης και οι απροσδιόριστες τα δεδομένα επαλήθευσης.

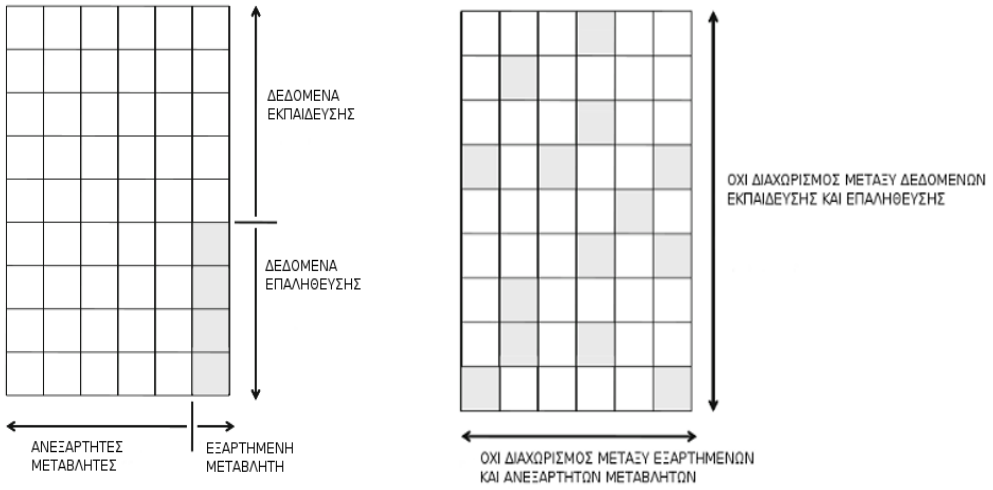
3. Σε ένα πρόβλημα ταξινόμησης οι στήλες του πίνακα αντιπροσωπεύουν χαρακτηριστικά ενώ οι γραμμές διαφορετικά παραδείγματα δεδομένων, ενώ ένα πρόβλημα συνεργατικής διήθησης μπορεί να προσεγγιστεί ως προς τις γραμμές ή ως προς τις στήλες του πίνακα βαθμολογιών για μοντέλο βασισμένο στο χρήστη ή στο αντικείμενο.

Μοντέλα Λανθανουσών Μεταβλητών

Τα *μοντέλα λανθανουσών μεταβλητών (latent factor models)* θεωρούνται τεχνολογία αιχμής (state-of-the-art) όσον αφορά τα συστήματα συστάσεων, αφού έχει αποδειχθεί ότι υπερτερούν σε αρκετούς τομείς έναντι των υπολοίπων τύπων συστημάτων. Τα μοντέλα αυτά κάνουν χρήση γνωστών τεχνικών μείωσης διαστάσεων (dimensionality reduction) για την εύρεση των ζητούμενων εγγραφών. Οι τεχνικές αυτές χρησιμοποιούνται ευρέως σε πολλές περιοχές της ανάλυσης δεδομένων για την αναπαράσταση των υπό εξέταση δεδομένων σε μικρότερο αριθμό διαστάσεων.

Η βασική ιδέα των τεχνικών μείωσης διαστάσεων είναι η περιστροφή του συστήματος αξόνων, έτσι ώστε οι συσχετίσεις μεταξύ των διαστάσεων ανά δύο να εξαλείφονται. Το βασικό πλεονέκτημα της μείωσης διαστάσεων έγκειται στο γεγονός ότι έχει τη δυνατότητα να μετατρέπει έναν όχι πλήρως προσδιορισμένο πίνακα δεδομένων σε μια μειωμένη, περιστραμμένη αλλά πλήρως προσδιορισμένη αναπαράστασή του. Από τη στιγμή που έχει εξασφαλιστεί η μειωμένη, πλήρως προσδιορισμένη αναπαράσταση, στη συνέχεια αυτή μπορεί να περιστραφεί πίσω στο αρχικό σύστημα αξόνων για να ανακατασκευαστεί ένας πλήρως προσδιορισμένος πίνακας δεδομένων [Agga01].

Με λίγα λόγια, οι τεχνικές μείωσης διαστάσεων εκμεταλλεύονται τις συσχετίσεις μεταξύ των γραμμών και στηλών του πίνακα προκειμένου να κατασκευάσουν μια μειωμένη και πλήρως προσδιορισμένη αναπαράστασή του. Η χρήση των συσχετίσεων αυτών είναι ούτως ή άλλως θεμελιώδης για



(a) Ταξινόμηση **(b)** Συνεργατική Διήθηση

Σχήμα 2.2: Σύνοψη διαφορών μεταξύ συνεργατικής διήθησης και ταξινόμησης

όλες τις μεθόδους συνεργατικής διήθησης, είτε αυτές βασίζονται στη μνήμη είτε σε μοντέλα. Για παράδειγμα, οι μέθοδοι που βασίζονται στο χρήστη εκμεταλλεύονται τις συσχετίσεις σε επίπεδο χρήστη ενώ αυτές που βασίζονται στο αντικείμενο κάνουν χρήση των συσχετίσεων σε επίπεδο αντικειμένου.

Οι τεχνικές *παραγοντοποίησης πίνακα (matrix factorization)* έχουν τον τρόπο να αξιοποιούν όλες τις συσχετίσεις μεταξύ των γραμμών και των στηλών σε ένα μόνο βήμα για την εκτίμηση ολόκληρου του πίνακα δεδομένων. Το συγκεκριμένο πλεονέκτημα της προσέγγισης αυτής αποτελεί και το λόγο για τον οποίο τα μοντέλα λανθανουσών μεταβλητών είναι από τα πιο αποτελεσματικά συστήματα συνεργατικής διήθησης.

Γεωμετρική διαίσθηση των μοντέλων λανθανουσών μεταβλητών

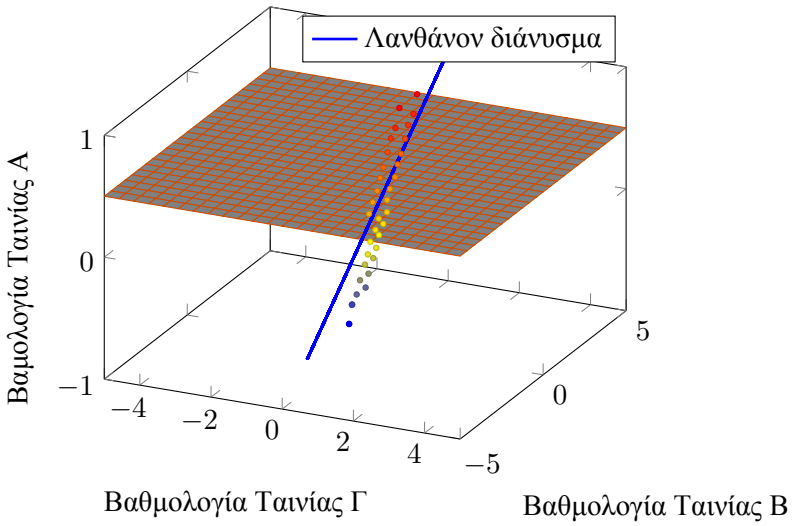
Προκειμένου να γίνουν κατανοητά αυτά που αναφέρθηκαν προηγουμένως, θεωρούμε έναν πίνακα βαθμολογιών με τρεις στήλες, που αντιπροσωπεύουν αντικείμενα, στον οποίο και τα τρία αντικείμενα είναι θετικά συσχετισμένα μεταξύ τους. Υποθέτουμε ότι τα αντικείμενα αποτελούν ταινίες ενώ για ευκολία στην οπτικοποίηση θεωρούμε ότι οι βαθμολογίες είναι συνεχείς τιμές που κυμαίνονται στο πεδίο τιμών το [-1,1].

Εφόσον λοιπόν οι βαθμολογίες των τριών αντικειμένων είναι θετικά συσχετισμένες, τότε το τρισδιάστατο διάγραμμα διασποράς μπορεί χονδρικά να προβληθεί πάνω σε μια μονοδιάστατη γραμμή όπως φαίνεται στο Σχήμα 2.3. Από τη στιγμή που οι βαθμολογίες τοποθετούνται κυρίως σε μια μονοδιάστατη γραμμή, αυτό σημαίνει ότι ο αρχικός πίνακας βαθμολογιών έχει κατά προσέγγιση τάξη ίση με 1, μετά την αφαίρεση των θορυβωδών δεδομένων. Αυτή η προσέγγιση τάξης-1 είναι το *λανθάνον διάνυσμα (latent vector)* το οποίο διέρχεται από το κέντρο των δεδομένων, όπως φαίνεται στο Σχήμα 2.3. Οι τεχνικές μείωσης διαστάσεων, όπως είναι η *Ανάλυση Πρωτευουσών Συνιστωσών (Principal Component Analysis ή PCA)* και η *Ανάλυση Πίνακα σε Ιδιάζουσες Τιμές (Singular Value Decomposition η SVD)*, χρησιμοποιούν την προβολή των δεδομένων πάνω σε τέτοια διανύσματα σαν προσέγγιση.

Επομένως, όταν ο $m \times n$ πίνακας βαθμολογιών έχει τάξη $p << \min\{m, n\}$ (μετά την αφαίρεση των θορυβωδών παρεκκλίσεων), τότε τα δεδομένα μπορούν προσεγγιστικά να αναπαρασταθούν σε ένα p-διάστατο υπερεπίπεδο. Αυτό αυτομάτως σημαίνει ότι οι άγνωστες βαθμολογίες ενός χρήστη μπορούν να εκτιμηθούν με μόνο p προσδιορισμένες εγγραφές, από τη στιγμή που το p-διάστατο υπερεπίπεδο είναι γνωστό. Για παράδειγμα, στην περίπτωση του Σχήματος 2.3, μόνο η βαθμολογία μιας από τις τρεις ταινίες χρειάζεται να είναι γνωστή προκειμένου να μπορούν να εκτιμηθούν οι υπόλοιπες δύο επειδή η τάξη του πίνακα βαθμολογιών είναι ίση με 1 μετά την αφαίρεση του θορύβου. Έτσι, εάν

η βαθμολογία της ταινίας Α ισούται με 0.5, τότε οι βαθμολογίες των ταινιών Β και Γ μπορούν να εκτιμηθούν από την τομή του μονοδιάστατου λανθάνοντος διανύσματος με το παράλληλο στους άξονες υπερεπίπεδο στο οποίο η βαθμολογία της ταινίας Α είναι σταθερή και ίση με 0.5. Το υπερεπίπεδο αυτό καθώς και το λανθάνον διάνυσμα φαίνονται ξεκάθαρα στο Σχήμα 2.3.

Στην παραπάνω περίπτωση, υποθέσαμε ότι ο πλήρως προσδιορισμένος $m \times n$ πίνακας βαθμολογιών ήταν διαθέσιμος προκειμένου να εκτιμηθεί το λανθάνον διάνυσμα. Ωστόσο, πρακτικά ο πίνακας βαθμολογιών δε χρειάζεται να είναι πλήρως προσδιορισμένος ώστε να εκτιμηθούν τα κυρίαρχα λανθάνοντα διανύσματα. Για την ακρίβεια, η ικανότητα να μπορούν να εκτιμηθούν τα διανύσματα αυτά παρά τις ελλιπείς εγγραφές είναι το κλειδί της επιτυχίας αυτών των μοντέλων. Η βασική ιδέα σε όλες αυτές τις μεθόδους είναι η εύρεση ενός συνόλου διανυσμάτων στα οποία η μέση τετραγωνική απόσταση του υπερεπιπέδου το οποίο ορίζουν αυτά από τα σημεία των δεδομένων (τα οποία αναπαριστούν βαθμολογίες χρηστών) να είναι η μικρότερη δυνατή. Επομένως, χρησιμοποιείται ένα μερικώς προσδιορισμένο σύνολο δεδομένων προκειμένου να βρεθεί το χαμηλής τάξης υπερεπίπεδο στο οποίο προσεγγιστικά κυμαίνονται τα δεδομένα.



Σχήμα 2.3: Πρόβλεψη των βαθμολογιών ενός χρήστη με τη βοήθεια του λανθάνοντος διανύσματος

Αλγεβρική επεξήγηση των μοντέλων λανθανουσών μεταβλητών

Ένας άλλος τρόπος να γίνει κατανοητή η αποτελεσματικότητα των μοντέλων λανθανουσών μεταβλητών είναι εξετάζοντας, από τη σκοπιά της γραμμικής άλγεβρας, το ρόλο που παίζει η παραγοντοποίηση σε τέτοιου είδους πίνακες. Η παραγοντοποίηση, γενικά, αποτελεί έναν πιο γενικό τρόπο προσέγγισης ενός πίνακα όταν αυτός είναι επιρρεπής στη μείωση διάστασης εξαιτίας συσχετίσεων μεταξύ των στηλών (ή γραμμών) του. Οι περισσότερες τεχνικές μείωσης διαστάσεων μπορούν να εκφραστούν ως παραγοντοποιήσεις πινάκων.

Αρχικά, ας υποτεθεί η απλή περίπτωση κατά την οποία τα στοιχεία του πίνακα R είναι προσδιορισμένα. Η βασική ιδέα είναι ότι οποιοσδήποτε $m \times n$ πίνακας R με τάξη $k << \min\{m, n\}$ μπορεί πάντα να εκφραστεί στην παρακάτω μορφή σαν γινόμενο παραγόντων τάξης-k:

$$R = UV^T \tag{2.20}$$

Στον παραπάνω τύπο, ο U είναι ένας $m \times k$ πίνακας, ενώ ο V είναι $n \times k$. Σημειώνεται ότι η τάξη του χώρου στηλών (column space) αλλά και γραμμών (row space) του πίνακα R είναι ίση με k. Κάθε στήλη του U μπορεί να θεωρηθεί ως ένα από τα k διανύσματα βάσης του k-διάστατου χώρου του R και η j-οστή γραμμή του V περιέχει τους αντίστοιχους συντελεστές, οι οποίοι συνδυάζουν αυτά τα διανύσματα βάσης για να παραχθεί η j-οστή στήλη του πίνακα R. Αντίστοιχα, κάποιος μπορεί

να θεωρήσει τις στήλες του V σαν διανύσματα βάσης του χώρου γραμμών του πίνακα R και τις γραμμές του U ως συντελεστές αυτών. Η ικανότητα παραγοντοποίησης κάθε πίνακα τάξης k στη μορφή αυτή αποτελεί θεμελιώδες θεώρημα της γραμμικής άλγεβρας καθώς και το γεγονός ότι υπάρχει άπειρος αριθμός τέτοιων παργοντοποιήσεων που αντιστοιχούν στα διάφορα σύνολα διανυσμάτων βάσης [Stra03]. Ο αλγόριθμος SVD αποτελεί ένα παράδειγμα τέτοιου είδους παραγοντοποίησης στην οποία τα διανύσματα βάσης που αναπαρίστανται από τις στήλες του U (και αυτές του V) είναι ανά δύο ορθογώνια μεταξύ τους.

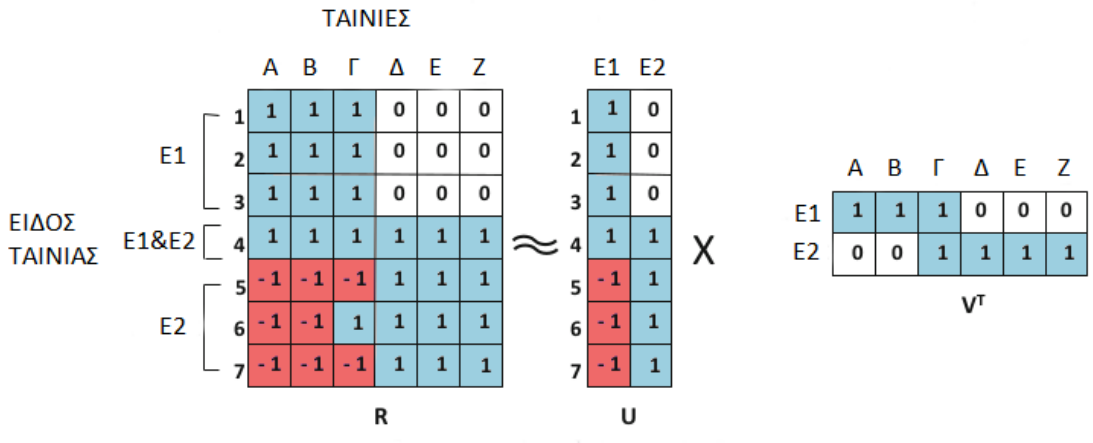
Ακόμα και στην περίπτωση που ο πίνακας R έχει τάξη μεγαλύτερη από k, μπορεί συνήθως να εκφραστεί προσεγγιστικά σαν γινόμενο παραγόντων k-τάξης:

$$R \approx UV^T \tag{2.21}$$

Όπως και πριν, ο U είναι ένας $m \times k$ πίνακας και ο V ένας $n \times k$. Το σφάλμα αυτής της προσέγγισης ισούται με $||R - UV^T||^2$, όπου το $||\cdot||^2$ παριστάνει το άθροισμα των τετραγώνων των εγγραφών του πίνακα υπολοίπων (residual matrix) $(R - UV^T)$. Η ποσότητα αυτή αναφέρεται επίσης σαν τετραγωνισμένη *νόρμα Frobenius* του πίνακα υπολοίπων. Ο πίνακας υπολοίπων τυπικά αναπαριστά το θόρυβο των βαθμολογιών του πίνακα R οι οποίες δεν μπορούν να μοντελοποιηθούν από τους παράγοντες χαμηλής τάξης.

Βασικές αρχές της παραγοντοποίησης πίνακα

Στο βασικό μοντέλο παραγοντοποίησης πίνακα, ο $m \times n$ πίνακας R παραγοντοποιείται προσεγγιστικά σε έναν $m \times k$ πίνακα U και έναν $n \times k$ πίνακα V σύμφωνα με την Εξίσωση 2.21. Κάθε στήλη του U (ή V) αναφέρεται ως *λανθάνον διάνυσμα (latent vector)* ή *λανθάνουσα συνιστώσα (latent component)*, ενώ κάθε γραμμή του U (ή V) αναφέρεται ως *λανθάνων παράγοντας (latent factor)*. Η i-οστή γραμμή $\overline{u_i}$ του U αναφέρεται σαν *παράγοντας χρήστη (user factor)* και περιέχει k εγγραφές που αντιστοιχούν στην οικειότητα του χρήστη i με τις k έννοιες του πίνακα βαθμολογιών R. Για παράδειγμα, στην περίπτωση του Σχήματος 2.4 το $\overline{u_i}$ είναι ένα δισδιάστατο διάνυσμα που δείχνει την οικειότητα του χρήστη i με τις ταινίες είδους E1 και E2 που προκύπτουν από τον πίνακα βαθμολογιών. Με παρόμοιο τρόπο, κάθε γραμμή $\overline{v_i}$ του V αποτελεί έναν παράγοντα αντικειμένου (item factor) και αντιπροσωπεύει την οικειότητα του i-οστού αντικειμένου με αυτές τις k έννοιες. Στο παράδειγμα του Σχήματος 2.4 ο παράγοντας αντικειμένου περιέχει την οικειότητα του αντικειμένου (ταινίας) ως προς τα δύο προαναφερθέντα είδη ταινιών.



Σχήμα 2.4: Παράδειγμα παραγοντοποίησης πίνακα 2ης τάξης

Από την εξίσωση 2.21 συνεπάγεται ότι κάθε βαθμολογία r_{ij} του πίνακα R μπορεί να εκφραστεί προσεγγιστικά σαν το εσωτερικό γινόμενο του i-οστού παράγοντα χρήστη με τον j-οστό παράγοντα αντικειμένου ως εξής:

$$r_{ij} \approx \overline{u_i} \cdot \overline{v_j} \tag{2.22}$$

Από τη στιγμή που οι λανθάνοντες παράγοντες $\overline{u_i} = (u_{i1}..u_{ik})$ και $\overline{v_i} = (v_{i1}..v_{ik})$ μπορούν να θεωρηθούν ως οι οικειότητες των χρηστών και αντικειμένων αντίστοιχα ως προς τις k διαφορετικές έννοιες, ένας διαισθητικός τρόπος έκφρασης της Εξίσωσης 2.21 είναι ο παρακάτω:

$$\begin{aligned} r_{ij} &\approx \sum_{s=1}^{k} u_{is} \cdot v_{js} \\ &= \sum_{s=1}^{k} (\text{Οικειότητα χρήστη i με την έννοια s}) \times (\text{Οικειότητα αντικειμένου j με την έννοια s}) \end{aligned} \tag{2.23}$$

Στην περίπτωση του παραδείγματος του Σχήματος 2.4, όπου οι δύο έννοιες (k=2) του παραπάνω αθροίσματος αντιστοιχούν στα είδη ταινιών E1 και E2, ο τύπος μεταφράζεται στη μορφή:

$$\begin{aligned} r_{ij} \approx\ & (\text{Οικειότητα χρήστη i με τις ταινίες είδους E1}) \times (\text{Οικειότητα ταινίας j με το είδος E1}) \\ & + (\text{Οικειότητα χρήστη i με τις ταινίες είδους E2}) \times (\text{Οικειότητα ταινίας j με είδος E2}) \end{aligned} \tag{2.24}$$

Εδώ αξίζει να σημειωθεί ότι οι έννοιες που αναφέρθηκαν προηγουμένως, όπως για παράδειγμα οι δύο κατηγορίες ταινιών του παραδείγματος, σπάνια μπορούν να ερμηνευθούν με κάποιο διαισθητικό τρόπο. Το λανθάνον διάνυσμα είναι συνήθως ένα αυθαίρετο διάνυσμα με θετικές και αρνητικές τιμές για το οποίο δύσκολα μπορεί να δοθεί μια σημασιολογική ερμηνεία. Ωστόσο, αυτό που είναι βέβαιο είναι ότι το διάνυσμα αυτό αναπαριστά ένα κυρίαρχο πρότυπο συσχέτισης μεταξύ των γραμμών και στηλών του πίνακα βαθμολογιών.

Υπάρχουν διάφορες μέθοδοι παραγοντοποίησης πίνακα, οι βασικές διαφορές μεταξύ των οποίων έγκεινται πρώτον στους περιορισμούς που επιβάλλονται στους πίνακες U και V (όπως η ορθογωνιότητα ή η μη-αρνητικότητα των λανθανόντων διανυσμάτων) και δεύτερον στη φύση της συνάρτησης βελτιστοποίησης (όπως η ελαχιστοποίηση της νόρμας Frobenius ή μεγιστοποίησης της πιθανοφάνειας). Επίσης, κάποιοι τύποι παραγοντοποίησης, όπως η μη-αρνητική παραγοντοποίηση πίνακα είναι ειδικά σχεδιασμένες προκειμένου να επιτυγχάνουν μεγαλύτερη ερμηνευτικότητα όσον αφορά τα λανθάνοντα διανύσματα. Ωστόσο, αυτές οι μέθοδοι ξεφεύγουν από την εμβέλεια της παρούσας διπλωματικής και για αυτό το λόγο δεν αναλύονται περαιτέρω.

2.2.5 Πλεονεκτήματα και μειονεκτήματα των μεθόδων βασισμένων στη μνήμη και βασισμένων σε μοντέλα

Οι μέθοδοι οι οποίες βασίζονται στη μνήμη παρουσιάζουν ορισμένα πλεονεκτήματα που σχετίζονται με την απλότητα και την εύκολα κατανοητή προσέγγισή τους. Ανάμεσα στα πλεονεκτήματα αυτά βρίσκονται τα παρακάτω:

- ✓ Είναι εύκολα στην υλοποίηση και στην αποσφαλμάτωσή τους.
- ✓ Μπορεί εύκολα να δικαιολογηθεί γιατί προτείνεται ένα συγκεκριμένο αντικείμενο, όπως π.χ. με την αναφορά των γειτονικών χρηστών.
- ✓ Οι συστάσεις των συστημάτων αυτών παραμένουν σχετικά σταθερές με την προσθήκη νέων αντικειμένων και χρηστών.

Από την άλλη μεριά, οι μέθοδοι αυτοί παρουσιάζουν σημαντικά μειονεκτήματα που έχουν να κάνουν κυρίως με την αποτελεσματικότητα τους όσον αφορά το χρόνο και τη μνήμη αλλά και την αραιότητα του πίνακα βαθμολογιών. Εξαιτίας των παραπάνω τα συστήματα αυτά:

- ✗ Πιθανόν να είναι μη πρακτικά σε συνθήκες μεγάλης κλίμακας. Η offline φάση σε συστήματα που βασίζονται στο χρήστη απαιτεί τουλάχιστον $\mathcal{O}(m^2)$ χρόνο και μνήμη, γεγονός που μπορεί να τα κάνει πολύ αργά ή απαιτητικά από άποψη μνήμης όταν το m είναι της τάξης του εκατομμυρίου.
- ✗ Παρουσιάζουν περιορισμένη κάλυψη (coverage) εξαιτίας της αραιότητας του πίνακα βαθμολογιών. Για παράδειγμα, όταν κανένας γειτονικός χρήστης του χρήστη-στόχου u δεν έχει βαθμολογήσει το αντικείμενο j τότε δεν μπορεί να παραχθεί πρόβλεψη.

Όσον αφορά τα συστήματα που βασίζονται σε μοντέλα, αυτά παρουσιάζουν κάποια σημαντικά πλεονεκτήματα έναντι των βασισμένων στη μνήμη συστημάτων όπως:

- ✓ Είναι αποτελεσματικά όσον αφορά το χώρο, αφού το μέγεθος του μοντέλου είναι πολύ μικρότερο από τον αρχικό πίνακα βαθμολογιών, σε αντίθεση με τα συστήματα που βασίζονται στη μνήμη τα οποία μπορεί να απαιτούν $\mathcal{O}(m^2)$ και $\mathcal{O}(n^2)$ χωρική πολυπλοκότητα για μοντέλο βασισμένο στο χρήστη και στο αντικείμενο αντίστοιχα.
- ✓ Είναι πιο γρήγορα κατά τη φάση της εκπαίδευσης και της πρόβλεψης. Κατά τη φάση της προεπεξεργασίας η χρονική πολυπλοκότητα των συστημάτων βασισμένων στη μνήμη είναι τετραγωνικής τάξης ενώ ένα μοντέλο συνήθως είναι πολύ πιο γρήγορο στο να κατασκευαστεί.
- ✓ Αντιμετωπίζουν πιο αποτελεσματικά την *υπερπροσαρμογή (overfitting)*. Η υπερπροσαρμογή συναντάται σε πολλούς αλγορίθμους μηχανικής μάθησης όταν η πρόβλεψη επηρεάζεται σημαντικά από τη μορφή του συνόλου δεδομένων. Η συνοπτική προσέγγιση των μεθόδων που βασίζονται σε μοντέλα καθώς και η κανονικοποίηση μπορεί να βοηθήσει σημαντικά ώστε να είναι τα μοντέλα πιο εύρωστα.

Επομένως, παρότι οι μέθοδοι που βασίζονται στη γειτονία ήταν από τις πρώτες και πιο δημοφιλείς τεχνικές συνεργατικής διήθησης χάρη στην απλότητά τους, δεν είναι απαραίτητα οι πιο ακριβείς μέθοδοι που χρησιμοποιούνται σήμερα. Για του λόγου το αληθές, ορισμένες από τις μεθόδους με την υψηλότερη ακρίβεια βασίζονται σε μοντέλα γενικά και πιο συγκεκριμένα σε μοντέλα λανθανουσών μεταβλητών [Agga16].

Κεφάλαιο 3

Κοινωνικά Δίκτυα

3.1 Ανάλυση κοινωνικών δικτύων

Η *ανάλυση κοινωνικών δικτύων (social network analysis ή SNA)* ορίζεται ως η διαδικασία εξερεύνησης κοινωνικών δομών μέσω της χρήσης των θεωριών δικτύων και γράφων [Otte02]. Η ανάλυση κοινωνικών δικτύων έχει εξελιχθεί σε μια μεθοδολογία κλειδί στη σύγχρονη κοινωνιολογία, παίζοντας επίσης πολύ σημαντικό ρόλο σε άλλους κλάδους όπως η ανθρωπολογία, η βιολογία, τα οικονομικά, η γεωγραφία, η πληροφορική, οι πολιτικές επιστήμες κ.α.

Η ανάλυση κοινωνικών δικτύων έχει τις θεωρητικές ρίζες της στη δουλειά κοινωνιολόγων, οι οποίοι έγραψαν για τη μεγάλη σημασία της μελέτης των προτύπων των σχέσεων μεταξύ κοινωνικών δραστών. Οι κοινωνικοί επιστήμονες χρησιμοποίησαν τον όρο "Κοινωνικά Δίκτυα" από τις αρχές τους 20ού αιώνα για να χαρακτηρίσουν πολύπλοκα σύνολα σχέσεων μεταξύ μελών κοινωνικών συστημάτων σε όλα τα επίπεδα, από διαπροσωπικές μέχρι διεθνείς. Τη δεκαετία του 1930 παρουσιάστηκαν οι πρώτες αναλυτικές μέθοδοι εξέτασης κοινωνικών δικτύων από τους Jacob Moreno και Helen Jennings [Free04]. Το 1954, ο John Arundel Barnes, ήταν ο πρώτος που χρησιμοποίησε τον όρο σε επιστημονικό πλαίσιο [Barn54] για να δηλώσει πρότυπα δεσμών, περικλείοντας όρους οι οποίοι χρησιμοποιούνταν στην καθημερινότητα αλλά και όρους που χρησιμοποιούνταν από κοινωνικούς επιστήμονες: οριοθετημένες ομάδες (π.χ. οικογένειες, φυλές) και κοινωνικές κατηγορίες (π.χ. γένος, εθνικότητα). Τις επόμενες δεκαετίες, πολλοί ερευνητές επέκτειναν το πεδίο της ανάλυσης κοινωνικών δικτύων εισάγοντας νέους όρους και μετρικές οι οποίες χρησιμοποιούνται και ερευνώνται ακόμα και σήμερα, τόσο σε θεωρητικό όσο και σε πρακτικό επίπεδο.

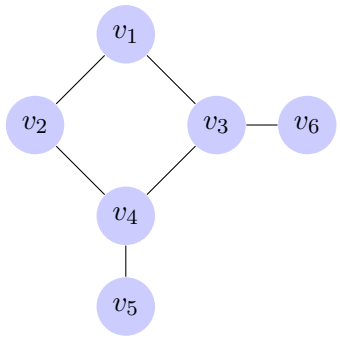
3.1.1 Βασικές έννοιες θεωρίας γράφων

Ο πιο συνηθισμένος τρόπος αναπαράστασης κοινωνικών δικτύων είναι με τη μορφή γράφων και γι' αυτό το λόγο η ανάλυσή τους βασίζεται σε μεγάλο βαθμό στη θεωρία γράφων. Επομένως, η εμβάθυνση στην ανάλυση κοινωνικών δικτύων προϋποθέτει την κατανόηση βασικών εννοιών της θεωρίας γράφων οι οποίες παρουσιάζονται παρακάτω. Οι έννοιες αυτές χρησιμοποιούνται εκτεταμένα στη συνέχεια του Κεφαλαίου 3 αλλά και στην περιγραφή της πειραματικής διαδικασίας που ακολουθεί στο Κεφάλαιο 4.

Ορισμός του γράφου

Γράφος $G(V, E)$ ονομάζεται ένα σύνολο από κορυφές (ή κόμβους) $v_1, v_2, ...v_n \in V$ οι οποίες ενώνονται μεταξύ τους με ακμές $e_1, e_2, ...e_m \in E$ και ορίζεται από τον τρόπο με τον οποίο συνδέονται οι κορυφές (κόμβοι). Αν οι ακμές προσανατολίζονται οριζόμενες από διατεταγμένα ζεύγη κόμβων, τότε ο γράφος αποκαλείται *κατευθυντικός (directed)*. Αν οι ακμές δεν προσανατολίζονται, οριζόμενες απλώς από διμελή σύνολα και όχι διατεταγμένα ζεύγη, τότε αποκαλείται *μη-κατευθυντικός (undirected)*. Στο Σχήμα 3.1a βλέπουμε το παράδειγμα ενός μη-κατευθυντικού γράφου 6 κόμβων 6 ακμών.

Επιπλέον στοιχεία για τον ορισμό ενός γράφου είναι η σύνδεση των ακμών του με κάποια αξία, οπότε αποκαλείται *σταθμισμένος (weighted)*. Αναφορικά με το πλήθος των ακμών του, ένας γράφος



(a) Γράφημα

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

(b) Πίνακας γειτνίασης

Σχήμα 3.1: Γράφημα και ο αντίστοιχος πίνακας γειτνίασης

αποκαλείται *πλήρης (complete)* όταν για κάθε ζεύγος κορυφών u, v υπάρχει ακμή $u-v$ που να συνδέει τις δυο κορυφές. Επίσης όταν ένας γράφος περιέχει μικρο αριθμό ακμών αποκαλείται *αραιός (sparse)* ενώ όταν αντίθετα έχει μεγάλο αριθμό ακμών αποκαλείται *πυκνος (dense)*.

Ο πίνακας γειτνίασης

Ο πίνακας γειτνίασης $A = [a_{ij}]$ ενός γράφου $G(V, E)$ με n κόμβους είναι ένας πίνακας $n \times n$ του οποίου τα στοιχεία a_{ij} υποδεικνύουν αν οι κόμβοι i και j συνδέονται μεταξύ τους. Πιο συγκεκριμένα, εάν ο κόμβος i συνδέεται με τον κόμβο j τότε $a_{ij} = 1$, αλλιώς $a_{ij} = 0$

Στην περίπτωση ενός πεπερασμένου απλού γράφου, ο πίνακας γειτνίασης είναι ένας πίνακας που αποτελείται αποκλειστικά από (0,1) με μηδενική διαγώνιο ($a_{ii} = 0$). Αν ο γράφος είναι μη-κατευθυντικός, τότε ο πίνακας γειτνίασης είναι συμμετρικός ($a_{ij} = a_{ji}$). Στο Σχήμα 3.1b φαίνεται ένα παράδειγμα πίνακα γειτνίασης ο οποίος αντιστοιχεί στο γράφο του Σχήματος 3.1a. Στο συγκεκριμένο παράδειγμα ο πίνακας γειτνίασης είναι συμμετρικός αφού ο γράφος είναι μη κατευθυνόμενος.

Η έννοια του βαθμού

Για ένα μη-κατευθυντικό γράφο, ο *βαθμός (degree)* μιας κορυφής είναι ο αριθμός των προσκείμενων σε αυτή ακμών. Ο βαθμός μιας κορυφής v συμβολίζεται με $\deg(v)$ ή $\deg v$. Ο *μέγιστος βαθμός (maximum degree)* ενός γράφου G, συμβολιζόμενος με $\Delta(G)$ και ο *ελάχιστος βαθμός (minimum degree)* ενός γράφου, συμβολιζόμενος με $\delta(G)$, είναι ο μέγιστος και ελάχιστος βαθμός των κορυφών του αντίστοιχα. Σε έναν *κανονικό γράφο (regular graph)* οι βαθμοί όλων των κορυφών είναι ίδιοι και επομένως μιλάμε για το βαθμό του γράφου. Για το άθροισμα των βαθμών ενός μη-κατευθυντικού γράφου ισχύει η εξίσωση:

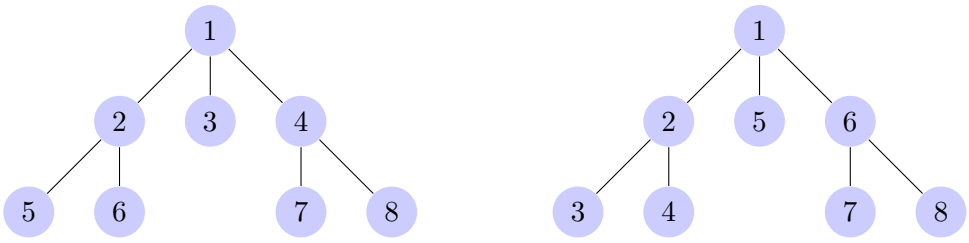
$$\sum_{v \in V} \deg^+(v) = 2|E| \tag{3.1}$$

Στην περίπτωση των κατευθυντικών γράφων, ορίζεται ο *εσωτερικός βαθμός (indegree)* και ο *εξωτερικό βαθμός (outdegree)* μιας κορυφής, που είναι ο αριθμός των ακμών που καταλήγουν και ο αριθμός των ακμών που ξεκινούν από την κορυφή αυτή αντίστοιχα. Για ένα γράφο $G(V, E)$ και μια κορυφή $v \in V$ ο εσωτερικός βαθμός της κορυφής v συμβολίζεται με $\deg^-(v)$ και ο εξωτερικός βαθμός της με $\deg^+(v)$. Επίσης για το άθροισμα των βαθμών ενός κατευθυντικού γράφου ισχύει:

$$\sum_{v \in V} \deg^-(v) = \sum_{v \in V} \deg^+(v) = |E| \tag{3.2}$$

Η έννοια της συνεκτικότητας

Ένας μη-κατευθυντικός γράφος $G(V, E)$ είναι *συνεκτικός (connected)* αν για κάθε ζεύγος κορυφών u, v υπάρχει μονοπάτι που να συνδέει τις κορυφές $u-v$. Ένας γράφος ο οποίος δεν είναι συνεκτι-



(a) Αναζήτηση κατά πλάτος **(b)** Αναζήτηση κατά βάθος

Σχήμα 3.2: Διάσχιση γράφου με αναζήτηση κατά πλάτος και αναζήτηση κατά βάθος

κός ονομάζεται *μη-συνεκτικός (disconnected)*. Δηλαδή, ένας γράφος $G(V, E)$ λέγεται μη-συνεκτικός αν υπάρχει ζεύγος κορυφών u, v τέτοιο ώστε να μην υπάρχει μονοπάτι στον G που να ξεκινάει από τη μια κορυφή και να καταλήγει στην άλλη. Σε ένα μη-κατευθυντικό γράφο *συνεκτική συνιστώσα (connected component)* ονομάζεται ένας *συνεκτικός* υπογράφος μεγιστοτικός ως προς τον αριθμό των κορυφών του.

Ένας κατευθυντικός γράφος ονομάζεται *ασθενώς συνεκτικός (weakly connected)* εάν αντικαθιστώντας όλες τις κατευθυντικές ακμές του με μη-κατευθυντικές παράγεται ένας (μη-κατευθυντικός) συνεκτικός γράφος. Ονομάζεται *συνεκτικός (connected)* όταν για κάθε ζεύγος κορυφών u, v υπάρχει ένα κατευθυντικό μονοπάτι από την u στην v *ή* ένα κατευθυντικό μονοπάτι από την v στην u. Ο γράφος είναι ισχυρά συνεκτικός όταν για κάθε ζεύγος κορυφών u, v υπάρχει ένα κατευθυντικό μονοπάτι από την u στην v *και* ένα κατευθυνόμενο μονοπάτι από την v στην u. Επίσης *ισχυρές συνιστώσες (strong components)* ενός κατευθυντικού γράφου ονομάζονται οι μέγιστοι ισχυρά συνεκτικοί υπογράφοι αυτού.

Γέφυρα (bridge) ή *ακμή τομής (edge cut)* ενός γράφου $G(V, E)$ ονομάζεται μια ακμή e η οποία όταν αφαιρεθεί αυξάνει το πλήθος των συνεκτικών συνιστωσών του γράφου. Αντίστοιχα, *σημείο άρθρωσης (joint)* ή *κορυφή τομής (vertex cut)* ονομάζεται μια κορυφή η οποία όταν αφαιρεθεί αυξάνει το πλήθος των συνεκτικών συνιστωσών του γράφου.

Διάσχιση γράφων

Η *διάσχιση γράφων (graph traversal)* αναφέρεται στη διαδικασία επίσκεψης (ελέγχου ή/και ανανέωσης) κάθε κόμβου ενός γράφου. Η διάσχιση γράφων χρησιμοποιείται κυρίως για την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τις ιδιότητες και με τη δομή ενός γράφου όπως για παράδειγμα τη συνεκτικότητα, την εύρεση σημείων τομής, την εύρεση συντομότερων διαδρομών κ.α. Οι διάφοροι τύποι διάσχισης γράφων κατατάσσονται ανάλογα με τη σειρά με την οποία προσπελάζονται οι κόμβοι. Οι πιο γνωστές μέθοδοι διάσχισης γράφων είναι οι ακόλουθες:

1. *Αναζήτηση κατά πλάτος (Breadth-First Search ή BFS)*: Στην αναζήτηση κατά πλάτος για κάθε κορυφή u την οποία επισκεπτόμαστε, γίνεται επίσκεψη σε όλες τις γειτονικές κορυφές αυτής προτού προχωρήσουμε σε κάποια επόμενη κορυφή. Στην αναζήτηση κατά πλάτος χρησιμοποιείται γενικά δομή ουράς (queue).

2. *Αναζήτηση κατά βάθος (Depth-First Search ή DFS)*: Στην αναζήτηση κατά βάθος, ξεκινώντας από μια κορυφή u, επισκεπτόμαστε το πρώτο της παιδί και στη συνέχεια προχωράμε στο παιδί της δεύτερης κ.ο.κ έως ότου συναντήσουμε κάποια κορυφή που δεν έχει παιδί. Έπειτα επιστρέφουμε στην πρώτη ανεπεξέργαστη κορυφή και ακολουθούμε την ίδια διαδικασία. Στην αναζήτηση κατά βάθος χρησιμοποιείται γενικά δομή στοίβας (stack).

Στο Σχήμα 3.2 φαίνεται ένα παράδειγμα διάσχισης γράφου με αναζήτηση κατά πλάτος (3.2a) και αναζήτηση κατά βάθος (3.2b). Ο αριθμός κάθε κόμβου υποδηλώνει τη σειρά επίσκεψής του θεωρώντας ότι αρχικός κόμβος είναι αυτός με τον αριθμό 1.

Το πρόβλημα συντομότερης διαδρομής

Στη θεωρία γράφων,- το πρόβλημα συντομότερης διαδρομής ορίζεται ως το πρόβλημα εύρεσης της διαδρομής με το μικρότερο κόστος μεταξύ ενός αρχικού κόμβου (source node) και ενός τελικού κόμβου (sink node). Στην περίπτωση ενός μη σταθμισμένου γράφου η συντομότερη διαδρομή είναι αυτή που διέρχεται από τον ελάχιστο αριθμό ακμών. Ένας από τους πιο γνωστούς και αποτελεσματικούς αλγορίθμους εύρεσης συντομότερων διαδρομών είναι ο *αλγόριθμος του Dijkstra* [Dijk59].

Ο αλγόριθμος αυτός εμφανίζεται σε διάφορες παραλλαγές. Η αρχική μορφή του αλγορίθμου υπολόγιζε τη συντομότερη διαδρομή μεταξύ δύο κόμβων αλλά η πιο συνηθισμένη παραλλαγή χρησιμοποιεί έναν μοναδικό κόμβο σαν αρχικό και υπολογίζει τις συντομότερες διαδρομές από τον κόμβο αυτό προς όλους τους υπολοίπους κόμβους του δικτύου παράγοντας έτσι ένα δένδρο συντομότερων διαδρομών. Τα βήματα του αλγορίθμου είναι τα παρακάτω :

1. Θεωρείται για κάθε κόμβο ένα διάνυσμα απόστασης $d[\star]$ με τιμή 0 στον αρχικό κόμβο και τιμή άπειρο σε όλους τους υπόλοιπους. Επίσης, θεωρείται ένα διάνυσμα προηγούμενου κόμβου $prev[\star]$ το οποίο αρχικοποιείται με κενή τιμή για όλους τους κόμβους. Το διάνυσμα αυτό χρειάζεται για τον υπολογισμό της ζητούμενης διαδρομής στο τέλος.

2. Σημειώνονται όλοι οι κόμβοι ως μη-επεξεργασμένοι ενώ ο αρχικός κόμβος θεωρείται ως ο τρέχων κόμβος (current node).

3. Για τον τρέχοντα κόμβο, εξετάζονται όλοι οι μη-επεξεργασμένοι γειτονικοί κόμβοι του και υπολογίζεται το συνολικό άθροισμα της απόστασής τους από τον αρχικό κόμβο. Για παράδειγμα, αν ο τρέχων κόμβος έχει απόσταση 6 από τον αρχικό και ο γειτονικός του τρέχοντος κόμβου, που εξετάζει αυτή τη στιγμή ο αλγόριθμος, έχει απόσταση 2 από τον τρέχων, το συνολικό άθροισμα απόστασης του γείτονα από τον αρχικό κόμβο είναι 6+2=8. Αν αυτή η απόσταση είναι μικρότερη από την τιμή του διανύσματος απόστασης που είχε σημειωθεί, τότε αυτή αντικαθίσταται από τη νέα υπολογισμένη τιμή και σημειώνεται ο τρέχων κόμβος στην ετικέτα προηγούμενου κόμβου.

4. Ο τρέχων κόμβος σημειώνεται ως επεξεργασμένος. Ένας επεξεργασμένος κόμβος δεν εξετάζεται ποτέ ξανά από τον αλγόριθμο. Το διάνυσμα απόστασής περιέχει την ελάχιστη τιμή και αυτή θα παραμείνει σταθερή.

5. Ο επόμενος τρέχων κόμβος θα είναι ο μη-επεξεργασμένος κόμβος με τη μικρότερη τιμή στο διάνυσμα απόστασης.

6. Αν όλοι οι κόμβοι έχουν σημειωθεί ως επεξεργασμένοι, ο αλγόριθμος προχωρά στο επόμενο βήμα. Διαφορετικά, ο αλγόριθμος επαναλαμβάνεται από το βήμα 3.

7. Για την εύρεση του συνόλου των κόμβων της συντομότερης διαδρομής μεταξύ ενός αρχικού κόμβου και ενός κόμβου-στόχου, ξεκινώντας από τον κόμβο-προορισμό (ο οποίος είναι ο τελευταίος τρέχων κόμβος) εκτυπώνεται ο κόμβος που αναγράφεται στο διάνυσμα προηγούμενου κόμβου (prev) επαναληπτικά μέχρι να βρεθεί μια άδεια τιμή.

3.1.2 Μετρικές ανάλυσης κοινωνικών δικτύων

Στην ανάλυση κοινωνικών δικτύων χρησιμοποιείται ένα σύνολο μετρικών οι οποίες έχουν προταθεί κατά καιρούς από την επιστημονική κοινότητα. Οι μετρικές αυτές μπορούν να χωριστούν σε τρεις κατηγορίες ανάλογα με τον το είδος τους. Οι κατηγορίες περιλαμβάνουν μετρικές οι οποίες αφορούν *συνδέσεις (connections)*, *κατανομές (distributions)* και *κατάτμηση (segmentation)*.

Συνδέσεις

- *Ομοφιλία (Homophily)*: Ο βαθμός στον οποίο οι κοινωνικοί δράστες συνάπτουν συνδέσεις με όμοιούς τους. Η ομοιότητα μπορεί να οριστεί με διάφορα κριτήρια όπως είναι το γένος, η φυλή, το επάγγελμα, η κοινωνική θέση, η εκπαίδευση και οποιαδήποτε άλλα προεξέχοντα χαρακτηριστικά [McPh01].
- *Πολυπλοκότητα (Multiplexity)*: Ο αριθμός των μορφών που περιλαμβάνονται σε ένα κοινωνικό δεσμό. Για παράδειγμα εάν δύο άνθρωποι είναι συγγενείς και δουλεύουν και μαζί τότε ο δεσμός τους έχει τιμή πολυπλοκότητας ίση με 2 [Podo97].
- *Αμοιβαιότητα (Reciprocity/Mutuality)*: Ο βαθμός στον οποίο οι κοινωνικοί δράστες ανταποδίδουν την αλληλεπίδραση (π.χ. ο Α είναι φίλος με τον Β αλλά και ο Β είναι φίλος με τον Α).
- *Εγγύτητα (Propinquity)*: Η τάση των κοινωνικών δραστών να έχουν περισσότερους δεσμούς με άλλους οι οποίο βρίσκονται τοπολογικά κοντά τους.

Κατανομές

- *Κεντρικότητα (Centrality)*: Η κεντρικότητα αναφέρεται σε ένα σύνολο μετρικών οι οποίες έχουν στόχο να ποσοτικοποιήσουν τη σημαντικότητα ή επιρροή ενός συγκεκριμένου κόμβου (ή ομάδας) σε ένα δίκτυο. Παραδείγματα γνωστών μετρικών κεντρικότητας αποτελούν η *κεντρικότητα βαθμού (degree centrality)*, η *ιδιοδιανυσματική κεντρικότητα (eigenvector centrality)*, η *κεντρικότητα κατά Katz (Katz centrality)*, η *κεντρικότητα ενδιαμεσότητας (betweenness centrality)* και η *κεντρικότητα εγγύτητας*.
- *Πυκνότητα (Density)*: Το ποσοστό των συνδέσεων ενός δικτύου σε σχέση με το τον αριθμό των δυνατών συνδέσεων. Η πυκνότητα χαρακτηρίζει ένα δίκτυο ως πυκνό ή αραιό σε περίπτωση μεγάλης και μικρής τιμής της πυκνότητας αντίστοιχα.
- *Απόσταση (Distance)*: Ο ελάχιστος αριθμός συνδέσεων που απαιτείται προκειμένου να συνδεθούν δύο συγκεκριμένοι κόμβοι. Με τη μετρική αυτή σχετίζεται και το *πείραμα του μικρού κόσμου (small-world experiment)*, γνωστό και ως *έξι βαθμοί διαχωρισμού (six degrees of separation)* [Trav67].

Κατάτμηση

- *Συντελεστής ομαδοποίησης (Clustering Coefficient)*: Η μέση πιθανότητα δύο κόμβων οι οποίοι συνδέονται με έναν τρίτο κόμβο να συνδέονται και μεταξύ τους. Τοπικά, ορίζεται ως ο αριθμός των τριγώνων που συνδέονται με έναν κόμβο u προς των αριθμό συνδεδεμένων τριάδων με κέντρο την κορυφή u [Watt98].
- *Συνοχή (Cohesion)*: Αναφέρεται στον ελάχιστο αριθμό μελών μιας συνεκτικής ομάδας ενός δικτύου τα οποία αν αφαιρεθούν αποσυνδέουν την ομάδα [Mood03].

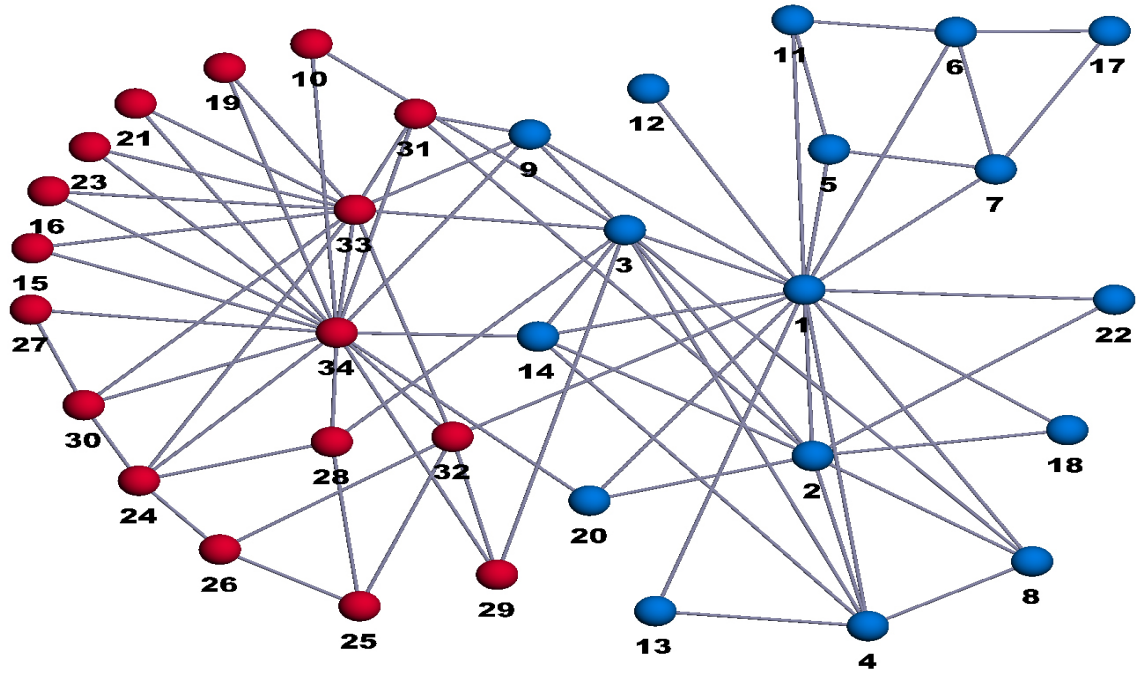
3.1.3 Εφαρμογές της ανάλυσης κοινωνικών δικτύων

Η ανάλυση κοινωνικών χρησιμοποιείται εκτεταμένα σε μια μεγάλη ποικιλία εφαρμογών. Σε αυτές περιλαμβάνονται η *ανίχνευση κοινοτήτων (community detection)* και η *πρόβλεψη συνδέσμου (link prediction)* σε δίκτυα, η ανάπτυξη συστημάτων συστάσεων, η εξόρυξη δεδομένων, η μοντελοποίηση δικτύων και η ανάλυση συμπεριφοράς χρήστη (user behavior analysis) [Golb13]. Επίσης, στον ιδιωτικό τομέα η ανάλυση κοινωνικών δικτύων χρησιμοποιείται από επιχειρήσεις για την υποστήριξη δραστηριοτήτων όπως η ανάλυση καταναλωτικής συμπεριφοράς και η αλληλεπίδραση με τους καταναλωτές (customer interaction and analysis), η ανάλυση ανάπτυξης πληροφοριακών συστημάτων (information system development analysis), το μάρκετινγκ και η επιχειρησιακή νοημοσύνη (business intelligence). Στην παρούσα διπλωματική δίνεται έμφαση στην ανίχνευση κοινοτήτων η οποία εφαρμόστηκε πρακτικά στο πειραματικό μέρος της εργασίας.

3.2 Ανίχνευση κοινοτήτων

Η μελέτη των πολύπλοκων δικτύων και κατ' επέκταση των κοινωνικών δικτύων έχει δείξει ότι τα συστήματα αυτά δεν είναι ούτε απολύτως κανονικά ούτε απολύτως τυχαία. Πρόκειται για συστήματα στα οποία συνυπάρχουν η τάξη με την αταξία. Σε αντίθεση με τους τυχαίους γράφους όπου η πιθανότητα να υπάρχει μια ακμή ανάμεσα σε δύο κόμβους είναι ίδια για όλα τα πιθανά ζεύγη κόμβων [Erdo59], τα πραγματικά δίκτυα παρουσιάζουν μεγάλες ανομοιογένειες στην κατανομή των ακμών. Οι ανομοιογένειες αυτές μάλιστα δεν εμφανίζονται μόνο καθολικά αλλά και τοπικά, με μεγάλες συγκεντρώσεις ακμών εντός συγκεκριμένων ομάδων από κόμβους και μικρές συγκεντρώσεις εκτός αυτών.

Το χαρακτηριστικό αυτό των πολύπλοκων δικτύων, που ονομάζεται *κοινοτική δομή* (community structure), αποτελεί ένα από τα πιο σημαντικά χαρακτηριστικά των γράφων που περιγράφουν πραγματικά συστήματα. Οι κοινότητες, που συχνά αναφέρονται και ως *συστάδες (clusters)*, είναι ομάδες κορυφών οι οποίες πιθανώς μοιράζονται παρόμοιες ιδιότητες ή/και παίζουν παρόμοιο ρόλο μέσα στο δίκτυο [Fort10]. Η ανίχνευση κοινοτήτων παίζει πολύ σημαντικό ρόλο στην πληροφορική, την κοινωνιολογία και τη βιολογία, τομείς όπου είναι σύνηθες τα συστήματα να αναπαρίστανται ως γράφοι. Ένα χαρακτηριστικό παράδειγμα απλού γράφου με κοινοτική δομή αποτελεί το *Karate Club* του Zachary [Zach77], το οποίο απεικονίζεται στο Σχήμα 3.3. Όπως φαίνεται στο εν λόγω σχήμα το κοινωνικό αυτό δίκτυο περιλαμβάνει δύο εμφανείς κοινότητες, οι οποίες έχουν ως κεντρικούς κόμβους τους 1 και 34 αντίστοιχα.



Σχήμα 3.3: Οι δύο κοινότητες του Karate Club του Zachary [Zach77]

3.2.1 Στοιχεία Ανίχνευσης κοινοτήτων

Το πρόβλημα της ανίχνευσης κοινοτήτων, παρότι διαισθητικό με μια πρώτη ματιά, στην πραγματικότητα δεν είναι σαφώς ορισμένο. Αυτό εξηγείται από το γεγονός ότι τα κύρια στοιχεία του προβλήματος, όπως είναι οι έννοιες της κοινότητας και της διαμέρισης, δεν είναι αυστηρώς ορισμένα αλλά εμπεριέχουν σε κάποιο βαθμό αυθαίρετη σκέψη.

Είναι σημαντικό να τονιστεί ότι ο προσδιορισμός των δομικών ομάδων είναι εφικτός μόνο όταν οι γράφοι είναι *αραιοί (sparse)*, δηλαδή όταν ο αριθμός των ακμών m είναι της τάξης του αριθμού των κόμβων n του γράφου. Στην περίπτωση όπου $m \gg n$ τότε η κατανομή των ακμών μεταξύ των

κόμβων είναι υπερβολικά ομογενής ώστε να έχουν νόημα οι κοινότητες και το πρόβλημα μετατρέπεται περισσότερο σε πρόβλημα ομαδοποίησης (clustering) το οποίο απαιτεί έννοιες και μεθόδους διαφορετικής φύσης.

Η έννοια της κοινότητας

Το πρωταρχικό πρόβλημα στην ανίχνευση κοινοτήτων είναι να δοθεί ένας ποσοτικός ορισμός της έννοιας της κοινότητας. Αξίζει να σημειωθεί πως στο συγκεκριμένο ζήτημα, κανένας ορισμός δεν είναι καθολικά αποδεκτός ενώ συχνά εξαρτάται από το εκάστοτε σύστημα ή την εφαρμογή που χρησιμοποιείται στην πράξη. Διαισθητικά, καταλαβαίνει κανείς ότι πρέπει να υπάρχουν περισσότερες ακμές "εντός" της κοινότητας από ακμές που συνδέουν κόμβους της κοινότητας με τον υπόλοιπο γράφο.

Έστω C ο υπογράφος ενός γράφου G με $|C| = n_c$ και $|G| = n$ πλήθος κορυφών αντίστοιχα. Ορίζουμε τον *εσωτερικό (internal degree)* και *εξωτερικό βαθμό (external degree)* της κορυφής $v \in C$, k_v^{int} και k_v^{ext} , ως το πλήθος των ακμών που συνδέουν τη v με άλλες κορυφές του C ή με τον υπόλοιπο γράφο αντίστοιχα. Στην περίπτωση που $k_v^{ext} = 0$, η κορυφή αυτή έχει γείτονες μόνο μέσα στον C, με αποτέλεσμα αυτός να είναι μια καλή ομάδα για την v. Αντίθετα, αν $k_v^{int} = 0$, τότε η κορυφή v είναι ασύνδετη με τον C και θα πρέπει καλύτερα να ανατεθεί σε άλλη ομάδα. Ο εσωτερικός βαθμός k_{int}^C του υπογράφου C ορίζεται ως το άθροισμα των εσωτερικών βαθμών των κορυφών του, ενώ αντίστοιχα ο εξωτερικός βαθμός k_{ext}^C ορίζεται ως το άθροισμα των εξωτερικών βαθμών των κορυφών του C. Ο *συνολικός βαθμός (total degree)* ορίζεται ως το άθροισμα των βαθμών όλων των κορυφών του C. Εξ' ορισμού, ισχύει ότι $k^C = k_{int}^C + k_{ext}^C$.

Ορίζουμε την *ενδο-συσταδική πυκνότητα (intra-cluster density)* $\delta_{int}(C)$ του υπογράφου C ως το λόγο του πλήθους των εσωτερικών ακμών του C προς το πλήθος όλων των δυνατών εσωτερικών ακμών, δηλαδή:

$$\delta_{int}(C) = \frac{\text{\#εσωτερικών ακμών του C}}{\frac{n_c(n_c-1)}{2}} \tag{3.3}$$

Αντίστοιχα, η *δια-συσταδική πυκνότητα (inter-cluster density)* $\delta_{ext}(C)$ είναι ο λόγος του πλήθους των ακμών που συνδέουν τις κορυφές του C με τον υπόλοιπο γράφο προς το πλήθος όλων των δυνατών τέτοιων ακμών, δηλαδή:

$$\delta_{ext}(C) = \frac{\text{\#δια-ομαδικών ακμών του C}}{\frac{n_c(n-n_c)}{2}} \tag{3.4}$$

Σύμφωνα με τα παραπάνω, προκειμένου να αποτελεί το C μια κοινότητα, αναμένουμε η ενδο-συσταδική πυκνότητα $\delta_{int}(C)$ να είναι αισθητά μεγαλύτερη από τη μέση πυκνότητα συνδέσμων $\delta(G)$ του G, η οποία δίνεται από το λόγο του πλήθους των ακμών του G προς το πλήθος όλων των δυνατών ακμών $n(n-1)/2$. Από την άλλη, η δια-συσταδική πυκνότητα $\delta_{ext}(C)$ αναμένεται να είναι πολύ μικρότερη από τη $\delta_{int}(G)$. Η αναζήτηση της καλύτερης αντιστάθμισης μεταξύ μιας μεγάλης $\delta(G)$ και μιας μικρής $\delta_{ext}(C)$ είναι άμεσα ή έμμεσα ο στόχος κάθε αλγορίθμου ομαδοποίησης.

Μια απαραίτητη ιδιότητα για μια κοινότητα είναι η *συνεκτικότητα (connectedness)* η οποία παρουσιάστηκε στην Ενότητα 3.1.1. Αυτό σημαίνει ότι για να αποτελεί ο υπογράφος C μια κοινότητα θα πρέπει να υπάρχει για κάθε ζεύγος κορυφών του ένα μονοπάτι που τις συνδέει και διέρχεται μόνο από κορυφές του C. Αυτό το χαρακτηριστικό απλοποιεί το πρόβλημα της ανίχνευσης κοινοτήτων σε μη-συνδεδεμένους γράφους, αφού σε αυτή την περίπτωση μπορεί κάποιος να επεξεργαστεί κάθε συνδεδεμένο μέρος του γράφου ξεχωριστά εκτός αν επιβάλλονται ειδικοί περιορισμοί ως προς τις τελικές ομάδες. Το χαρακτηριστικό της συνεκτικότητας, όπως θα δούμε σε επόμενο κεφάλαιο, παίζει σημαντικό ρόλο και στο πειραματικό κομμάτι αυτής της εργασίας.

Παραπάνω, δόθηκαν κάποιοι βασικοί ορισμοί σχετικά με την έννοια της κοινότητας. Ωστόσο, διάφοροι άλλοι ορισμοί έχουν δοθεί από αναλυτές κοινωνικών δικτύων, επιστήμονες πληροφορικής και φυσικούς. Οι ορισμοί αυτοί μπορούν να χωριστούν σε τρεις κατηγορίες: τοπικοί, καθολικοί και βασισμένοι στην ομοιότητα κορυφών [Fort10].

Η έννοια της διαμέρισης

Διαμέριση ονομάζεται μια διαίρεση του γράφου σε ομάδες, έτσι ώστε κάθε κορυφή να ανήκει σε μια ομάδα. Στα πραγματικά συστήματα, μια κορυφή μπορεί να ανήκει σε παραπάνω από μια κοινότητες. Μια διαίρεση του γράφου σε επικαλυπτόμενες ή ασαφείς κοινότητες ονομάζεται *κάλυψη (cover)*.

Ο αριθμός των πιθανών διαμερίσεων σε k ομάδες ενός γράφου με n κόμβους ισούται με τον αριθμό Stirling δεύτερου είδους (Stirling number of the second kind) $S(n,k)$. Ο συνολικός αριθμός των πιθανών διαμερίσεων ισούται με το n-οστό αριθμό Bell $B_n = \sum_{k=0}^{n} S(n,k)$ [Andr76]. Ο παραπάνω τύπος, στο όριο για πολύ μεγάλες τιμές του n παίρνει την παρακάτω ασυμπτωτική μορφή [Lova93]:

$$B_n \sim \frac{1}{\sqrt{n}}[\lambda(n)]^{n+1/2} e^{\lambda(n)-n-1} \tag{3.5}$$

όπου $\lambda(n) = e^{W(n)} = n/W(n)$, με το $W(n)$ να είναι η *συνάρτηση Lambert W*. Για το λόγο αυτό, το B_n αυξάνεται ταχύτερα από εκθετικά με το μέγεθος του γράφου n, το οποίο σημαίνει ότι η απαρίθμηση ή αξιολόγηση όλων των διαμερίσεων ενός γράφου δεν είναι εφικτή παρά μόνο αν αυτός αποτελείται από ένα πολύ μικρό πλήθος κόμβων.

Οι διαμερίσεις μπορεί να είναι ιεραρχικά διατεταγμένες αν ο γράφος έχει διαφορετικά επίπεδα οργάνωσης/δομής σε διαφορετικές κλίμακες. Σε αυτή την περίπτωση, οι ομάδες παρουσιάζουν και οι ίδιες κοινοτική δομή, περιέχοντας μικρότερες κοινότητες, οι οποίες με τη σειρά τους μπορούν να περιέχουν μικρότερες κοινότητες κ.ο.κ. Ένα φυσικό τρόπο αναπαράστασης της ιεραρχικής δομής ενός γράφου αποτελεί το δενδρόγραμμα.

Συναρτήσεις ποιότητας: Τμηματικότητα

Οι αξιόπιστοι αλγόριθμοι ανίχνευσης κοινοτήτων θεωρητικά θα πρέπει να επιστρέφουν "καλές" διαμερίσεις. Ωστόσο, αυτό δε συμβαίνει πάντα και γι' αυτό το λόγο είναι χρήσιμο να υπάρχει ένα ποσοτικό κριτήριο προκειμένου να μπορεί κάποιος να εκτιμήσει το πόσο "καλή" είναι μια διαμέριση. Το κριτήριο αυτό είναι συνήθως μια *συνάρτηση ποιότητας (quality function)*.

Μια συνάρτηση ποιότητας είναι μια συνάρτηση η οποία αντιστοιχίζει έναν αριθμό σε κάθε διαμέριση ενός γράφου. Με αυτό τον τρόπο μπορεί κάποιος να κατατάξει τις διαμερίσεις με βάση τη βαθμολογία που της δόθηκε από τη συνάρτηση ποιότητας. Έτσι, οι διαμερίσεις με υψηλές βαθμολογίες είναι "καλές" ενώ αυτή με τη μεγαλύτερη βαθμολογία θα είναι εξ' ορισμού η καλύτερη. Εντούτοις, πρέπει να έχουμε υπόψη ότι το πρόβλημα του κατά πόσο μια διαμέριση είναι "καλύτερη" από μια άλλη δεν είναι πλήρως ορισμένο και εξαρτάται πάντα από το εκάστοτε είδος της κοινότητας ή/και της συνάρτησης ποιότητας που χρησιμοποιείται.

Μια συνάρτηση ποιότητας Q είναι *αθροιστική (additive)* αν υπάρχει μια βοηθητική συνάρτηση q έτσι ώστε για κάθε διαμέριση P ενός γράφου

$$Q(P) = \sum_{c \in P} q(C) \tag{3.6}$$

όπου C είναι μια παραγόμενη ομάδα (cluster) της διαμέρισης P. Η παραπάνω εξίσωση δηλώνει ότι η ποιότητα μιας διαμέρισης δίνεται από το άθροισμα των ποιοτήτων των επιμέρους ομάδων. Οι περισσότερες συναρτήσεις ποιότητας που χρησιμοποιούνται στη βιβλιογραφία είναι αθροιστικές χωρίς ωστόσο αυτό να αποτελεί απαραίτητη προϋπόθεση.

Ένα παράδειγμα συνάρτησης ποιότητας είναι η *επίδοση (performance) P*, η οποία μετρά το πλήθος των ορθά "ερμηνευμένων" ζευγών κορυφών, δηλαδή κορυφών οι οποίες ανήκουν στην ίδια κοινότητα και συνδέονται με ακμή ή κορυφών που δεν ανήκουν στην ίδια κοινότητα και δε συνδέονται με ακμή. Ο ορισμός της συνάρτησης της επίδοσης για μια διαμέριση P είναι ο παρακάτω:

$$P(\wp) = \frac{|(i,j) \in E, C_i = C_j| + |(i,j) \notin E, C_i \neq C_j|}{n(n-1)/2} \tag{3.7}$$

Εξ' ορισμού ισχύει ότι $0 \leq P() \leq 1$. Ένα άλλο παράδειγμα αποτελεί η *κάλυψη (coverage)*, δηλαδή ο λόγος του πλήθους των ενδοκοινοτικών ακμών προς το συνολικό πλήθος ακμών.

Η πιο δημοφιλής συνάρτηση ποιότητας είναι η *τμηματικότητα (modularity)* των Newman και Girvan [Fort10]. Η μετρική αυτή βασίζεται στην ιδέα ότι ένας τυχαίος γράφος δεν αναμένεται να έχει κοινοτική δομή, επομένως η πιθανή ύπαρξη κοινοτήτων μπορεί να αποκαλυφθεί από τη σύγκριση μεταξύ της πραγματικής πυκνότητας των ακμών σε έναν υπογράφο και της αναμενόμενης πυκνότητας των ακμών του υπογράφου, εάν οι κορυφές συνδέονταν ασχέτως κοινοτικής δομής. Αυτή η αναμενόμενη πυκνότητα των ακμών εξαρτάται από το επιλεγμένο μοντέλο αναφοράς (null model), δηλαδή το αντίγραφο του αρχικού γράφου, το οποίο διατηρεί κάποιες από τις δομικές του ιδιότητες αλλά στερείται κοινοτικής δομής. Η τμηματικότητα μπορεί να γραφτεί όπως παρακάτω:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j) \tag{3.8}$$

, όπου το άθροισμα περιλαμβάνει όλα τα ζεύγη κορυφών, A είναι ο πίνακας γειτνίασης, m ο συνολικός αριθμός ακμών του γράφου και P_{ij} ο αναμενόμενος αριθμός ακμών μεταξύ των κορυφών i και j στο μοντέλο αναφοράς. Η συνάρτηση δ παίρνει την τιμή 1 αν οι κορυφές i και j ανήκουν στην ίδια κοινότητα ($C_i = C_j$) και μηδέν αλλιώς. Η επιλογή του μοντέλου αναφοράς είναι γενικά αυθαίρετη και υπάρχουν πολλές δυνατότητες, ωστόσο προτιμάται αυτό να έχει την ίδια κατανομή βαθμών με την αντίστοιχη του αρχικού γράφου [Fort10].

Η τμηματικότητα μπορεί να γραφεί και ως εξής:

$$Q = \sum_{c=1}^{n_c} [\frac{l_c}{m} - (\frac{d_c}{2m})^2] \tag{3.9}$$

, όπου n_c είναι ο αριθμός των ομάδων, l_c ο συνολικός αριθμός ακμών που συνδέουν κορυφές της κοινότητας c και d_c το άθροισμα των βαθμών των κορυφών του c. Στην παραπάνω εξίσωση ο πρώτος όρος του αθροίσματος παριστάνει το ποσοστό των ακμών του γράφου που βρίσκονται εντός της κοινότητας c, ενώ ο δεύτερος όρος το αναμενόμενο ποσοστό των ακμών εντός της κοινότητας αν ο γράφος ήταν τυχαίος με τον ίδιο αναμενόμενο βαθμό για κάθε κορυφή.

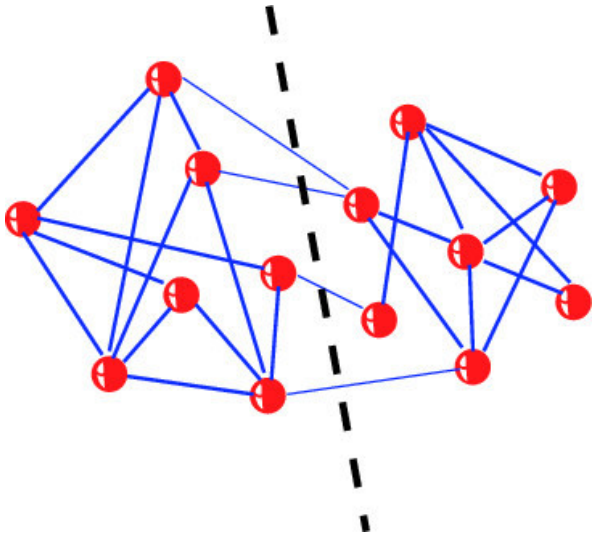
Σύμφωνα με την Εξίσωση 3.9 ένας υπογράφος αποτελεί κοινότητα αν η συνεισφορά του στο άθροισμα της τμηματικότητας είναι θετική. Όσο περισσότερο ο αριθμός των εσωτερικών ακμών της κοινότητας ξεπερνάει τον αναμενόμενο, τόσο καλύτερα ορισμένη είναι η κοινότητα. Συνήθως μεγάλες θετικές τιμές της τμηματικότητας υποδεικνύουν "καλές" διαμερίσεις. Ωστόσο, αξίζει να σημειωθεί ότι η μέγιστη τιμή της τμηματικότητας ενός γράφου γενικά μεγαλώνει όταν το μέγεθος του γράφου ή/και ο αριθμός των καλά διαχωρισμένων ομάδων μεγαλώνουν [Good10]. Επομένως, η τμηματικότητα δε θα πρέπει να χρησιμοποιείται ως μέτρο σύγκρισης της ποιότητας της κοινοτικής δομής μεταξύ γράφων πολύ διαφορετικών μεγεθών.

3.2.2 Κλασσικές μέθοδοι ανίχνευσης κοινοτήτων

Διαμερισμός γράφου

Το πρόβλημα του διαμερισμού ενός γράφου $G = (V, E)$ συνίσταται στη διαίρεση των κορυφών του σε g ομάδες προκαθορισμένου μεγέθους με σκοπό ο αριθμός των ακμών μεταξύ των ομάδων αυτών να είναι ο ελάχιστος δυνατός. Ο αριθμός των ακμών οι οποίες συνδέουν τις διαφορετικές ομάδες ονομάζεται *μέγεθος τομής (cut size)*. Μια τομή αποκαλείται *ελάχιστη (minimum cut)* όταν δεν υπάρχει δυνατή τομή με μικρότερο μέγεθος και αντίστοιχα *μέγιστη (maximum cut)* όταν δεν υπάρχει δυνατή τομή με μεγαλύτερο μέγεθος. Το Σχήμα 3.3 απεικονίζει τη λύση του προβλήματος της διαμέρισης ενός γράφου με 14 κορυφές για $g = 2$ ομάδες ίδιου μεγέθους με την ελάχιστη τομή. Το μέγεθος τομής στο παράδειγμα αυτό είναι 4.

Το βασικό πρόβλημα των αλγορίθμων διαμερισμού γράφου είναι ότι απαιτούν ως είσοδο τον αριθμό των ομάδων και κάποιες φορές το μέγεθος αυτών. Αυτό το πρόβλημα μπορεί να αποφευχθεί επιλέγοντας κάποιο άλλο μέτρο για τη βελτιστοποίηση του διαμερισμού αντί για το μέγεθος των



Σχήμα 3.4: Διαμερισμός γράφου σε δύο ομάδες με ελάχιστο μέγεθος τομής

ομάδων. Οι περισσότεροι από τους αλγορίθμους αυτούς πραγματοποιούν μια διχοτόμηση του γράφου, ενώ διαμερίσεις σε περισσότερες από δύο ομάδες συνήθως επιτυγχάνονται με επαναληπτικές διχοτομήσεις. Τέλος, στις περισσότερες περιπτώσεις επιβάλλεται ο περιορισμός οι ομάδες να έχουν ίδιο μέγεθος. Το πρόβλημα αυτό είναι γνωστό και ως πρόβλημα *ελάχιστης διχοτόμησης (minimum bisection)* ή τομή (2,1).

Ο αλγόριθμος Kernighan-Lin

Ο αλγόριθμος Kernighan-Lin [Kern70] αποτελεί μια από τις πιο παλιές μεθόδους διαμερισμού γράφου η οποία χρησιμοποιείται ακόμα και σήμερα, συνήθως σε συνδυασμό με άλλες μεθόδους. Ο αλγόριθμος αυτός μάλιστα έχει σημαντικές εφαρμογές στον τομέα της σχεδίασης ψηφιακών κυκλωμάτων VLSI [Ravi96]. Οι συγγραφείς του αλγορίθμου εμπνευστήκαν από το πρόβλημα του διαμερισμού ηλεκτρονικών κυκλωμάτων σε πλακέτες ούτως ώστε οι κόμβοι που εμπεριέχονται σε διαφορετικές πλακέτες να συνδέονται μεταξύ τους με το μικρότερο δυνατό αριθμό συνδέσεων.

Ο σκοπός του αλγορίθμου αυτού είναι η βελτιστοποίηση μιας συνάρτησης κέρδους Q, η οποία παριστάνει τη διαφορά μεταξύ του πλήθους των ακμών που βρίσκονται μέσα στις ομάδες του γράφου και των ακμών που βρίσκονται ανάμεσα σε αυτές. Τα βήματα του αλγορίθμου συνοψίζονται ως εξής:

1. Πραγματοποιείται μια αρχική διαμέριση του γράφου σε δύο ομάδες προκαθορισμένου μεγέθους, η οποία μπορεί να είναι είτε τυχαία ή προτεινόμενη από κάποια πληροφορία σχετικά με τη δομή του γράφου.

2. Υποσύνολα που αποτελούνται από ίδιο αριθμό κορυφών ανταλλάσσονται μεταξύ των δύο ομάδων έτσι ώστε να επιτυγχάνεται η μέγιστη αύξηση της συνάρτησης Q.

3. Επιλέγεται η διαμέριση με τη μεγαλύτερη τιμή της Q προκειμένου να χρησιμοποιηθεί ως αφετηρία μιας νέας σειράς επαναλήψεων του βήματος 2.

Προκειμένου να μειωθεί ο κίνδυνος να παγιδευθεί ο αλγόριθμος σε ένα τοπικό μέγιστο, η διαδικασία του βήματος 2 περιλαμβάνει και ανταλλαγές οι οποίες μειώνουν την τιμή της Q.

Ο αλγόριθμος Kernighan-Lin είναι αρκετά γρήγορος, έχοντας πολυπλοκότητα $O(n^2 \log n)$ (όπου n ο αριθμός των κορυφών) στην περίπτωση που ένας σταθερός αριθμός ανταλλαγών πραγματοποιείται σε κάθε επανάληψη. Το πιο ακριβό υπολογιστικά κομμάτι του αλγορίθμου είναι η εύρεση των υποσυνόλων προς ανταλλαγή, το οποίο απαιτεί τη σύγκριση των κερδών/απωλειών μεταξύ όλων των υποψήφιων υποσυνόλων.

Οι τελικές διαμερίσεις που επιστρέφονται από τον αλγόριθμο εξαρτώνται έντονα από την αρχική διάταξη. Έτσι, είναι προτιμητέο να ξεκινάει κανείς με μια καλή εκτίμηση της ζητούμενης διαμέρισης αλλιώς το αποτέλεσμα είναι πιθανό να μην είναι ικανοποιητικό. Εξαιτίας του προβλήματος της αρχικοποίησης, ο αλγόριθμος αυτός χρησιμοποιείται συχνά για τη βελτίωση των διαμερίσεων που εντοπίζονται με τη χρήση άλλων τεχνικών, οι οποίες χρησιμοποιούνται ως αρχικές διαμορφώσεις (configurations) για τον αλγόριθμο. Ο αλγόριθμος των Kernighan-Lin έχει επεκταθεί ώστε να εξάγει διαμερίσεις οποιουδήποτε αριθμού ομάδων [Suar88], ωστόσο ο χρόνος εκτέλεσης και το κόστος σε μνήμη αυξάνονται γρήγορα με τον αριθμό των ομάδων.

3.2.3 Ιεραρχική Συσταδοποίηση

Γενικότερα, είναι δύσκολο να γνωρίζει κανείς εκ των προτέρων πράγματα σχετικά με την κοινοτική δομή ενός γράφου, όπως για παράδειγμα τον αριθμό των ομάδων στις οποίες αυτός χωρίζεται. Όταν ισχύει αυτό, διαδικασίες ομαδοποίησης όπως οι μέθοδοι διαμερισμού γράφου που περιγράφηκαν παραπάνω δύσκολα μπορούν να φανούν χρήσιμες και είναι αναγκαίο να γίνουν κάποιες λογικές υποθέσεις όσον αφορά τον αριθμό και το μέγεθος των ομάδων που παραμένουν απροσδιόριστες. Από την άλλη, ένας γράφος μπορεί να έχει ιεραρχική δομή, δηλαδή να παρουσιάζει διάφορα επίπεδα ομαδοποίησης των κορυφών του με μικρές ομάδες να συμπεριλαμβάνονται μέσα σε μεγαλύτερες. Σε τέτοιες περιπτώσεις, μπορεί κάποιος να χρησιμοποιήσει *αλγορίθμους ιεραρχικής συσταδοποίησης (hierarchical clustering algorithms)* [Frie01], δηλαδή αλγορίθμους που αποκαλύπτουν την πολυεπίπεδη δομή ενός γράφου. Η ιεραρχική συσταδοποίηση χρησιμοποιείται ευρέως στην ανάλυση κοινωνικών δικτύων, στη βιολογία, στη μηχανική, στο μάρκετινγκ κ.α.

Η αφετηρία κάθε μεθόδου ιεραρχικής συσταδοποίησης είναι ο ορισμός ενός μέτρου ομοιότητας μεταξύ των κόμβων του γράφου. Εν συνεχεία, υπολογίζεται η ομοιότητα μεταξύ κάθε ζεύγους κόμβων είτε αυτοί συνδέονται είτε όχι και έτσι προκύπτει ένας $n \times n$ πίνακας X ο οποίος ονομάζεται πίνακας ομοιότητας. Οι αλγόριθμοι ιεραρχικής συσταδοποίησης μπορούν να χωριστούν σε δύο κατηγορίες:

1. *Σωρευτικοί αλγόριθμοι (agglomerative algorithms)*, στους οποίους κάθε κορυφή του γράφου ξεκινά ως αυτόνομη ομάδα και οι ομάδες συγχωνεύονται όταν η ομοιότητά τους είναι ικανοποιητικά υψηλή.

2. *Διαιρετικοί αλγόριθμοι (divisive algorithms)*, στους οποίους όλες οι κορυφές του γράφου ξεκινούν σε μια ομάδα και οι ομάδες χωρίζονται αφαιρώντας ακμές οι οποίες συνδέουν ομάδες με χαμηλή ομοιότητα.

Οι δύο παραπάνω κατηγορίες αλγορίθμων αναφέρονται σε αντίθετες διαδικασίες. Οι σωρευτικοί αλγόριθμοι είναι τύπου από-κάτω-προς-τα-πάνω (bottom-up) εφόσον ξεκινούν από το χαμηλότερο επίπεδο ιεραρχίας, όπου ο κάθε κόμβος αποτελεί μια δική του ομάδα. Αντίθετα, οι διαιρετικοί αλγόριθμοι είναι τύπου από-πάνω-προς-τα-κάτω (top-down) αφού ξεκινούν από το ανώτερο επίπεδο ιεραρχίας και καταλήγουν σε χαμηλότερα. Οι διαιρετικοί αλγόριθμοι ιεραρχικής συσταδοποίησης έχουν χρησιμοποιηθεί σπάνια στο παρελθόν ωστόσο όπως θα δούμε παρακάτω άλλες διαιρετικές μέθοδοι όπως ο αλγόριθμος Girvan-Newman έχουν αρχίσει να γίνονται πιο δημοφιλείς.

Στους σωρευτικούς αλγορίθμους, εφόσον οι ομάδες συγχωνεύονται με βάση την ομοιότητά τους, είναι απαραίτητος ο ορισμός ενός μέτρου το οποίο να εκτιμά την ομοιότητα των ομάδων με βάση τον πίνακα ομοιότητας X. Η συγκεκριμένη ομαδοποίηση επιτυγχάνεται με διάφορους τρόπους όπως λ.χ. η *συσταδοποίηση απλής διασύνδεσης (single linkage clustering)*, όπου η ομοιότητα μεταξύ των δύο ομάδων είναι το ελάχιστο στοιχείο x_{ij} με το i να ανήκει στη μία ομάδα και το j στην άλλη. Αντίθετα, το μέγιστο στοιχείο x_{ij} για κορυφές διαφορετικών ομάδων χρησιμοποιείται στην περίπτωση της *συσταδοποίησης πλήρους διασύνδεσης (complete linkage clustering)*. Αντίστοιχα, στην *συσταδοποίηση μέσης διασύνδεσης (average linkage clustering)* χρησιμοποιείται ο μέσος όρος των x_{ij}.

Η ιεραρχική συσταδοποίηση έχει το πλεονέκτημα ότι δεν απαιτεί κάποια πρότερη γνώση σχετικά με τον αριθμό και το μέγεθος των κοινοτήτων. Ωστόσο, δεν υπάρχει κάποιος κοινά αποδεκτός τρόπος διαχωρισμού των διάφορων διαμερίσεων που μπορεί να προκύψουν από τη διαδικασία, προκειμένου

να επιλεγεί αυτή ή αυτές που παριστάνουν καλύτερα την κοινοτική δομή ενός γράφου. Τα αποτελέσματα της μεθόδου εξαρτώνται σε μεγάλο βαθμό από το εκάστοτε μέτρο ομοιότητας που έχει υιοθετηθεί. Επίσης, η μέθοδος αυτή μπορεί, από την κατασκευή της, να δημιουργεί μια ιεραρχική δομή η οποία είναι μάλλον τεχνητή τις περισσότερες φορές, αφού ο γράφος που εξετάζεται μπορεί να μην έχει καν ιεραρχική δομή. Άλλα προβλήματα των μεθόδων αυτών είναι ότι μπορεί οι κορυφές μια κοινότητας να μην ταξινομούνται σωστά ενώ είναι συχνό φαινόμενο να αγνοούνται κόμβοι ακόμα και αν αυτοί έχουν σημαντικό ρόλο στις κοινότητες τους [Newm04a]. Επιπλέον, κορυφές οι οποίες έχουν μόνο ένα γείτονα συχνά ταξινομούνται σαν ξεχωριστές ομάδες, το οποίο τις περισσότερες φορές δεν έχει νόημα. Τέλος, ένα σημαντικό μειονέκτημα των σωρευτικών αλγορίθμων ιεραρχικής συσταδοποίησης είναι η πολυπλοκότητά τους, η οποία στην καλύτερη περίπτωση είναι $O(n^2)$ για απλή διασύνδεση και $O(n^2 \log n)$ για πλήρη και μέση διασύνδεση, ενώ μπορεί να είναι πολύ μεγαλύτερη όταν ο υπολογισμός της επιλεγμένης συνάρτησης ομοιότητας είναι δαπανηρός.

3.2.4 Φασματική Συσταδοποίηση

Οι αλγόριθμοι φασματικής συσταδοποίησης συγκαταλέγονται στις παραδοσιακές μεθόδους ανίχνευσης κοινοτήτων. Ας υποθέσουμε ότι έχουμε ένα σύνολο από n αντικείμενα $x_1, x_2..., x_n$ με μια συνάρτηση S που ορίζεται ανά ζεύγη, η οποία είναι συμμετρική και μη αρνητική (π.χ. $S(x_i, x_j) = S(x_j, x_i) \geq 0, \forall i, j = 1, ..n$). Στην περίπτωση των γράφων ο πίνακας γειτνίασης αποτελεί ένα παράδειγμα τέτοιας συνάρτησης αφού πληρεί τις παραπάνω ιδιότητες. Η *Φασματική συσταδοποίηση (spectral clustering)* περιλαμβάνει όλες τις μεθόδους και τεχνικές οι οποίες διαμερίζουν το σύνολο σε ομάδες χρησιμοποιώντας τις ιδιοτιμές πινάκων, όπως του πίνακα γειτνίασης ή άλλων πινάκων που παράγονται από αυτόν.

Συγκεκριμένα, τα αντικείμενα μπορεί να είναι σημεία σε κάποιο μετρικό χώρο ή κορυφές ενός γράφου. Η φασματική συσταδοποίηση πραγματοποιεί ένα μετασχηματισμό του αρχικού συνόλου αντικειμένων σε ένα σύνολο σημείων στο χώρο, των οποίων οι συντεταγμένες είναι στοιχεία των ιδιοδιανυσμάτων. Στη συνέχεια, το σύνολο των σημείων ταξινομείται σε ομάδες με τη χρήση γνωστών τεχνικών συσταδοποίησης, όπως η *συσταδοποίηση k-μέσων (k-means clustering)*. Η βασική ιδέα της φασματικής συσταδοποίησης έγκειται στο ότι η μείωση διαστάσεων που προκύπτει από τον παραπάνω μετασχηματισμό αποκαλύπτει την κοινοτική δομή του γράφου με μεγαλύτερη σαφήνεια [VonL06].

Ο πίνακας που χρησιμοποιείται κατά κύριο λόγο στη φασματική συσταδοποίηση είναι ο Λαπλασιανός πίνακας L. Έστω A ο πίνακας γειτνίασης ενός γράφου $G = (V, E)$ με n κορυφές και m ακμές και D ο διαγώνιος πίνακας που περιέχει τον βαθμό των κορυφών δηλαδή $D_i = deg(i),\ i = 1..n$. Ο μη-κανονικοποιημένος πίνακας Λαπλασιανός πίνακας L δίνεται από τη σχέση $L = D - A$. Ο (κανονικοποιημένος) Λαπλασιανός πίνακας Λ δίνεται από την εξίσωση $\Lambda = D^{-1/2}(D - A)D^{-1/2} = I - D^{-1/2}AD^{-1/2}$. Μπορεί να επαληθευτεί ότι οι πίνακες L και Λ είναι συμμετρικοί και θετικά ορισμένοι, επομένως έχουν πραγματικές και θετικές ιδιοτιμές [Chun97]. Ο Λαπλασιανός πίνακας έχει πάντα τουλάχιστον μια μηδενική ιδιοτιμή ενώ το πλήθος των μηδενικών ιδιοτιμών ισούται με το πλήθος των συνεκτικών συνιστωσών του γράφου.

Το βασικό μειονέκτημα των φασματικών αλγορίθμων είναι η υπολογιστική τους πολυπλοκότητα και γι' αυτό το λόγο πρακτικά είναι δύσκολη η εφαρμογή φασματικής ομαδοποίησης σε πολύ μεγάλα δίκτυα (της τάξης των εκατοντάδων χιλιάδων κόμβων) χωρίς τη χρήση παράλληλων αλγορίθμων.

3.2.5 Διαιρετικοί αλγόριθμοι

Ένας απλός τρόπος να χωριστεί ένας γράφος σε κοινότητες είναι να εντοπιστούν οι ακμές που συνδέουν κορυφές διαφορετικών κοινοτήτων και να αφαιρεθούν, έτσι ώστε οι ομάδες που εμπλέκονται να αποσυνδεθούν μεταξύ τους. Αυτή είναι η φιλοσοφία των διαιρετικών αλγορίθμων ιεραρχικής συσταδοποίησης. Το σημείο-κλειδί είναι η εύρεση μιας ιδιότητας των ακμών που συνδέουν διαφορετικές κοινότητες προκειμένου αυτές να μπορούν να εντοπιστούν. Η κύρια διαφορά των διαιρετικών αλγορίθμων με τους διαιρετικός αλγορίθμους ιεραρχικής συσταδοποίησης έγκειται στο ότι εδώ αφαιρούνται δια-ομαδικές ακμές αντί για ακμές μεταξύ κορυφών που παρουσιάζουν χαμηλή ομοιότητα.

Σε κάποιες περιπτώσεις είναι πιθανό να αφαιρούνται κορυφές (με όλες τις ακμές τους) ή ολόκληροι υπογράφοι, αντί για απλές ακμές.

Ο αλγόριθμος Girvan-Newman

Ο αλγόριθμος Girvan-Newman (ονομάστηκε έτσι από τους Michelle Girvan και Mark Newman) είναι ένας διαιρετικός αλγόριθμος που χρησιμοποιείται για την ανίχνευση κοινοτήτων σε πολύπλοκα συστήματα [Newm04c]. Ο αλγόριθμος αυτός ανιχνεύει κοινότητες αφαιρώντας προοδευτικά ακμές από το αρχικό δίκτυο. Τα συνδεδεμένα μέρη του εναπομείναντος δικτύου αποτελούν τις κοινότητες.

Ο αλγόριθμος Girvan-Newman χρησιμοποιεί την έννοια του *βαθμού ενδιαμεσότητας των ακμών (edge betweenness)* του γράφου σε αντιστοιχία με το *βαθμό ενδιαμεσότητας των κορυφών (vertex betweenness)* του Freeman [Free77]. Ο βαθμός ενδιαμεσότητας των κορυφών αναδεικνύει τους κεντρικούς κόμβους ενός δικτύου ως εξής : Για κάθε κόμβο n ενός δικτύου $G = (V, E)$ με n κόμβους και m ακμές, ο βαθμός ενδιαμεσότητας των κορυφών ορίζεται ως ο αριθμός των συντομότερων διαδρομών μεταξύ κάθε ζεύγους κόμβων του δικτύου που διέρχονται μέσω του κόμβου n. Ο αλγόριθμος Girvan-Newman επεκτείνει αυτό τον ορισμό στην περίπτωση των ακμών ορίζοντας το βαθμό ενδιαμεσότητας των ακμών για μια ακμή m ως τον αριθμό των συντομότερων διαδρομών μεταξύ κάθε ζεύγους κόμβων του δικτύου που διέρχονται μέσω της ακμής m. Στην περίπτωση που υπάρχει παραπάνω από μια σύντομη διαδρομή μεταξύ δύο κόμβων, τότε σε κάθε διαδρομή ανατίθεται τέτοιο βάρος ούτως ώστε το συνολικό βάρος των διαδρομών αυτών να αθροίζει στη μονάδα. Εάν ένα δίκτυο αποτελείται από κοινότητες οι οποίες συνδέονται μεταξύ τους με ένα μικρό αριθμό ακμών, τότε όλες οι συντομότερες διαδρομές μεταξύ των κοινοτήτων θα διέρχονται μέσω κάποιας εξ' αυτών των ακμών. Έτσι, οι ακμές που συνδέουν κοινότητες θα έχουν μεγάλο βαθμό ενδιαμεσότητας (τουλάχιστον μια από αυτές). Αφαιρώντας τις ακμές αυτές, οι ομάδες χωρίζονται η μία από την άλλη αποκαλύπτοντας την κοινοτική δομή που υπάρχει στο δίκτυο.Τα βήματα του αλγορίθμου συνοψίζονται παρακάτω:

1. Υπολογίζεται ο βαθμός ενδιαμεσότητας όλων των ακμών.

2. Η ακμή με το μεγαλύτερο βαθμό ενδιαμεσότητας αφαιρείται.

3. Ο βαθμός ενδιαμεσότητας όλων των ακμών που επηρεάζονται από την αφαίρεση επανυπολογίζεται.

4. Τα βήματα 2-3 επαναλαμβάνονται μέχρι να φτάσουμε στον επιθυμητό αριθμό κοινοτήτων.

Το γεγονός ότι ο βαθμός ενδιαμεσότητας υπολογίζεται μόνο στις ακμές οι οποίες επηρεάζονται από την αφαίρεση μιας ακμής μπορεί να επιταχύνει την εκτέλεση του αλγορίθμου στον υπολογιστή. Ωστόσο, ο βαθμός ενδιαμεσότητας των ακμών πρέπει να επανυπολογίζεται σε κάθε βήμα γιατί αλλιώς μπορεί να προκύψουν σοβαρά λάθη. Για παράδειγμα, αν δύο κοινότητες συνδέονται με περισσότερες από δύο ακμές δεν είναι σίγουρο ότι και οι δύο ακμές αυτές θα έχουν μεγάλο βαθμό ενδιαμεσότητας. Σύμφωνα με τον αλγόριθμο, γνωρίζουμε ότι τουλάχιστον μια από τις ακμές αυτές θα έχει μεγάλη τιμή, αλλά τίποτα περισσότερο. Επανυπολογίζοντας όμως το βαθμό ενδιαμεσότητας μετά την αφαίρεση της κάθε ακμής, βεβαιώνεται ότι τουλάχιστον μια από τις εναπομείνασες ακμές θα έχει μεγάλη τιμή ενδιαμεσότητας.

Ο υπολογισμός του βαθμού ενδιαμεσότητας των ακμών μπορεί να πραγματοποιηθεί με διάφορους τρόπους, που κατά βάση περιστρέφονται γύρω από την ίδια ιδέα. Εάν δύο κοινότητες συνδέονται με λίγες μόνο ακμές, τότε όλες οι διαδρομές του δικτύου, από κόμβους της μιας κοινότητας προς κόμβους της άλλης, θα πρέπει να διέρχονται μέσω μιας εξ' αυτών των ακμών. Δεδομένου ενός συνόλου διαδρομών, μετρώντας πόσες από αυτές διέρχονται μέσω κάθε ακμής του γράφου, αναμένεται οι ακμές μεταξύ των κοινοτήτων να εμφανίζονται περισσότερες φορές, δίνοντας έτσι μια μέθοδο για τον εντοπισμό των κοινοτήτων αυτών. Οι συγγραφείς αναφέρουν τρεις εναλλακτικούς ορισμούς του βαθμού ενδιαμεσότητας ακμών :

a) Ο *γεωδαιτικός βαθμός ενδιαμεσότητας ακμών (shortest-path betweenness)*: Πρόκειται για το απλούστερο παράδειγμα υπολογισμού του βαθμού ενδιαμεσότητας ακμών, το οποίο βασίζεται στις συντομότερες (γεωδαιτικές) διαδρομές. Αρχικά, υπολογίζονται οι συντομότερες διαδρομές μεταξύ κάθε ζεύγους κόμβων του δικτύου και έπειτα, για κάθε διαδρομή, μετριέται πόσες φορές προσπελάζεται η κάθε ακμή.

b) *Ο βαθμός ενδιαμεσότητας τυχαίου περιπάτου (random-walk betweenness)*: Στην περίπτωση αυτή ο βαθμός ενδιαμεσότητας μιας ακμής ισούται με τη συχνότητα των περασμάτων από την ακμή αυτή ενός τυχαίου περιπατητή ο οποίος κινείται στο γράφο. Για τον υπολογισμό της επιλέγεται τυχαία ένα ζεύγος κορυφών s (αφετηρία) και t (στόχος) και θεωρούμε έναν τυχαίο περιπατητή ο οποίος ξεκινάει από την κορυφή s και κινείται σε κάθε γειτονική ακμή με την ίδια πιθανότητα έως ότου φτάσει στην κορυφή t, όπου σταματάει. Ο τελικός βαθμός ενδιαμεσότητας προκύπτει από το μέσο όρο των πιθανοτήτων διάσχισης κάθε ακμής για κάθε ζεύγος αφετηρίας-στόχου s-t.

c) *Ο βαθμός ενδιαμεσότητας ροής ρεύματος (current-flow betweenness)*: Για τον υπολογισμό του θεωρούμε το γράφο σαν ένα ηλεκτρικό κύκλωμα με τις ακμές να έχουν μοναδιαία τιμή αντίστασης. Εφαρμόζοντας διαφορά τάσης μεταξύ δύο κορυφών, κάθε ακμή διαρρέεται από κάποια ποσότητα ρεύματος. Ο μέσος όρος της απόλυτης τιμής των ρευμάτων που διαρρέουν κάθε ακμή, αν επαναλάβουμε τη διαδικασία για κάθε ζεύγος κορυφών, ισούται με το βαθμό ενδιαμεσότητας ροής ρεύματος.

Στον αλγόριθμο των Girvan-Newman η επιλογή του μέτρου του βαθμού ενδιαμεσότητας δεν είναι ιδιαίτερα κρίσιμη, αφού οι διαφορετικές προσεγγίσεις του βαθμού ενδιαμεσότητας οδηγούν σε παρόμοιες κοινοτικές δομές.

Παρότι ο αλγόριθμος φαίνεται να επιστρέφει καλά αποτελέσματα σε αρκετές περιπτώσεις, υπάρχουν δύο πρωταρχικά μειονεκτήματα. Το πρώτο είναι ότι δεν παρέχει σαφή καθοδήγηση σχετικά με τον αριθμό των κοινοτήτων στις οποίες πρέπει να διαιρεθεί ένα δίκτυο. Για να επιλύσουν αυτό το πρόβλημα, οι συγγραφείς του αλγορίθμου πρότειναν ότι η διαίρεση ενός δικτύου μπορεί να αξιολογηθεί με τη χρήση της τμηματικότητας, η οποία παρουσιάστηκε στην Ενότητα 3.2.1. Στην αλγεβρική της μορφή, για μια διαίρεση ενός γράφου G σε g κοινότητες, ορίζεται ένας πίνακας e μεγέθους $g \times g$ του οποίου το στοιχείο e_{ij} είναι το κλάσμα των ακμών του αρχικού δικτύου (όπως ήταν πριν την αφαίρεση ακμών) οι οποίες συνδέουν κόμβους που ανήκουν στην κοινότητα i με κόμβους που ανήκουν στην κοινότητα j. Έτσι η τιμή της τμηματικότητας μπορεί να πάρει την παρακάτω μορφή :

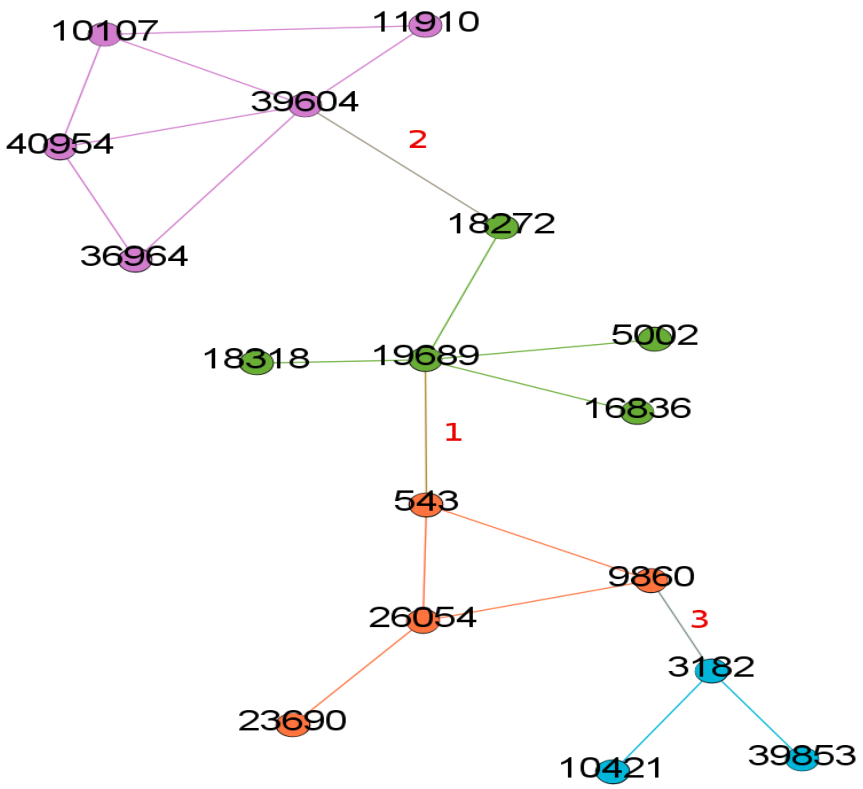
$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki} = Tre - \|e^2\| \tag{3.10}$$

, όπου το $\|e\|$ συμβολίζει το άθροισμα όλων των στοιχείων του e. Το Q πρακτικά είναι το κλάσμα των ακμών που βρίσκονται εσωτερικά των κοινοτήτων μείον την αναμενόμενη τιμή της ίδιας ποσότητας σε έναν γράφο με κορυφές ίδιου βαθμού στον οποίο οι ακμές είναι τυχαία τοποθετημένες. Μια τιμή $Q = 0$ υποδηλώνει ότι η κοινοτική δομή δεν είναι ισχυρότερη από αυτή που θα αναμενόταν σε περίπτωση τυχαίας τοποθέτησης των ακμών ενώ όσο πιο κοντά στη μονάδα είναι η τιμή της τόσο ισχυρότερη θεωρείται η κοινοτική δομή του γράφου. Στην πράξη οι τιμές συνήθως κυμαίνονται στο εύρος από 0.3 έως 0.7 ενώ υψηλότερες τιμές είναι σπάνιες.

Το άλλο κύριο μειονέκτημα του αλγορίθμου είναι το γεγονός ότι είναι αργός. Δεδομένου ότι πρέπει να αφαιρεθούν m ακμές συνολικά και κάθε επανάληψη του αλγορίθμου απαιτεί χρόνο (mn), τότε στη χειρότερη περίπτωση απαιτείται χρόνος $O(m^2 n)$ ή (n^3) για αραιούς γράφους ($m \simeq n$). Για την αντιμετώπιση της αργής ταχύτητας του αλγορίθμου ένας αριθμός συγγραφέων έχουν προτείνει παραλλαγές του βασικού αλγορίθμου όπως οι αλγόριθμοι του Tyler et al. [Tyle05] και Radicchi [Radi04].

Στο Σχήμα 3.5 απεικονίζεται ένα παράδειγμα εφαρμογής του αλγορίθμου Girvan-Newman σε μια από τις συνεκτικές συνιστώσες του γράφου των χρηστών του συνόλου δεδομένων που χρησιμοποιήθηκε στο πειραματικό μέρος της εργασίας. Όπως φαίνεται στο εν λόγω σχήμα ο αλγόριθμος διαμέρισε

το γράφο σε τέσσερις κοινότητες (σημειωμένες με διαφορετικό χρώμα η καθεμία) αφαιρώντας διαδοχικά τις τρεις ακμές οι οποίες είναι αριθμημένες με κόκκινο χρώμα.



Σχήμα 3.5: Παράδειγμα εφαρμογής του αλγορίθμου Girvan-Newman σε μια συνεκτική συνιστώσα του γράφου με 17 κόμβους

3.2.6 Αλγόριθμοι βασισμένοι στην τμηματικότητα

Η τμηματικότητα των Newman-Girvan, η οποία αρχικά παρουσιάστηκε ως ένα κριτήριο τερματισμού για τον αλγόριθμο των δύο συγγραφέων, εξελίχθηκε γρήγορα σε απαραίτητο στοιχείο πολλών αλγορίθμων ομαδοποίησης. Η τμηματικότητα είναι μακράν η πιο γνωστή και ευρέως χρησιμοποιημένη συνάρτηση ποιότητας [Fort10]. Ήταν επίσης η πρώτη προσπάθεια μιας κατ' αρχήν κατανόησης του προβλήματος της ομαδοποίησης και ενσωματώνει στη συμπαγή μορφή της όλα τα απαραίτητα συστατικά και ορισμούς, από τον ορισμό της κοινότητας, την επιλογή ενός μοντέλου αναφοράς, έως την έννοια της "δυνατής" κοινότητας και διαμέρισης.

Ο Newman πρότεινε έναν άπληστο σωρευτικό αλγόριθμο ιεραρχικής συσταδοποίησης ο οποίος βασίζεται στην αύξηση της τμηματικότητας [Newm04b]. Η βασική ιδέα του αλγορίθμου είναι η εξής: Ξεκινώντας από μια αρχική κατάσταση στην οποία κάθε κορυφή του γράφου είναι το μοναδικό μέλος μιας από τις n κοινότητες, οι κοινότητες συγχωνεύονται επαναληπτικά σε ζεύγη επιλέγοντας κάθε φορά τη συγχώνευση η οποία επιφέρει τη μεγαλύτερη αύξηση (ή μικρότερη μείωση) της τμηματικότητας. Στη διαδικασία αυτή δε λαμβάνονται υπόψη ζεύγη κορυφών οι οποίες δε συνδέονται μεταξύ τους, καθώς η συγχώνευση των ομάδων αυτών δε μπορεί να αυξήσει την τμηματικότητα. Επομένως ο μέγιστος αριθμός ζευγών που θα πρέπει να θεωρηθούν ισούται με m, δηλαδή τον αριθμό των ακμών του γράφου. Αν συνυπολογιστεί ότι μετά από κάθε συγχώνευση πρέπει να ενημερωθεί ο πίνακας e_{ij} που χρειάζεται για τον υπολογισμό της τμηματικότητας, το οποίο χρειάζεται στο χειρότερο σενάριο χρόνο $O(n)$, και ότι ο αλγόριθμος χρειάζεται συνολικά $n-1$ επαναλήψεις (συγχωνεύσεις κοινοτήτων) για να ολοκληρωθεί, τότε είναι εύκολο να καταλάβει κανείς ότι η πολυπλοκότητα του αλγορίθμου εί-

ναι $O((m+n)n)$ γενικά ή $O(n^2)$ για αραιούς γράφους (όπου ισχύει $m \simeq n$). Οι Clauset κ.α. [Clau04] βελτίωσαν την πολυπλοκότητα του αλγορίθμου χρησιμοποιώντας αποδοτικές δομές δεδομένων όπως σωρούς μεγίστων.

3.3 Τα κοινωνικά δίκτυα στα συστήματα συστάσεων

Η γενικευμένη χρήση των μέσων κοινωνικής δικτύωσης τα τελευταία χρόνια παράγει *κοινωνική πληροφορία (social information)* με ταχύτατους ρυθμούς. Αυτό έχει ως αποτέλεσμα την ταχεία ανάπτυξη των *συστημάτων κοινωνικών συστάσεων (social recommender systems)*, δηλαδή των συστημάτων τα οποία χρησιμοποιούν την κοινωνική πληροφορία για την παράγωγη συστάσεων. Η κοινωνική σύσταση μελετάται από το 1997 [Kaut97] και προσελκύει όλο και περισσότερο ενδιαφέρον με τη ραγδαία ανάπτυξη των κοινωνικών δικτύων.

Τα συστήματα συστάσεων επωφελούνται από τα κοινωνικά δίκτυα και αντιστρόφως σε μεγάλο βαθμό. Από τη μια μεριά τα κοινωνικά δίκτυα εισάγουν νέους τύπους δεδομένων και μετα-δεδομένων (metadata) στα συστήματα συστάσεων όπως για παράδειγμα επισημάνσεις (tags), σχόλια, ψήφους και γενικότερα εμφανείς ή υπονοούμενες κοινωνικές σχέσεις. Από την άλλη, τα συστήματα συστάσεων έχουν μεγάλο αντίκτυπο στην επιτυχία των κοινωνικών συστημάτων, βεβαιώνοντας ότι οι χρήστες έχουν συνέχεια στη διάθεσή τους αντικείμενα σχετικά με τα ενδιαφέροντα τους και τις ανάγκες τους.

3.3.1 Ορισμοί της κοινωνικής σύστασης

Η κοινωνική σύσταση δε διαθέτει κάποιον κοινώς αποδεκτό ορισμό. Για το λόγο αυτό παρουσιάζονται παρακάτω δύο ορισμοί, ένας στενός και ένας πιο ευρύς.

Ορισμός 3.3.1.1. Ο στενός ορισμός της κοινωνικής σύστασης την ορίζει ως οποιαδήποτε σύσταση η οποία χρησιμοποιεί τις κοινωνικές σχέσεις σαν επιπλέον είσοδο. Οι σχέσεις αυτές μπορεί να περιλαμβάνουν οποιαδήποτε μορφή αλληλεξάρτησης όπως εμπιστοσύνη, φιλία, συγγένεια, γνωριμία, επικοινωνία, οικονομικές συναλλαγές, σχέσεις συμμετοχής, αντιπάθεια κ.α.

Σε αυτό τον ορισμό τα κοινωνικά συστήματα συστάσεων υποθέτουν ότι οι χρήστες συσχετίζονται όταν τους συνδέουν κοινωνικές σχέσεις. Για παράδειγμα, οι προτιμήσεις των χρηστών είναι πιθανό να επηρεάζονται ή να είναι παρόμοιες με αυτές των φίλων τους σε ένα κοινωνικό δίκτυο. Κάνοντας αυτή την υπόθεση, η κοινωνική σύσταση αξιοποιεί τις συσχετίσεις μεταξύ των χρηστών, οι οποίες υποδηλώνονται από τις κοινωνικές σχέσεις, προκειμένου να βελτιώσει την ποιότητα της σύστασης.

Ορισμός 3.3.1.2. Ο ευρύς ορισμός [Guy15] της κοινωνικής σύστασης αναφέρει ότι κοινωνικό σύστημα συστάσεων είναι κάθε σύστημα συστάσεων που στοχεύει σε *τομείς (domains)* κοινωνικών μέσων. Αυτός ο ορισμός θεωρεί συστήματα κοινωνικής σύστασης εκείνα των οποίων οι συστάσεις αφορούν οντότητες κοινωνικών δικτύων όπως για παράδειγμα ανθρώπους, ετικέτες (tags), αντικείμενα, κοινότητες, συνδέσεις (links) κ.α.

Οι πηγές που χρησιμοποιούν τα συστήματα αυτά περιλαμβάνουν, εκτός από online κοινωνικές σχέσεις, διάφορους τύπους δεδομένων από κοινωνικά δίκτυα όπως *κοινωνική επισήμανση (social tagging)*, *συμπεριφορά επιλογών (click behaviours)* και άλλες αλληλεπιδράσεις μεταξύ των χρηστών.

Στην ενότητα αυτή γίνεται εστίαση σε συστήματα κοινωνικής σύστασης τα οποία ανταποκρίνονται στον πρώτο ορισμό μιας και το σύστημα που υλοποιήθηκε στην πειραματική διαδικασία ανήκει σε αυτή την κατηγορία συστημάτων. Ωστόσο, παρόμοιες τεχνικές όπως αυτές που περιγράφουμε παρακάτω μπορούν να εφαρμοστούν και σε συστήματα που αφορούν το δεύτερο ορισμό.

3.3.2 Βασική ιδέα της κοινωνικής σύστασης

Η βασική ιδέα όσον αφορά τα συστήματα κοινωνικής σύστασης είναι ότι οι κοινωνικές σχέσεις ενός χρήστη έχουν μεγάλο αντίκτυπο στις προτιμήσεις, τις επιλογές και την καταναλωτική συμπεριφορά του [Weng10]. Οι χρήστες πολύ συχνά συμβουλεύονται τους φίλους τους για προτάσεις σχετικά

με ταινίες, ταξίδια και καταναλωτικά προϊόντα. Αυτή η ομοιότητα στις προτιμήσεις των χρηστών εξηγείται με τις έννοιες της *ομοφιλίας (homophily)* και της *εμπιστοσύνης (trust)*, δύο έννοιες οι οποίες σχετίζονται μεταξύ τους αλλά και διαφέρουν.

Η ομοφιλία αναφέρεται στο γεγονός ότι συνδεδεμένοι χρήστες σε κοινωνικά δίκτυα είναι πιθανότερο να να συμπεριφέρονται με παρόμοιο τρόπο όσον αφορά προτιμήσεις και ενδιαφέροντα. Από την άλλη, η εμπιστοσύνη αναφέρεται στο ότι οι χρήστες είναι πιθανότερο να εμπιστευτούν προτιμήσεις και συστάσεις των συνδεδεμένων με αυτούς χρηστών. Σε πολλές περιπτώσεις, η εμπιστοσύνη αποτελεί συνέπεια της ομοφιλίας αφού, από τη στιγμή που οι συνδεδεμένοι χρήστες τείνουν να συμπεριφέρονται παρόμοια, τείνουν να εμπιστεύονται τα ενδιαφέροντα και τις προτιμήσεις ο ένας του άλλου. Η ισχυρή συσχέτιση μεταξύ της ομοφιλίας και της εμπιστοσύνης έχει μελετηθεί στα [Golb09],[Zieg07].

Για την ενσωμάτωση των παραπάνω εννοιών στα κοινωνικά συστήματα συστάσεων, εκτός από το γνωστό $m \times n$ πίνακα βαθμολογιών R, όπου m ο αριθμός των χρηστών και n των αντικειμένων, χρησιμοποιείται επίσης και ένας πίνακας χρήστη-χρήστη $T \in \mathbb{R}^{m \times m}$ του οποίου οι τιμές περιγράφουν τις κοινωνικές συνδέσεις μεταξύ των χρηστών. Το εύρος τιμών του πίνακα T μπορεί να διαφέρει ανάλογα με την εκάστοτε εφαρμογή. Δηλαδή σε κάποιες περιπτώσεις αυτός λαμβάνει τις τιμές $T_{ij} = 1$ αν ο χρήστης i συνδέεται με το χρήστη j και $T_{ij} = 0$ διαφορετικά, αλλά μπορεί επίσης το πεδίο τιμών του να είναι το [0,1] ή [-1,1] με την τιμή T_{ij} να εκφράζει το βαθμό της εμπιστοσύνης (ή μη εμπιστοσύνης αν η τιμή είναι αρνητική) του χρήστη i προς το χρήστη j.

3.3.3 Υφιστάμενα συστήματα κοινωνικής σύστασης

Όπως αναφέρθηκε στην Ενότητα 2.2 τα συστήματα συνεργατικής διήθησης είναι τα πλέον διαδεδομένα συστήματα συστάσεων και έτσι τα περισσότερα κοινωνικά συστήματα συστάσεων βασίζονται και αυτά σε τεχνικές συνεργατικής διήθησης. Στην πλειοψηφία των περιπτώσεων τα συστήματα κοινωνικής σύστασης χρησιμοποιούν κάποιο σύστημα συνεργατικής διήθησης ως το βασικό αλγόριθμο για την παραγωγή συστάσεων και προτείνουν προσεγγίσεις για τη σύλληψη της κοινωνικής πληροφορίας οι οποίες βασίζονται στην ανάλυση κοινωνικών δικτύων. Επομένως, στη γενική του μορφή ένα σύστημα κοινωνικής σύστασης το οποίο βασίζεται στην συνεργατική διήθηση έχει την παρακάτω μορφή:

$$\begin{aligned}&\text{Σύστημα κοινωνικής σύστασης βασισμένο στη συνεργατική διήθηση =}\\&\text{Βασικό σύστημα συνεργατικής διήθησης + Μοντέλο κοινωνικής πληροφορίας}\end{aligned} \tag{3.11}$$

Τα κοινωνικά συστήματα συστάσεων βασισμένα στη συνεργατική διήθηση, όπως και τα κανονικά, μπορούν να χωριστούν σε βασισμένα στη μνήμη και στο μοντέλο, ανάλογα με τι τύπου είναι το βασικό σύστημα συνεργατική διήθησης που χρησιμοποιούν.

Κοινωνικά συστήματα συστάσεων βασισμένα στη μνήμη

Τα κοινωνικά συστήματα συστάσεων που βασίζονται στη μνήμη (ή γειτονιά) χρησιμοποιούν μεθόδους συνεργατικής διήθησης βασισμένες στο χρήστη, οι οποίες περιγράφηκαν στην αντίστοιχη παράγραφο της Ενότητας 2.2.1. Η διαφορά τους από τα συστήματα αυτά έγκειται στο ότι η γειτονιά του κάθε χρήστη δεν υπολογίζεται μόνο με βάση τις αξιολογήσεις του αλλά και βάσει των συνδέσεων του χρήστη στο κοινωνικό δίκτυο στο οποίο ανήκει, δηλαδή της κοινωνικής πληροφορίας του. Επομένως, τα κοινωνικά συστήματα συστάσεων της κατηγορίας αυτής διαφέρουν από τα κλασσικά συστήματα ως προς τον υπολογισμό της γειτονιάς και της ομοιότητας των χρηστών, ενώ ακολουθούν την ίδια λογική κατά τον προσδιορισμό των ελλιπών βαθμολογιών, κάνοντας χρήση των συναρτήσεων πρόβλεψης οι οποίες αναλύθηκαν στη σχετική παράγραφο της Ενότητας 2.2.1. Παρακάτω παρουσιάζονται ορισμένα παραδείγματα τέτοιων συστημάτων με τα βασικά χαρακτηριστικά τους.

Απλό κοινωνικό σύστημα σύστασης

Το πιο απλό σύστημα της κατηγορίας αυτής λαμβάνει ως γειτονιά N_u του χρήστη u όλους τους χρήστες οι οποίοι συνδέονται άμεσα με τον u με κάποιου είδους κοινωνική σχέση (φιλία, εμπιστοσύνη κλπ). Επομένως εδώ δεν περιλαμβάνεται κάποιος υπολογισμός ομοιότητας αλλά η πρόβλεψη της βαθμολογίας του χρήστη u για ένα αντικείμενο j υπολογίζεται ως ο μέσος όρος των βαθμολογιών των συνδεδεμένων με αυτόν χρηστών. Επομένως, εάν $N_u(j)$ το σύνολο των γειτονικών χρηστών του χρήστη u οι οποίοι έχουν προσδιορίσει βαθμολογία για το αντικείμενο j, τότε η προβλεπόμενη βαθμολογία του χρήστη u για το αντικείμενο j ορίζεται ως εξής:

$$\widehat{r_{uj}} = \frac{\sum\limits_{k \in N_u(j)} r_{kj}}{|N_u(j)|} \tag{3.12}$$

για τον απλό μέσο όρο, ή όπως παρακάτω:

$$\widehat{r_{uj}} = \mu_u + \frac{\sum\limits_{k \in N_u(j)} (r_{kj} - \mu_k)}{|N_u(j)|} \tag{3.13}$$

για τον κεντραρισμένο στη μέση τιμή μέσο όρο όπου μ_k η μέση τιμή των βαθμολογιών κάθε χρήστη k (τύπος του Resnick).

Συστήματα που βασίζονται σε δίκτυα εμπιστοσύνης

Μια άλλη προσέγγιση μπορεί να υπάρξει όταν είναι διαθέσιμο ένα δίκτυο εμπιστοσύνης, δηλαδή εκτός από τις συνδέσεις μεταξύ των χρηστών, έχει προσδιοριστεί και μια τιμή για την εμπιστοσύνη που έχει (ή δεν έχει) ένας χρήστης στους χρήστες του κοινωνικού δικτύου με τους οποίους συνδέεται. Χρησιμοποιώντας τον πίνακα εμπιστοσύνης $T = [t_{ij}]$, ο οποίος περιγράφηκε στην Ενότητα 3.3.2, η τιμή t_{ij} εκφράζει την εμπιστοσύνη του χρήστη i στο χρήστη j. Σε αυτή την περίπτωση, η προβλεπόμενη βαθμολογία $\widehat{r}_{uj}$ του χρήστη u για το αντικείμενο j προκύπτει από το μέσο όρο των βαθμολογιών των χρηστών οι οποίοι έχουν βαθμολογήσει το αντικείμενο j, για τους οποίους ο χρήστης u έχει μια τιμή εμπιστοσύνης μεγαλύτερη ενός δοσμένου κατωφλίου θ. Επομένως, η συνάρτηση πρόβλεψης θα έχει την παρακάτω μορφή:

$$\widehat{r_{uj}} = \frac{\sum\limits_{k \in N_u(j,\theta)} t_{uk} r_{kj}}{\sum\limits_{k \in N_u(j,\theta)} t_{uk}} \tag{3.14}$$

Αυτή η προσέγγιση επομένως μπορεί να θεωρηθεί ως μια βασισμένη στο χρήστη μέθοδος στην οποία οι τιμές εμπιστοσύνης χρησιμοποιούνται αντί για κάποιο συντελεστή ομοιότητας (από αυτούς που παρουσιάστηκαν στην Ενότητα 2.2.1). Ο τύπος αυτός είναι γνωστός και ως *σταθμισμένος μέσος όρος εμπιστοσύνης (trust weighted mean)* [Agga16]. Προφανώς ο τύπος μπορεί να τροποποιηθεί όπως και παραπάνω ώστε οι βαθμολογίες να κεντραριστούν στη μέση τιμή και επομένως να λάβει την παρακάτω μορφή:

$$\widehat{r_{uj}} = \mu_u + \frac{\sum\limits_{k \in N_u(j,\theta)} t_{uk} (r_{kj} - \mu_k)}{\sum\limits_{k \in N_u(j,\theta)} t_{uk}} \tag{3.15}$$

Ωστόσο, τα δίκτυα εμπιστοσύνης είναι πολλές φορές αραιά, δηλαδή δεν είναι απαραίτητο ότι όλα τα ζεύγη χρηστών έχουν προσδιορισμένες τιμές εμπιστοσύνης μεταξύ τους. Για το λόγο αυτό τα κοινωνικά συστήματα συστάσεων χρησιμοποιούν την έννοια της *μεταβατικότητας (transitivity)* των σχέσεων εμπιστοσύνης, δηλαδή της δυνατότητας να συμπεραίνονται οι απροσδιόριστες τιμές εμπιστοσύνης μέσω των τελεστών της *διάδοσης (propagation)* και της *συνάθροισης (aggregation)*.

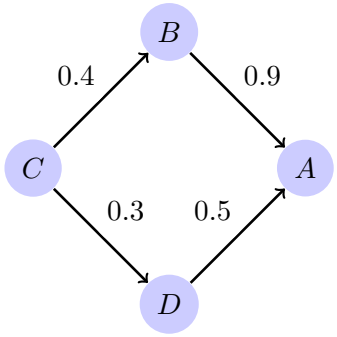
Για να γίνει κατανοητή η έννοια της μεταβατικότητας έστω ότι έχουμε τρεις χρήστες A,B και C. Στην περίπτωση όπου ο B εμπιστεύεται τον A και ο C εμπιστεύεται τον B, μπορεί κανείς να συμπεράνει ότι ο C εμπιστεύεται τον A και επομένως να κάνει συστάσεις στον C με βάση τις προτιμήσεις του A. Με άλλα λόγια είναι απαραίτητο να καθοριστούν μονοπάτια στο δίκτυο εμπιστοσύνης προκειμένου να συμπεράνει κάποιος τις ελλιπείς τιμές εμπιστοσύνης. Ο προσδιορισμός της άγνωστης τιμής εμπιστοσύνης μεταξύ δύο κόμβων στο τέλος ενός μονοπατιού αναφέρεται ως *διάδοση εμπιστοσύνης (trust propagation)*. Παρόλα αυτά, συνήθως υπάρχουν περισσότερα από ένα μονοπάτια μεταξύ δύο χρηστών σε ένα δίκτυο εμπιστοσύνης. Για παράδειγμα έστω το απλό δίκτυο εμπιστοσύνης του Σχήματος 3.6 όπου το βάρος μιας κατευθυνόμενης ακμής από ένα χρήστη u σε ένα χρήστη v δείχνει κατά πόσο ο u εμπιστεύεται τον v. Στο δίκτυο αυτό υπάρχουν δύο μονοπάτια μεταξύ των χρηστών C και A, ένα μέσω του B και ένα μέσω του D, και έτσι οι διαδιδόμενες τιμές εμπιστοσύνης πρέπει να συναθροιστούν πάνω στα δύο μονοπάτια. Οι τελεστές της διάδοσης και της συνάθροισης ορίζονται όπως περιγράφεται παρακάτω:

- *Διάδοση εμπιστοσύνης κατά μήκος ενός μονοπατιού:* Για τον υπολογισμό της τιμής εμπιστοσύνης μεταξύ δύο κόμβων οι οποίοι δε συνδέονται άμεσα με ακμή αλλά ανάμεσά τους παρεμβάλλονται ένας ή περισσότεροι άλλοι κόμβοι, οι τιμές εμπιστοσύνης των ενδιάμεσων ακμών πολλαπλασιάζονται και το αποτέλεσμα που προκύπτει θεωρείται ως η διαδιδόμενη τιμή εμπιστοσύνης μεταξύ του αρχικού και τελικού κόμβου του μονοπατιού [Guha04]. Για παράδειγμα, στο Σχήμα 3.6, θεωρώντας το μονοπάτι $C \rightarrow B \rightarrow A$, πολλαπλασιάζοντας τις τιμές εμπιστοσύνης κατά μήκος του μονοπατιού προκύπτει ότι η τιμή εμπιστοσύνης του χρήστη C στον A ισούται με $0.4 \times 0.9 = 0.36$. Πολλές φορές χρησιμοποιείται και ένας συντελεστής απόσβεσης $\beta \leq 1$ για να μειώσει την τιμή της εμπιστοσύνης σε μεγάλα μονοπάτια, πολλαπλασιάζοντας την τελική τιμή με β^l, όπου l το μήκος του μονοπατιού. Έτσι στο παραπάνω παράδειγμα, μιας και το μήκος του μονοπατιού ισούται με 2, η τελική τιμή της διαδεδομένης εμπιστοσύνης θα ήταν $0.36 \times \beta^l$.

- *Συνάθροιση εμπιστοσύνης μεταξύ πολλαπλών μονοπατιών:* Από τη στιγμή που ανάμεσα σε δύο κόμβους μπορεί να υπάρχουν περισσότερα από ένα μονοπάτια, οι διαδιδόμενες τιμές των μονοπατιών αυτών πρέπει να συναθροιστούν ώστε να προκύψει μια μοναδική τιμή εμπιστοσύνης. Υπάρχουν διάφορες εναλλακτικές μέθοδοι για τον τελεστή της συνάθροισης όπως η ελάχιστη τιμή, η μέγιστη τιμή, η μέση τιμή, η σταθμισμένη μέση τιμή και το σταθμισμένο άθροισμα. Στο σταθμισμένο μέσο όρο και στο σταθμισμένο άθροισμα κάποια μονοπάτια, όπως για παράδειγμα μικρότερα μονοπάτια, θεωρούνται πιο σημαντικά από άλλα. Αυτή η στάθμιση μπορεί να επιτευχθεί προφανώς και με τη χρήση του συντελεστή απόσβεσης κατά τη διάδοση της εμπιστοσύνης στα επιμέρους μονοπάτια. Η τελική τιμή της συνάθροισης μπορεί να διαφέρει αρκετά ανάλογα με τη μέθοδο που χρησιμοποιείται. Για παράδειγμα, για τον υπολογισμό της εμπιστοσύνης του χρήστη C στον A στο δίκτυο του Σχήματος 3.6, χρησιμοποιώντας τη μέγιστη τιμή έχουμε την εκτίμηση $\max(0.36, 0.15) = 0.36$ ενώ με το μέσο όρο έχουμε $(0.36 + 0.15)/2 = 0.255$.

Από τη στιγμή που έχουν υπολογιστεί οι ελλιπείς τιμές εμπιστοσύνης μεταξύ των χρηστών με τη βοήθεια της διάδοσης και της συνάθροισης, επιλέγονται οι χρήστες με τιμή εμπιστοσύνης μεγαλύτερης ενός δοσμένου κατωφλίου θ και οι προβλεπόμενες βαθμολογίες υπολογίζονται από τις Εξισώσεις 3.14 και 3.15.

Επομένως οι παραπάνω έννοιες αποτελούν τη βάση όλων των κοινωνικών συστημάτων συστάσεων που χρησιμοποιούν δίκτυα εμπιστοσύνης. Από εκεί και πέρα, έχουν προταθεί διάφοροι αλγόριθμοι εμπλουτισμένοι και με άλλα εργαλεία της θεωρίας δικτύων όπως π.χ. τυχαίους περιπάτους. Οι πιο γνωστοί από αυτούς είναι οι αλγόριθμοι *Tidaltrust* [Golb06], *MoleTrust* [Mass04] και *TrustWalker* [Jama09].



Σχήμα 3.6: Απλό δίκτυο εμπιστοσύνης

Συστήματα που βασίζονται στην ανάλυση κοινωνικών δικτύων

Τα συστήματα αυτά αξιοποιούν το δίκτυο των χρηστών προκειμένου να υπολογίσουν την ομοιότητα μεταξύ τους χρησιμοποιώντας γνωστές μετρικές της θεωρίας γραφημάτων όπως ο συντελεστής ομοιότητας Jaccard [Jacc01], ο συντελεστής των Adamic-Adar [Adam03] και ο συντελεστής κεντρικότητας του Katz [Katz53]. Για παράδειγμα, έστω ένα σύστημα το οποίο χρησιμοποιεί το συντελεστή ομοιότητας Jaccard. Για δύο χρήστες u και v, με N_u και N_v να είναι τα σύνολα με τους γειτονικούς κόμβους των δύο χρηστών αντίστοιχα, η ομοιότητα μεταξύ των χρηστών αυτών υπολογίζεται από τον παρακάτω τύπο:

$$\text{Jaccard(u,v)} = \frac{|N_u \cap N_v|}{|N_u \cup N_v|} \tag{3.16}$$

Ο συντελεστής Jaccard προφανώς λαμβάνει τιμές στο κλειστό διάστημα [0,1] αφού από τη θεωρία συνόλων ισχύει ότι $|A \cap B| \leq |A \cup B|$. Ουσιαστικά ο συντελεστής αυτός, στη συγκεκριμένη εφαρμογή, εκφράζει το ποσοστό των κοινών γειτόνων μεταξύ δύο χρηστών και θα έχει την τιμή 1 στην περίπτωση που όλοι οι γείτονες των δύο χρηστών είναι κοινοί και την τιμή 0 αν οι δύο χρήστες δεν έχουν κανένα κοινό γείτονα.

Ο συντελεστής των Adamic-Adar αποτελεί μια παρόμοια μετρική με το συντελεστή Jaccard, η οποία όμως δίνει έμφαση στους κοινούς γείτονες οι οποίοι έχουν μικρότερο βαθμό (degree) στο δίκτυο, δηλαδή μικρότερο αριθμό γειτόνων. Ο συντελεστής ορίζεται σύμφωνα με τον τύπο που ακολουθεί:

$$\text{Adamic-Adar(u,v)} = \sum_{z \in N(u) \cap N(v)} \frac{1}{\log |N(z)|} \tag{3.17}$$

Ο συντελεστής αυτός είναι προφανώς θετικός όπως φαίνεται στην Εξίσωση 3.17 ωστόσο δεν είναι πάντα μικρότερος της μονάδας. Για αυτό το λόγο και επειδή στα συστήματα συστάσεων συνεργατικής διήθησης ορίζεται συνήθως ένα κατώφλι ομοιότητας ο συντελεστής των Adamic-Adar μπορεί να κανονικοποιηθεί ώστε να λαμβάνει τιμές στο κλειστό διάστημα [0,1]. Η κανονικοποιημένη μορφή του συντελεστή ορίζεται ως εξής:

$$\text{Normalized Adamic-Adar(u,v)} = \frac{1}{|N(u) \cap N(v)|} \sum_{z \in N(u) \cap N(v)} \frac{\log 2}{\log |N(z)|} \tag{3.18}$$

Με τη νέα αυτή μορφή του συντελεστή, στην περίπτωση που δύο χρήστες u,v έχουν μόνο ένα κοινό χρήστη z και οι μοναδικοί γείτονες του z είναι οι χρήστες u και v (άρα $|N(z)| = 2$) ο συντελεστής αυτός θα ισούται με τη μονάδα ενώ σε οποιαδήποτε άλλη περίπτωση θα είναι μικρότερος αυτής. Προφανώς εάν οι χρήστες u,v δεν έχουν κανένα κοινό γείτονα ο συντελεστής θα ισούται με το μηδέν.

Εκτός από τους συντελεστές ομοιότητας Jaccard και Adamic-Adar μπορούν να χρησιμοποιηθούν και άλλες μετρικές που χρησιμοποιούνται στη θεωρία δικτύων όπως είναι η μετρική κεντρικότητας

του Katz η οποία παρουσιάστηκε στην Ενότητα 2.2.3 και υπολογίζεται ανάμεσα σε δύο κόμβους-χρήστες από την Εξίσωση 2.18 ή για κάθε ζεύγος κόμβων με τη χρήση του πίνακα γειτνίασης όπως φαίνεται στην Εξίσωση 2.19.

Ένα άλλο κομμάτι της ανάλυσης κοινωνικών δικτύων το οποίο μπορεί να βρει εφαρμογή σε κοινωνικά συστήματα συστάσεων είναι η ανίχνευση κοινοτήτων, η οποία παρουσιάστηκε αναλυτικά στην Ενότητα 3.2. Αυτό συμβαίνει γιατί όταν σε ένα κοινωνικό δίκτυο υπάρχει κοινοτική δομή είναι πιθανό οι χρήστες που ανήκουν σε κάποια κοινότητα να μοιράζονται κοινά ενδιαφέροντα και επομένως οι προτιμήσεις του ενός να επηρεάζονται άμεσα από αυτές του άλλου. Με βάση αυτή την υπόθεση, ο περιορισμός των υποψήφιων γειτονικών χρηστών ενός χρήστη-στόχου θα μπορούσε να βελτιώσει την ακρίβεια των συστάσεων. Εκτός από την ακρίβεια είναι δεδομένο ότι ο περιορισμός αυτός βελτιώνει και την ταχύτητα της φάσης υπολογισμού των γειτονικών χρηστών ενός χρήστη στόχου σε ένα σύστημα συνεργατικής διήθησης, αφού δε χρειάζεται να υπολογιστεί η ομοιότητα με όλους τους χρήστες αλλά μόνο με αυτούς που μοιράζονται την ίδια κοινότητα με αυτόν.

Στα συστήματα αυτά οι κοινότητες του δικτύου που θα προκύψουν από τον αλγόριθμο ανίχνευσης κοινοτήτων μπορούν να αξιοποιηθούν με διαφορετικούς τρόπους. Στην πιο απλή προσέγγιση όλοι οι χρήστες που ανήκουν στην κοινότητα ενός χρήστη-στόχου μπορούν να θεωρηθούν ως οι γειτονικοί χρήστες του και επομένως η βαθμολογία του χρήστη για ένα αντικείμενο να προκύψει από το μέσο όρο των βαθμολογιών όλων των χρηστών οι οποίοι ανήκουν στην κοινότητά του και έχουν βαθμολογήσει το συγκεκριμένο αντικείμενο. Οπότε σε αντιστοιχία με τις Εξισώσεις 3.12 και 3.13 οι συναρτήσεις πρόβλεψης για τη βαθμολογία ενός χρήστη u σε ένα αντικείμενο j θα έχουν τη μορφή:

$$\widehat{r_{uj}} = \frac{\sum\limits_{k \in N_u^*(j)} r_{kj}}{|N_u^*(j)|} \tag{3.19}$$

για τον απλό μέσο όρο και

$$\widehat{r_{uj}} = \mu_u + \frac{\sum\limits_{k \in N^*{}_u(j)} (r_{kj} - \mu_k)}{|N_u^*(j)|} \tag{3.20}$$

με τον τύπο του Resnick, όπου $N_u^*(j)$ είναι το σύνολο των χρηστών οι οποίοι ανήκουν στην ίδια κοινότητα με το χρήστη-στόχο u και έχουν βαθμολογήσει το αντικείμενο j.

Μια άλλη προσέγγιση θα ήταν οι χρήστες που ανήκουν στην ίδια κοινότητα με το χρήστη u να μην είναι οι τελικά γειτονικοί του χρήστες αλλά οι υποψήφιοι και οι τελικά όμοιοι χρήστες να υπολογίζονται με τις κλασσικές συναρτήσεις ομοιότητας όπως η ομοιότητα συνημιτόνου, ο συντελεστής ομοιότητας Pearson ή Spearman κλπ. Αυτή η προσέγγιση είναι σίγουρα πιο πλήρης από την παραπάνω καθώς λαμβάνει υπόψη και την κοινοτική δομή του κοινωνικού δικτύου αλλά και τις συσχετίσεις των βαθμολογιών μεταξύ των χρηστών. Ένα τέτοιο σύστημα είναι και αυτό που παρουσιάζουμε στο πειραματικό μέρος της εργασία στο Κεφάλαιο 4.

Κεφάλαιο 4

Πειραματική Διαδικασία και Αποτελέσματα

4.1 Η συλλογή δεδομένων

Για την υλοποίηση του πειραματικού μέρους της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκε η συλλογή δεδομένων `MovieTweetings` [Doom13], η οποία είναι μια συνεχώς ανανεούμενη συλλογή δεδομένων που λαμβάνεται καθημερινά από το Twitter και περιέχει βαθμολογίες ταινιών από χρήστες. Προκείμενου οι βαθμολογίες αυτές να εξαχθούν σωστά, λαμβάνονται υπόψη αποκλειστικά και μόνο καλά δομημένα tweets χρηστών του Twitter. Τα καλά δομημένα αυτά tweets προέρχονται από τις εφαρμογές του *IMDb (Internet Movie Database)*[1], της μεγαλύτερης κινηματογραφικής βάσης δεδομένων του διαδικτύου. Στις εφαρμογές αυτές, όταν ο χρήστης βαθμολογεί μια ταινία, του προτείνεται να μοιραστεί την εμπειρία του στο Twitter, με ένα καλά δομημένο tweet στη μορφή:

"I rated The Matrix 9/10 http://www.imdb.com/title/tt0133093/ #IMDb"

Έτσι, το API του Twitter διερωτάται σε καθημερινή βάση για τον όρο *"I rated #IMDb"* και τα tweets που εντοπίζονται δέχονται την κατάλληλη επεξεργασία και οι βαθμολογίες τους ενσωματώνονται στο dataset. Οι βαθμολογίες κυμαίνονται στην κλίμακα $0-10$ και είναι πάντα ακέραιοι αριθμοί.

Πιο συγκεκριμένα, χρησιμοποιήθηκε το *στιγμιότυπο* (snapshot) της $3^{\eta\varsigma}$ Φεβρουαρίου του 2016, τα χαρακτηριστικά του οποίου συνοψίζονται στον Πίνακα 4.1.

Πίνακας 4.1: Η συλλογή δεδομένων `MovieTweetings`

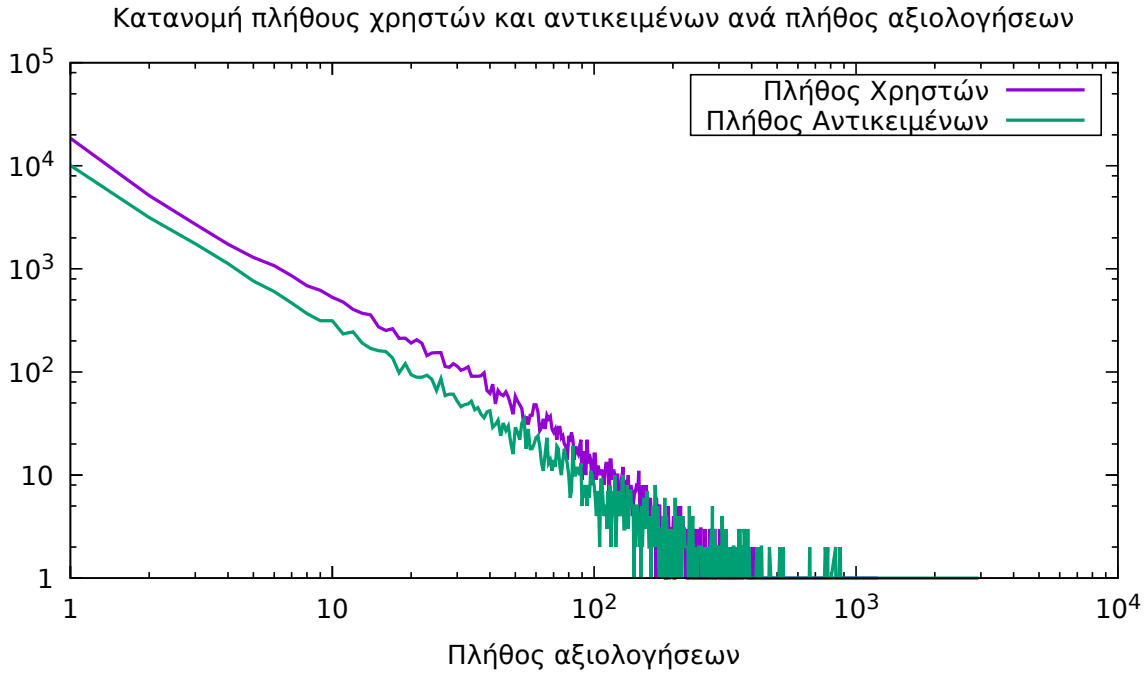
Χαρακτηριστικό	Τιμή
Βαθμολογίες	466.545
Δεσμοί μεταξύ των Χρηστών	205.122
Χρήστες	41.226
Ταινίες	23.536

Αυτό σημαίνει ότι εάν m είναι ο αριθμός των χρηστών και n ο αριθμός των αντικειμένων οι προσδιορισμένες τιμές του $m \times n$ πίνακα βαθμολογιών αποτελούν μονάχα ένα ποσοστό $\frac{466.545}{41.226 \times 23.536} \simeq 0,048\%$. Επομένως ο πίνακας βαθμολογιών είναι πολύ αραιός και μάλιστα η *πυκνότητά βαθμολογίας (rating density)* είναι πάνω από 10 τάξεις μικρότερη του 1% που είναι μια συνηθισμένη τιμή πυκνότητας για έναν πίνακα βαθμολογιών ενός συστήματος σύστασης.

Επίσης, ένα άλλο στατιστικό στο οποίο αξίζει να δοθεί έμφαση είναι οι κατανομές των βαθμολογιών ανά τους χρήστες και ανά τα αντικείμενα. Η (στρογγυλοποιημένη) μέση τιμή του πλήθους βαθμολογήσεων ανά χρήστη είναι 11 ενώ η αντίστοιχη μέση τιμή ανά αντικείμενο είναι 20. Αυτό σημαίνει ότι θα υπάρχει μεγάλος αριθμό χρηστών οι οποίοι έχουν ελάχιστα βαθμολογήσει και μεγάλος αριθμός ταινιών οι οποίες έχουν ελάχιστα βαθμολογηθεί. Η γραφική παράσταση του Σχήματος 1.3 της Ενότητας 1.3, η οποία χρησιμοποιήθηκε για την περιγραφή του φαινομένου long-tail, έχει προκύψει από τη συγκεκριμένη συλλογή δεδομένων. Επομένως είναι φανερό ότι η συλλογή δεδομένων

[1] http://www.imdb.com/

που χρησιμοποιήθηκε για την πειραματική διαδικασία αποτελεί χαρακτηριστικό παράδειγμα δεδομένων των οποίων η κατανομή ακολουθεί νόμο δύναμης. Αυτό επιβεβαιώνεται και από το Σχήμα 4.1 το όποιο δείχνει την ίδια κατανομή σε λογαριθμικούς άξονες. Όταν σε μια τέτοια γραφική παράσταση η κατανομή είναι σχεδόν ευθεία γραμμή τότε έχουμε να κάνουμε με δεδομένα που ακολουθούν νόμο δύναμης.



Σχήμα 4.1: Το φαινόμενο long-tail σε λογαριθμικούς άξονες

4.2 Πειραματική Διαδικασία

4.2.1 Το δίκτυο των χρηστών

Όπως αναφέρθηκε στην Ενότητα 1.4 το πειραματικό μέρος της παρούσας διπλωματικής εργασίας περιλαμβάνει την ανίχνευση των κοινοτήτων στις οποίες πιθανώς ανήκουν οι χρήστες του Twitter που περιλαμβάνονται στη συλλογή δεδομένων που περιγράφηκε στην Ενότητα 4.1. Εδώ πρέπει να σημειωθεί ότι η συγκεκριμένη συλλογή δεδομένων περιέχει για κάθε χρήστη που έχει βαθμολογήσει ταινίες το μοναδικό ID του στο Twitter.

Προκειμένου να προσομοιωθεί το δίκτυο των χρηστών χρησιμοποιήθηκε το *Twitter API* για να βρεθούν οι συνδέσεις μεταξύ τους. Έτσι δημιουργήθηκε ένας γράφος στη γραφο-θεωρητική βάση δεδομένων *Titan*[2], αποτελούμενος από $m = 41.226$ κόμβους και $n = 205.122$ ακμές. Οι κόμβοι του γράφου αναπαριστούν τους χρήστες και οι ακμές τις συνδέσεις μεταξύ τους. Στο σημείο αυτό πρέπει να τονιστεί ότι το Twitter είναι ένα κατευθυντικό δίκτυο, πράγμα που σημαίνει ότι αν ο χρήστης A ακολουθεί το χρήστη B, δεν είναι υποχρεωτικό και ο χρήστης B να ακολουθεί τον A.

Ένα σημαντικό χαρακτηριστικό του γράφου των χρηστών είναι η *πυκνότητα (graph density)*, η οποία δείχνει το πόσο απέχει ο συγκεκριμένος γράφος από το να γίνει κλίκα *κλίκα (clique)* (όπου όλοι οι χρήστες συνδέονται μεταξύ τους) και που για έναν κατευθυντικό γράφο ορίζεται ως εξής:

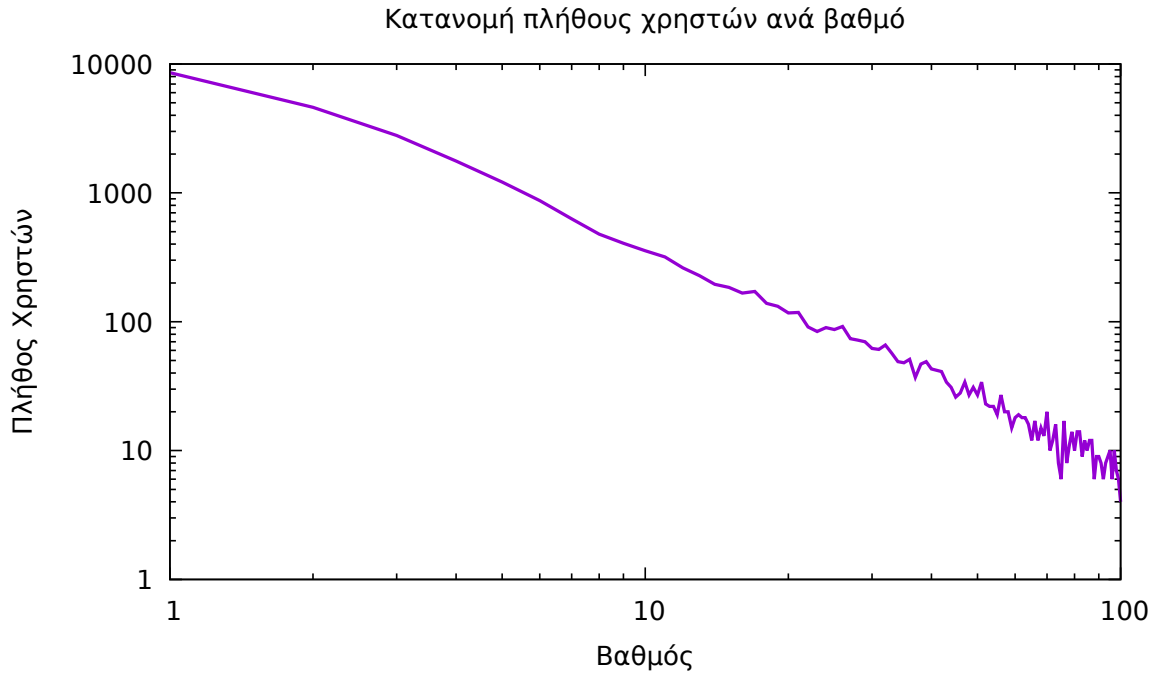
$$D = \frac{E}{V(V-1)} \tag{4.1}$$

[2] http://titan.thinkaurelius.com/

με το E να είναι το πλήθος των ακμών και το V το πλήθος των κόμβων. Στην περίπτωση του υπό εξέταση γράφου η πυκνότητα έχει τιμή $D = \frac{E}{V(V-1)} = \frac{205122}{41226*41225} = 0,000120693 \simeq 0,012\%$. Αυτό σημαίνει ότι έχουμε να κάνουμε με έναν πολύ αραιό γράφο.

Για το λόγο αυτό και σε μια προσπάθεια να αυξηθεί η πυκνότητα του, ο συγκεκριμένος γράφος αντιμετωπίστηκε ως μη-κατευθυντικός στην πειραματική διαδικασία. Άλλοι παράγοντες που συνηγόρησαν στην επιλογή αυτή είναι η επιτάχυνση του αλγορίθμου Girvan-Newman, όπως επίσης και το γεγονός πως σε αρκετές περιπτώσεις είναι καλύτερο να αγνοείται η κατευθυντική φύση ενός δικτύου όταν πρόκειται για τον εντοπισμό κοινοτικής δομής [Newm04c].

Ένα ακόμα χαρακτηριστικό του γράφου το οποίο υπήρξε ανασταλτικός παράγοντας στη διαδικασία ανίχνευσης κοινοτήτων είναι η *κατανομή βαθμών (degree distribution)* ανάμεσα στους κόμβους του γράφου. Όπως φαίνεται στο Σχήμα 4.2, η κατανομή αυτή ακολουθεί νόμο δύναμης αφού η μεγάλη πλειοψηφία των χρηστών συνδέονται με έναν ή ελάχιστους χρήστες, ενώ είναι λίγοι οι χρήστες οι οποίοι έχουν βαθμό πάνω από 100.



Σχήμα 4.2: Η κατανομή βαθμών κορυφών του γράφου των χρηστών

4.2.2 Η διαδικασία ανίχνευσης κοινοτήτων

Για το σκοπό της ανίχνευσης των κοινοτήτων στο δίκτυο των χρηστών χρησιμοποιήθηκε ο αλγόριθμος των Girvan-Newman ο οποίος παρουσιάστηκε αναλυτικά στην Ενότητα 3.2.5. Ο αλγόριθμος αυτός συνοπτικά εντοπίζει τις κοινότητες ενός δικτύου αφαιρώντας προοδευτικά τις ακμές οι οποίες συναντώνται περισσότερο σε διαδρομές συντομότερων μονοπατιών (shortest paths).

Εύρεση συνεκτικών συνιστωσών του γράφου

Από τη στιγμή που το δίκτυο των χρηστών έχει εντελώς τυχαία δομή, καθώς αποτελείται απλώς από χρήστες οι οποίοι έχουν βαθμολογήσει ταινίες αλλά χωρίς κάποια επιπλέον πληροφορία, είναι αναμενόμενο να υπάρχουν χρήστες ή ομάδες χρηστών οι οποίες να μη συνδέονται καθόλου με το υπόλοιπο δίκτυο χρηστών του dataset. Αυτό σημαίνει ότι, πρώτον, είναι πιθανό να υπάρχουν ήδη διαμορφωμένες κοινότητες στο δίκτυο των χρηστών και δεύτερον ότι είναι απαραίτητο να βρεθούν οι

συνεκτικές συνιστώσες (connected components) (Ενότητα 3.1.1) του δικτύου στις οποίες θα εφαρμοστεί ο αλγόριθμος ανίχνευσης κοινοτήτων. Το δεύτερο είναι απαραίτητη προϋπόθεση καθώς μόνο σε ένα συνεκτικό (connected) γράφο έχει νόημα να μιλάμε για ανίχνευση κοινοτήτων.

Προκειμένου να βρεθούν οι συνεκτικές συνιστώσες του γράφου εφαρμόστηκε *αναζήτηση κατά πλάτος (breadth-first search)* η οποία περιγράφηκε στην Ενότητα 3.1.1. Η δομή του γράφου, όπως προέκυψε από την αναζήτηση κατά πλάτος, συνοψίζεται στον Πίνακα 4.2

Πίνακας 4.2: Συνεκτικές συνιστώσες κοινωνικού δικτύου του `MovieTweetings`

Πλήθος συνεκτικών συνιστωσών	**Αριθμός Χρηστών**
1	24.632
1	18
1	17
1	16
3	8
2	7
7	6
17	5
30	4
94	3
625	2

Τέλος υπάρχουν και 14.726 χρήστες που δε συνδέονται με κανέναν άλλο χρήστη του δικτύου. Οι συνεκτικές συνιστώσες αποτελούμενες από 2 έως 18 χρήστες θεωρήθηκαν ως υπάρχουσες κοινότητες οι οποίες χρησιμοποιήθηκαν ως έχουν στο κομμάτι της σύστασης. Αντίθετα, η μεγάλη συνεκτική συνιστώσα του γράφου αποτελούμενη από 24.632 κόμβους-χρήστες ήταν αυτή στην οποία εφαρμόστηκε ο αλγόριθμος των Girvan-Newman για την ανίχνευση κοινοτητων. Επομένως στη συνέχεια της Ενότητας 4.2.2 όταν αναφερόμαστε σε δίκτυο ή γράφο θα εννοούμε τον υπογράφο της μεγάλης συνεκτικής συνιστώσας του δικτύου.

Σε συνέχεια της ανάλυσης που έγινε αναφορικά με το Σχήμα 4.2, αξίζει να σημειωθεί ότι η μεγάλη συνεκτική συνιστώσα του γράφου περιλάμβανε ένα χρήστη ο οποίος είχε βαθμό 7.130, αποτελούσε δηλαδή αυτό που θα μπορούσε να ονομαστεί "υπερκόμβος" για το γράφο. Πρόκειται για τον επίσημο λογαριασμό στο Twitter της ίδιας της IMDb. Είναι φανερό πως, από τη στιγμή που συνδεόταν σχεδόν με το $1/3$ τον χρηστών, δε συνεισέφερε στη διαδικασία ανίχνευσης κοινοτήτων και γι' αυτό το λόγο κρίθηκε σκόπιμο να αφαιρεθεί πριν την εφαρμογή του αλγορίθμου.

Εφαρμογή του αλγορίθμου Girvan-Newman

Όπως αναφέρθηκε στην Ενότητα 3.2.5, ο αλγόριθμος Girvan-Newman αποτελείται από τα εξής βήματα:

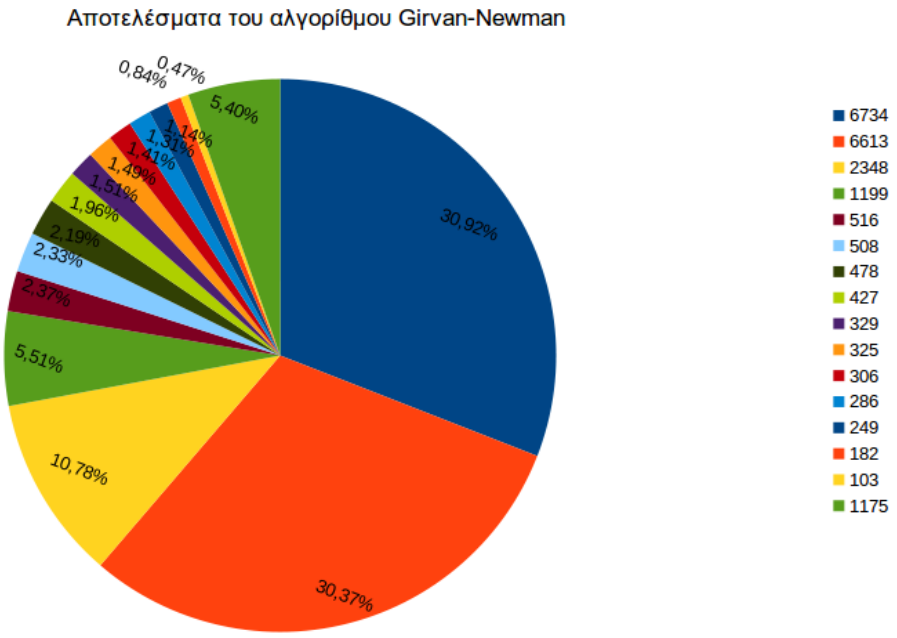
1. Υπολογισμός του βαθμού ενδιαμεσότητας όλων των ακμών.
2. Αφαίρεση της ακμής με το μεγαλύτερο βαθμό ενδιαμεσότητας.
3. Επανυπολογισμός του βαθμού ενδιαμεσότητας όλων των ακμών που επηρεάζονται.
4. Επανάληψη των βημάτων 2-3 μέχρι να φτάσουμε στον επιθυμητό αριθμό κοινοτήτων.

Για το πρώτο βήμα του αλγορίθμου Girvan-Newman, δηλαδή τον υπολογισμό του βαθμού ενδιαμεσότητας όλων των ακμών του δικτύου χρησιμοποιήθηκε ο γεωδαιτικός βαθμός ενδιαμεσότητας, ο οποίος παρουσιάστηκε στην Ενότητα 3.2.5. Για την εύρεση των συντομότερων διαδρομών μεταξύ

όλων των ζευγών κορυφών του δικτύου χρησιμοποιήθηκε ο αλγόριθμος του Dijkstra, ο οποίος περιγράφεται αναλυτικά στην Ενότητα 3.1.1. Έτσι, σε κάθε επανάληψη του αλγορίθμου υπολογίζονται οι συντομότερες διαδρομές μεταξύ όλων των ζευγών κορυφών και αφαιρείται η ακμή με τις περισσότερες προσπελάσεις.

Σαν επόμενο βήμα, μετά από την αφαίρεση κάθε ακμής, εξετάζεται αν ο γράφος συνεχίζει να είναι συνεκτικός, το οποίο σημαίνει ότι ο γράφος δεν έχει χωριστεί σε δύο μέρη, ή αν δεν είναι πια συνεκτικός, πράγμα το οποίο σημαίνει ότι η αφαίρεση της ακμής προκάλεσε το "σπάσιμο" του εξεταζόμενου δικτύου σε δύο μέρη. Στη δεύτερη περίπτωση, δηλαδή όταν μετά την αφαίρεση κάποιας ακμής ο γράφος χωρίζεται σε δύο υπογράφους, επιλέγεται ο υπογράφος με το μεγαλύτερο πλήθος κόμβων ως υπό εξέταση γράφος στον οποίο εφαρμόζεται εκ νέου ο αλγόριθμος ανίχνευσης κοινοτήτων. Για παράδειγμα, εάν στο αρχικό δίκτυο των 24.632 χρηστών μετά από κάποιες αφαιρέσεις ακμών αυτό χωριστεί σε δύο συνεκτικές συνιστώσες των 20.000 και 4.632 χρηστών αντίστοιχα, τότε ο αλγόριθμος θα προχωρήσει σε ανίχνευση κοινοτήτων στη συνιστώσα των 20.000 χρηστών. Σε κάθε περίπτωση δηλαδή ο αλγόριθμος εξετάζει ποιος είναι ο υπογράφος με το μεγαλύτερο πλήθος κόμβων και το θέτει ως τον υπό εξέταση υπογράφο.

Σχετικά με το κριτήριο τερματισμού του αλγορίθμου, εάν δεν υπάρχει καθορισμένος αριθμός επιθυμητών κοινοτήτων, όπως στην παρούσα εργασία, χρησιμοποιείται η συνάρτηση της τμηματικότητας η οποία παρουσιάστηκε στην Ενότητα 3.2.1. Ωστόσο, εξαιτίας του μεγάλου μεγέθους του γράφου, ο υπολογισμός της τιμής της τμηματικότητας σε κάθε επανάληψη του αλγορίθμου θα αύξανε σε μεγάλο βαθμό την πολυπλοκότητα του αλγορίθμου και γι' αυτό το λόγο δε χρησιμοποιήθηκε. Αντίθετα, κριτήριο τερματισμού του αλγορίθμου θεωρήθηκε ο αριθμός των κόμβων μιας κοινότητας αλλά και το πόσο συμπαγής ήταν αυτή. Δηλαδή, σε κοινότητες των $100 - 200$ κόμβων δεν επιδιώχθηκε επιπλέον "σπάσιμο" μιας και θεωρήθηκαν ικανοποιητικού μεγέθους, ώστε να αποτελούν μια αυτοτελή κοινότητα στο σύνολο των 24.632 κόμβων.



Σχήμα 4.3: Η κατανομή των κόμβων στις κοινότητες που δημιουργήθηκαν από τον αλγόριθμο Girvan-Newman

Επίσης, κάποιες κοινότητες οι οποίες προέκυψαν κατά την εφαρμογή του αλγορίθμου είχαν κατανομή βαθμών η οποία ακολουθούσε νόμο δύναμης. Δηλαδή λίγοι κόμβοι ήταν οι πιο δημοφιλείς συγκεντρώνοντας το μεγαλύτερο μέρος των ακμών ενώ σχεδόν όλοι οι υπόλοιποι κόμβοι είχαν ελάχιστες ακμές, οι περισσότερες από τις οποίες ήταν προς τους δημοφιλείς αυτούς κόμβους. Δίκτυα σαν τις κοινότητες αυτές ονομάζονται *δίκτυα ελεύθερης κλίμακας (scale-free networks)* και πολλά δίκτυα έχει

αποδειχθεί να είναι αυτού του τύπου [Clau09]. Αυτές τις κοινότητες ο αλγόριθμος Girvan-Newman δεν μπορούσε να τις "σπάσει" ικανοποιητικά σε μικρότερες, επομένως παρέμειναν ως είχαν παρά το μεγάλο μέγεθός τους.

Οι τελικές κοινότητες που προέκυψαν από τον αλγόριθμο Girvan-Newman καθώς και το πλήθος τους φαίνονται αναλυτικά στο Σχήμα 4.3. Όπως είναι φανερό στο εν λόγω σχήμα, υπάρχουν δύο κοινότητες οι οποίες μαζί αποτελούν περίπου το 60% του συνολικού αριθμού κόμβων (από τη συνεκτική συνιστώσα των 24.632 κόμβων). Αυτές οι δύο κοινότητες αποτελούν χαρακτηριστικό παράδειγμα δικτύων ελεύθερης κλίμακας τα οποία αναφέρθηκαν παραπάνω, δηλαδή η κατανομή των βαθμών των κόμβων τους ακολουθεί νόμο δύναμης και επομένως δεν εμφανίζουν κοινοτική δομή.

4.2.3 Η διαδικασία της αξιολόγησης

Για την παραγωγή συστάσεων υλοποιήθηκαν διάφοροι τύποι συστημάτων συστάσεων τα οποία συγκρίθηκαν με το κοινωνικό σύστημα το οποίο προτείνουμε στην παρούσα διπλωματική εργασία. Όπως σε κάθε σύστημα πρόβλεψης ή ταξινόμησης, προκειμένου να γίνει αξιολόγηση του μοντέλου, το σύνολο δεδομένων χωρίζεται σε δεδομένα εκπαίδευσης και δεδομένα επαλήθευσης. Στην περίπτωση των συστημάτων συνεργατικής διήθησης, ο διαχωρισμός σε δεδομένα εκπαίδευσης και επαλήθευσης είναι ουσιαστικά ο διαχωρισμός μεταξύ προσδιορισμένων και απροσδιόριστων εγγραφών του $m \times n$ πίνακα βαθμολογιών.

Διασταυρούμενη Αντεπικύρωση

Η πιο διαδεδομένη μέθοδος αξιολόγησης συστημάτων πρόβλεψης και γενικότερα συστημάτων μηχανικής μάθησης είναι η λεγόμενη *διασταυρούμενη αντεπικύρωση (cross-validation)*. Η μέθοδος αυτή αποτελεί μια τεχνική επαλήθευσης μοντέλων για να εκτιμήσει κανείς το κατά πόσο μια στατιστική ανάλυση μπορεί να γενικευθεί σε ένα ανεξάρτητο σύνολο δεδομένων [Koha95]. Στόχος της είναι να ορίσει ένα υποσύνολο δεδομένων για να ελέγξει το μοντέλο στη φάση της εκπαίδευσης, προκειμένου να περιορίσει φαινόμενα όπως η υπερπροσαρμογή (overfitting) και να δώσει μια ρεαλιστική εκτίμηση για την απόδοση του μοντέλου σε ένα καινούργιο σύνολο δεδομένων.

Ένας κύκλος διασταυρούμενης αντεπικύρωσης περιλαμβάνει τη διαμέριση του συνόλου δεδομένων σε συμπληρωματικά υποσύνολα, εφαρμόζοντας την ανάλυση κάθε φορά σε ένα από αυτά, το οποίο ονομάζεται σύνολο εκπαίδευσης (training set), και την επαλήθευση της ανάλυσης στο άλλο υποσύνολο, που ονομάζεται σύνολο επαλήθευσης (validation ή testing set). Για τη μείωση της μεταβλητότητας, πραγματοποιούνται αρκετοί κύκλοι διασταυρούμενης αντεπικύρωσης, χρησιμοποιώντας διαφορετικές διαμερίσεις και το τελικό αποτέλεσμα της αξιολόγησης προκύπτει από το μέσο όρο των αποτελεσμάτων όλων των κύκλων. Η μέθοδος της διασταυρούμενης αντεπικύρωσης μπορεί να χωριστεί σε δύο κατηγορίες, την *εξαντλητική (exhaustive)* και τη *μη-εξαντλητική (non-exhaustive)*, οι οποίες με τη σειρά τους περιλαμβάνουν διάφορες υποκατηγορίες.

Οι τύποι εξαντλητικής διασταυρούμενης αντεπικύρωσης εκπαιδεύουν και επαληθεύουν το μοντέλο με όλους τους δυνατούς τρόπους διαίρεσης του αρχικού συνόλου δεδομένων σε δεδομένα εκπαίδευσης και επαλήθευσης. Τέτοιους τύπους αποτελούν η *διασταυρούμενη αντεπικύρωση πλην p (leave-p-out cross-validation)*, η οποία περιλαμβάνει την εξαίρεση p παρατηρήσεων ως δεδομένων επαλήθευσης, ενώ όλες οι υπόλοιπες παρατηρήσεις ορίζουν τα δεδομένα εκπαίδευσης, με τη διαδικασία αυτή να επαναλαμβάνεται για όλα τα πιθανά υποσύνολα p παρατηρήσεων. Χαρακτηριστικό παράδειγμα αποτελεί η *διασταυρούμενη αντεπικύρωση πλην ενός (leave-one-out cross-validation)*, όπου η διαδικασία επαναλαμβάνεται n φορές, με το n να αποτελεί το πλήθος των παρατηρήσεων του συνόλου δεδομένων (κάθε φορά χρησιμοποιείται ως σύνολο επαλήθευσης μια μοναδική παρατήρηση). Ωστόσο, το μειονέκτημα της διασταυρούμενης αντεπικύρωσης πλην p είναι ότι για $p > 1$ και ειδικά για μεγαλύτερες τιμές του p γίνεται υπολογιστικά σχεδόν αδύνατη, αφού η μέθοδος χρειάζεται συνολικά C_p^n επαναλήψεις για ένα σύνολο δεδομένων με n παρατηρήσεις, όπου C ο διωνυμικός συντελεστής. Για παράδειγμα, για n=100 και και p=30 χρειάζονται $C_{30}^{100} \approx 3 \times 10^{25}$ κύκλοι διασταυρούμενης αντεπικύρωσης.

Αντίθετα, στη μη-εξαντλητική διασταυρούμενη αντεπικύρωση δεν υπολογίζονται όλοι οι δυνατοί τρόποι χωρισμού του συνόλου δεδομένων, αλλά οι μέθοδοι αυτές αποτελούν προσεγγίσεις της διασταυρούμενης αντεπικύρωσης πλην p. Η πιο γνωστή μέθοδος μη εξαντλητικής διασταυρούμενης αντεπικύρωσης και αυτή που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία είναι η *διασταυρούμενη αντεπικύρωση k διπλωμάτων (k-fold cross validation)*. Στη μέθοδο αυτή το αρχικό σύνολο δεδομένων χωρίζεται τυχαία σε k ίσα υποσύνολα και ένα από τα k αυτά υποσύνολα χρησιμοποιείται ως δεδομένα επαλήθευσης ενώ τα υπόλοιπα k-1 ως δεδομένα εκπαίδευσης. Στη συνέχεια η διαδικασία επαλήθευσης επαναλαμβάνεται k φορές (τα ονομαζόμενα "διπλώματα"), με κάθε ένα από τα k υποσύνολα να χρησιμοποιείται ακριβώς μια φορά ως σύνολο επαλήθευσης. Στο τέλος μπορεί να χρησιμοποιηθεί ο μέσος όρος από τα αποτελέσματα των k διπλωμάτων για να προκύψει μια μοναδική εκτίμηση. Προφανώς, στην περίπτωση που $k = n$ η μέθοδος είναι ακριβώς ίδια με τη διασταυρούμενη αντεπικύρωση πλην ενός, η οποία περιγράφηκε παραπάνω. Στο Σχήμα 4.4 απεικονίζεται η μέθοδος της διασταυρούμενης αντεπικύρωσης 4 διπλωμάτων.

Σχήμα 4.4: Αναπαράσταση k-fold cross validation με k=4. Από Fabian Flöck - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=51562781

Γενικά, δεν υπάρχει κάποιος βέλτιστος αριθμός k που να εγγυάται καλύτερα αποτελέσματα από άλλες επιλογές. Μια συνήθης επιλογή είναι το k να ισούται με 10 [McLa05]. Για την αξιολόγηση των συστημάτων που υλοποιήθηκαν στην πειραματική μας διαδικασία, η τιμή του k τέθηκε ίση με 5.

4.3 Μετρικές

Όπως αναφέρθηκε στην Ενότητα 1.1 το πρόβλημα της σύστασης διατυπώνεται με δύο τρόπους, την πρόβλεψη και την κατάταξη. Για αυτό το λόγο, οι κλασσικές μετρικές που χρησιμοποιούνται για την αξιολόγηση των συστημάτων συστάσεων μπορούν να χωριστούν σε *μετρικές ακρίβειας πρόβλεψης (prediction accuracy metrics)* και σε *μετρικές κατάταξης* ή *top-N μετρικές*, οι οποίες αφορούν προτεινόμενες λίστες αντικειμένων (recommended lists).

Οι πιο γνωστές μετρικές ακρίβειας της πρόβλεψης που έχουν χρησιμοποιηθεί στα συστήματα

συστάσεων είναι το το *μέσο απόλυτο σφάλμα (mean absolute error ή MAE)* και η *ρίζα του μέσου τετραγωνικού σφάλματος (root mean squared error ή MSE)*. Αν και οι συγκεκριμένες μετρικές συγκαταλέγονται ανάμεσα στις πρώτες που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης των συστημάτων συστάσεων, πλέον δεν θεωρούνται επαρκείς και έχουν επί της ουσίας εγκαταλειφθεί από την επιστημονική κοινότητα [Shan11].

4.3.1 Ακρίβεια και Ανάκληση

Από τις μετρικές κατάταξης, αυτές που είναι οι πλέον διαδεδομένες είναι η *Ακρίβεια (Precision)* και η *Ανάκληση (Recall)* και χρησιμοποιούνται για την αξιολόγηση της πραγματικής κατανάλωσης των αντικειμένων. Για παράδειγμα ένα σύστημα σύστασης μπορεί να προτείνει σε ένα χρήστη μια λίστα με κατάταξη από ταινίες, αλλά στην πραγματικότητα ο χρήστης πιθανότατα θα επιλέξει ένα υποσύνολο των ταινιών αυτών. Τα αντικείμενα τα οποία τελικά καταναλώνονται αναφέρονται ως *ground-truth positives* ή *true positives*. Στόχος των μετρικών αυτών είναι να υπολογίσουν το ποσοστό των σχετικών αντικειμένων από τη συνολική λίστα προτεινόμενων αντικειμένων του συστήματος σύστασης. Βεβαίως, είναι λογικό, μεταβάλλοντας το μέγεθος της προτεινόμενης λίστας να μεταβάλλεται αυτό το ποσοστό.

Η βασική ιδέα είναι ότι όλα τα αντικείμενα μπορούν να καταταχθούν με βάση τη βαθμολογία τους η οποία προκύπτει από τον εκάστοτε αλγόριθμο που χρησιμοποιεί το σύστημα σύστασης. Ωστόσο, μόνο τα top-k αντικείμενα (όπου k το μέγεθος της προτεινόμενης λίστας) συστήνονται στο χρήστη. Έτσι, μεταβάλλοντας το μέγεθος της λίστας μπορεί να εξετάσει κάποιος το ποσοστό των σχετικών αντικειμένων τα οποία βρίσκονται στη λίστα (*true-positives*) και το ποσοστό των σχετικών αντικειμένων τα οποία λείπουν από τη λίστα (*false-negatives*). Στην περίπτωση που η προτεινόμενη λίστα είναι πολύ μικρή τότε ο αλγόριθμος πιθανότατα θα χάσει κάποια σχετικά αντικείμενα, ενώ αν η λίστα έχει μεγάλο μέγεθος θα υπάρχουν συστάσεις αντικειμένων τα οποία δεν καταναλώνονται από το χρήστη (*false-positives*). Χρειάζεται επομένως ένας συμβιβασμός μεταξύ των false-positive και false-negative αντικειμένων. Το πρόβλημα είναι ότι το ιδανικό μέγεθος προτεινόμενης λίστας δε μπορεί να είναι γνωστό σε ένα πραγματικό σύστημα.

Έστω ότι επιλέγονται τα top-k αντικείμενα της προτεινόμενης λίστας για να προταθούν στο χρήστη. Για οποιαδήποτε τιμή μεγέθους της λίστας k, το σύνολο των προτεινόμενων αντικειμένων θα είναι $S(t)$ και θα ισχύει $|S(t)| = t$. Επίσης, έστω ένα σύνολο G, το οποίο περιλαμβάνει τα σχετικά αντικείμενα (ground-truth positives ή true positives) τα οποία έχουν όντως καταναλωθεί/βαθμολογηθεί από το χρήστη. Έτσι, για οποιαδήποτε τιμή του k, η Ακρίβεια ορίζεται ως το ποσοστό των σχετικών αντικειμένων από όλα τα προτεινόμενα αντικείμενα σύμφωνα με τον παρακάτω τύπο:

$$Precision(k) = 100 \cdot \frac{|S(k) \cap G|}{|S(k)|} \tag{4.2}$$

Η τιμή της παραπάνω Εξίσωσης εξαρτάται προφανώς από την επιλογή του k. Ωστόσο, πρέπει να σημειωθεί ότι η τιμή της ακρίβειας δεν είναι μονότονη συνάρτηση του k, καθώς ο αριθμητής και ο παρανομαστής μπορεί να αλλάξουν διαφορετικά μεταβάλλοντας το k. Η Εξίσωση 4.2 αποτελεί μια ειδική μορφή της ακρίβειας, εξαρτώμενη από την τιμή της προτεινόμενης λίστας k. Η γενική μορφή της ακρίβειας, η οποία απεικονίζεται παραστατικά στο Σχήμα 4.5 είναι η παρακάτω:

$$Precision = \frac{\textit{true positives}}{\textit{true positives} + \textit{false positives}} \tag{4.3}$$

Αντίστοιχα, ορίζεται η *Ανάκληση (Recall)* ως το ποσοστό των σχετικών αντικειμένων τα οποία τελικά προτάθηκαν από το σύστημα συστάσεων. Επομένως, σε αντιστοιχία με τον τύπο της ακρίβειας, η ανάκληση εκφράζεται από την παρακάτω Εξίσωση

$$Recall(k) = 100 \cdot \frac{|S(k) \cap G|}{|G|} \tag{4.4}$$

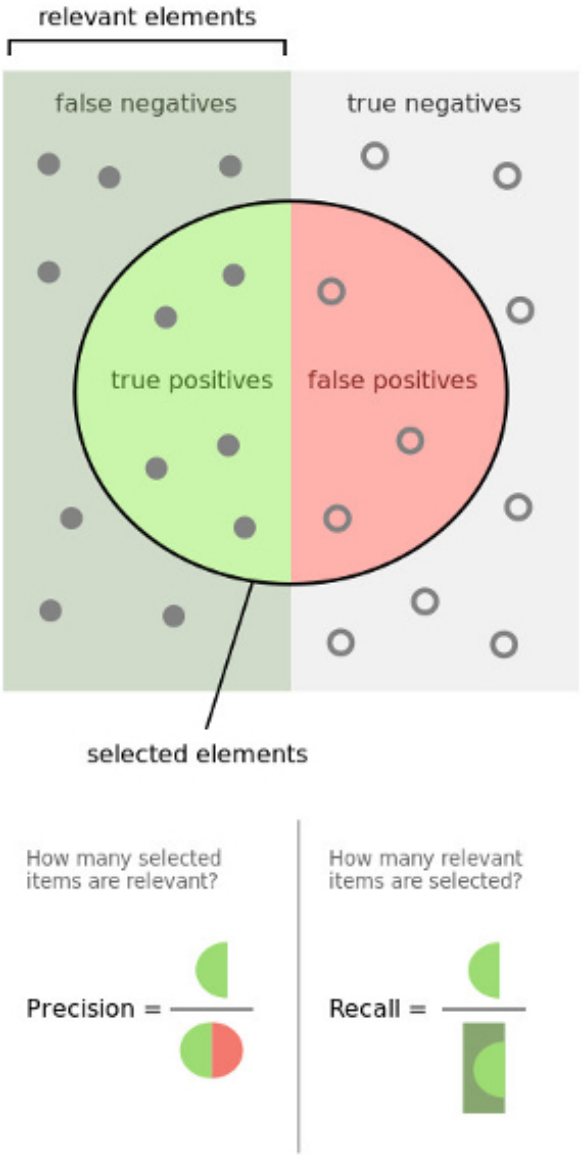
Η γενική μορφή της ανάκλησης, η οποία απεικονίζεται και στο Σχήμα 4.5 είναι η παρακάτω:

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{4.5}$$

Γενικά, υπάρχει ένα συμβιβασμός (trade-off) μεταξύ ακρίβειας και ανάκλησης ο οποίος όμως δεν είναι πάντα μονότονος. Δηλαδή, δεν είναι απαραίτητο ότι μια αύξηση της ακρίβειας θα οδηγεί πάντα σε μείωση της ανάκλησης και το αντίστροφο. Μια μετρική η οποία συνδυάζει την ακρίβεια και την ανάκληση είναι η F_1 score (γνωστή και ως F-score ή F-measure), η οποία είναι ο αρμονικός μέσος των δύο μετρικών με την υψηλότερη τιμή του να είναι το 1 και η χαμηλότερη το 0. Η μετρική F_1 score ορίζεται ως εξής:

$$F_1 \text{ score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.6}$$

Παρότι η μετρική F_1 score συνδυάζει την ακρίβεια και την ανάκληση, έχει δεχτεί κριτική σε συγκεκριμένες περιπτώσεις εξαιτίας της πόλωσης της [Powe11]. Η ακρίβεια, η ανάκληση και το F_1 score



Σχήμα 4.5: Σχηματική απεικόνιση των μετρικών της ακρίβειας και ανάκλησης. Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283

υπολογίζονται ξεχωριστά για κάθε χρήστη και στη συνέχεια συνήθως χρησιμοποιείται ο μέσος όρος τους στο σύνολο των χρηστών για να προκύψει μια καθολική τιμή για το σύστημα.

Το μειονέκτημα που παρουσιάζουν η ακρίβεια και η ανάκληση είναι ότι δίνουν μια μοναδική τιμή για ολόκληρη την προτεινόμενη λίστα και δεν δίνουν έμφαση στη σειρά με την οποία κατατάσσονται τα αντικείμενα. Για το λόγο αυτό μια πιο κατάλληλη μετρική που χρησιμοποιείται στα συστήματα συστάσεων, όπου είναι πολύ σημαντική η σειρά κατάταξης των αντικειμένων, είναι η *μέση ακρίβεια (average precision)*. Η μετρική αυτή υπολογίζει ουσιαστικά την τιμή της ακρίβειας και της ανάκλησης σε κάθε θέση της προτεινόμενης λίστας από 1 έως n, όπου n το μέγεθος αυτής. Έχοντας επομένως μια τιμή ακρίβειας p και μια τιμή ανάκλησης r για κάθε θέση στη λίστα των αντικειμένων, μπορεί κάποιος να σχεδιάσει μια καμπύλη ακρίβειας-ανάκλησης, με την ακρίβεια να είναι μια συνάρτηση της ανάκλησης $p(r)$. Η μέση ακρίβεια υπολογίζει τη μέση τιμή της $p(r)$ πάνω στο διάστημα από $r = 0$ έως $r = 1$ δηλαδή [Zhan12]:

$$\text{AvgPrecision} = \int_0^1 p(r)d(r) \tag{4.7}$$

Το παραπάνω ολοκλήρωμα είναι ουσιαστικά η περιοχή κάτω από την καμπύλη ακρίβειας-ανάκλησης. Το ολοκλήρωμα αυτό μπορεί να αντικατασταθεί από ένα πεπερασμένο άθροισμα πάνω σε κάθε θέση της κατάταξης των αντικειμένων στην προτεινόμενη λίστα ως εξής:

$$\text{AvgPrecision} = \sum_{k=1}^{n} P(k)\Delta r(k) \tag{4.8}$$

όπου k είναι η κατάταξη στην προτεινόμενη λίστα n αντικειμένων, $P(k)$ είναι η ακρίβεια των top-k αντικειμένων και $\Delta r(k)$ είναι η αλλαγή στην ανάκληση από $k-1$ σε k αντικείμενα.

Επειδή η μέση ακρίβεια υπολογίζεται σε επίπεδο χρήστη, για να προκύψει μια καθολική τιμή τα συστήματα συστάσεων χρησιμοποιούν την ονομαζόμενη *Μέση Αντιπροσωπευτική Ακρίβεια (Mean Average Precision ή MAP)*, στην οποία υπολογίζεται ο μέσος όρος της μέσης ακρίβειας στο σύνολο των χρηστών m. Δηλαδή:

$$\text{MAP} = \frac{\sum_{u=1}^{m} \text{AvgPrecision(u)}}{m} \tag{4.9}$$

Η παραπάνω μορφή της ακρίβειας λαμβάνει υπόψη όλα τα αντικείμενα, ανεξάρτητα αν έχουν λάβει υψηλή ή χαμηλή βαθμολογία. Μια άλλη επέκταση της μετρικής λαμβάνει υπόψη μόνο τα "καλά" αντικείμενα, δηλαδή αντικείμενα τα οποία έχουν λάβει υψηλή βαθμολογία, και για αυτό αναφέρεται με το όνομα *MAP of good*. Το ποια βαθμολογία θεωρείται καλή και κακή είναι μια παράμετρος που πρέπει να οριστεί από το χρήστη και προφανώς μπορεί να επηρεάσει σημαντικά την τιμή της μετρικής.

4.3.2 Κάλυψη

Για να είναι επιτυχημένο ένα σύστημα σύστασης δεν αρκεί μόνο να είναι ακριβές στις προβλέψεις του, αλλά θα πρέπει να έχει τη δυνατότητα να εξασφαλίζει ότι είναι πάντα σε θέση να κάνει συστάσεις για τη μεγάλη πλειοψηφία των αντικειμένων και των χρηστών. Η μετρική που σχετίζεται με αυτή την ικανότητα είναι γνωστή ως *Κάλυψη (Coverage)*. Το γεγονός ότι οι αλγόριθμοι των συστημάτων συστάσεων δεν μπορούν να κάνουν συστάσεις για όλα τα αντικείμενα είναι αποτέλεσμα της αραιότητας των πινάκων βαθμολογιών. Παρόλα αυτά, αρκετά συστήματα εμφανίζουν συνήθως 100% κάλυψη επειδή χρησιμοποιούν τιμές αναφοράς (π.χ. τον μέσο όρο των βαθμολογιών του αντικειμένου) ή προεπιλεγμένες τιμές για βαθμολογίες που δεν μπορεί να προβλέψει ο αλγόριθμος. Για το λόγο αυτό είναι πολύ σημαντικός ο συμβιβασμός μεταξύ της ακρίβειας και της κάλυψης ενός συστήματος.

Η κάλυψη γενικά μπορεί να μετρηθεί στο σύνολο των προβλέψεων για τους διάφορους συνδυασμούς χρήστη-αντικειμένου ή να χωριστεί σε *κάλυψη χώρου χρήστη (user-space coverage)* και σε *κάλυψη χώρου αντικειμένου (item-space coverage)*. Η κάλυψη χώρου χρήστη μετράει το ποσοστό των χρηστών για τους οποίους τουλάχιστον k βαθμολογίες μπορούν να προβλεφθούν. Η τιμή του k ουσιαστικά είναι το μέγεθος της λίστας συστάσεων. Αντίστοιχα, η κάλυψη χώρου αντικειμένου μετράει το ποσοστό των αντικειμένων για τα οποία οι προβλέψεις τουλάχιστον k χρηστών είναι εφικτές.

Ωστόσο, η τελευταία μορφή της κάλυψης σπάνια χρησιμοποιείται αφού τα συστήματα συστάσεων παρέχουν προτεινόμενες λίστες για χρήστες και όχι για αντικείμενα.

Μια πιο συνηθισμένη μορφή της κάλυψης χώρου αντικειμένου είναι γνωστή ως *κάλυψη καταλόγου (catalog coverage ή CC)*, η οποία εφαρμόζεται μόνο σε λίστες συστάσεων [Agga16]. Η μετρική αυτή είναι χρήσιμη για περιπτώσεις όπου παρότι είναι εφικτή η πρόβλεψη των περισσότερων εγγραφών του πίνακα βαθμολογιών, η προτεινόμενη λίστα περιλαμβάνει πάντα το ίδιο σύνολο αντικειμένων. Αυτό έχει ως αποτέλεσμα μολονότι η κάλυψη χώρου αντικειμένου έχει υψηλή τιμή, από τη μια οι συστάσεις στους χρήστες να μην έχουν ποικιλία και από την άλλη ο κατάλογος των αντικειμένων να μην καλύπτεται πλήρως. Έτσι, εάν T_u είναι η λίστα των top-k αντικειμένων που προτείνονται στο χρήστη $u \in 1...m$, η κάλυψη καταλόγου ορίζεται όπως παρακάτω:

$$CC = \frac{|\cup_{u=1}^{m} T_u|}{n} \tag{4.10}$$

Η κάλυψη καταλόγου δηλαδή εκφράζει το ποσοστό των αντικειμένων τα οποία έχουν προταθεί σε τουλάχιστον ένα χρήστη.

4.4 Αποτελέσματα

Ο Πίνακας 4.3 συνοψίζει τα συστήματα που υλοποιήθηκαν στο πειραματικό κομμάτι της εργασίας. Για την εκπαίδευση των συστημάτων επιλέχθηκε η διασταυρούμενη αντεπικύρωση 5 διπλωμάτων (Ενότητα 4.2) και τέλος για την αξιολόγηση των παραγόμενων συστάσεων χρησιμοποιήθηκε η Μέση Αντιπροσωπευτική Ακρίβεια (Εξίσωση 4.9) για τα "καλά" αντικείμενα λίστας ελέγχου N (όσα δηλαδή έχουν βαθμολογία πάνω από ένα όριο), καθώς και η Κάλυψη (Ενότητα 4.3.2).

Πίνακας 4.3: Υλοποιούμενα συστήματα συστάσεων στην πειραματική διαδικασία

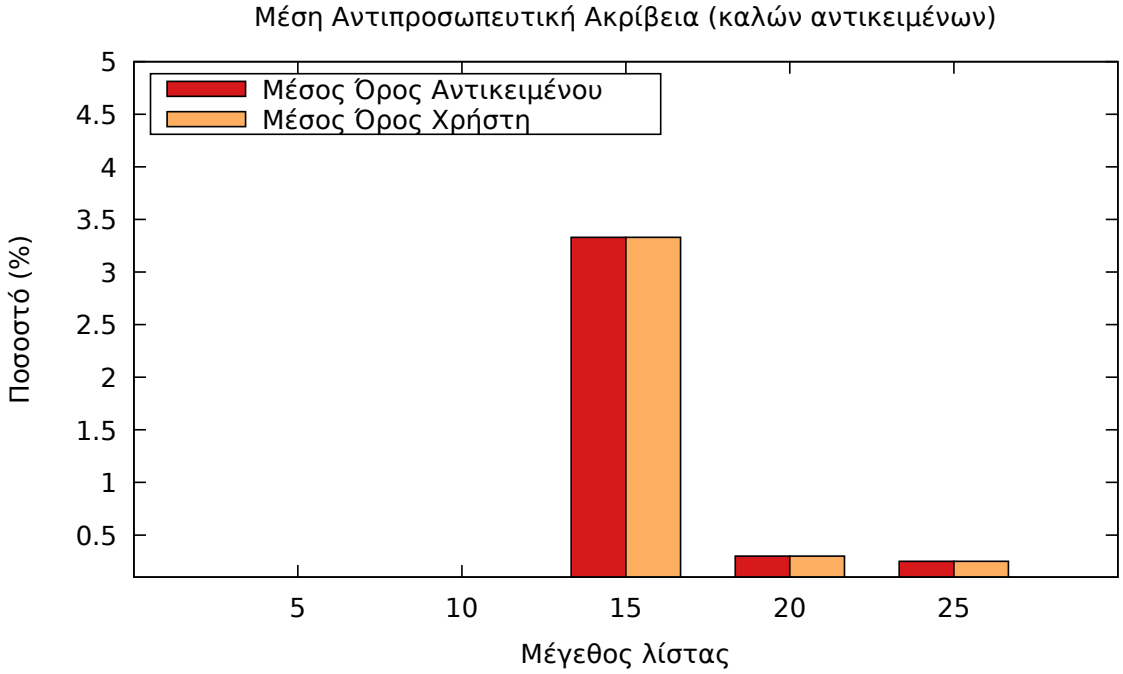
Κατηγορία Συστήματος	Σύστημα
Συστήματα Αναφοράς	Μέσος Όρος Χρήστη Μέσος Όρος Αντικειμένου
Συστήματα Συνεργατικής Διήθησης	Βασισμένης στο Χρήστη (User-based CF) Βασισμένης στο Αντικείμενο (Item-based CF)
Συστήματα Κοινωνικής Συνεργατικής Διήθησης	Ανάλυσης Κοινωνικού Δικτύου Ανίχνευσης Κοινοτήτων

Όπως θα γίνει περισσότερο εμφανές στην ανάλυση που θα ακολουθήσει, τα συστήματα του Πίνακα 4.3 εξετάστηκαν για διαφορετικές συναρτήσεις ομοιότητας και πρόβλεψης και σε συνάρτηση με το μέγεθος της γειτονιάς, το κατώφλι ομοιότητας και το μέγεθος της προτεινόμενης λίστας στην περίπτωση της Μέσης Αντιπροσωπευτικής Ακρίβειας.

4.4.1 Συστήματα αναφοράς

Στην περίπτωση των συστημάτων αναφοράς που αποτελούνται από το μέσο όρο χρήστη και τον μέσο όρο αντικειμένου δεν υφίσταται το μέγεθος γειτονιάς και το κατώφλι ομοιότητας σαν παράμετρος μιας και δεν υπολογίζεται κάποια ομοιότητα μεταξύ χρηστών ή αντικειμένων. Επίσης, όπως είναι λογικό, η κάλυψη θα είναι ίση με 100% για τα δύο αυτά συστήματα αφού πάντα είναι υπολογίσιμος ο μέσος όρος της βαθμολογίας του ίδιου του χρήστη και του ίδιου αντικειμένου, από τη στιγμή που στη συγκεκριμένη συλλογή δεδομένων δεν υπάρχουν χρήστες οι οποίοι δεν έχουν βαθμολογήσει ούτε ένα αντικείμενο ή αντικείμενα τα οποία δεν έχουν βαθμολογηθεί.

Στο Σχήμα 4.6 απεικονίζεται η μέση αντιπροσωπευτική ακρίβεια (για τα "καλά" αντικείμενα) σε συνάρτηση με το μέγεθος της προτεινόμενης λίστας. Είναι λογικό για τα συστήματα αυτά η MAP να έχει πολύ χαμηλές τιμές αφού σε όλους τους χρήστες προτείνονται τα ίδια αντικείμενα, και συγκεκριμένα τα k αντικείμενα τα οποία έχουν λάβει την υψηλότερη βαθμολογία (όπου k το μέγεθος της



Σχήμα 4.6: Μέση Αντιπροσωπευτική Ακρίβεια για τα Συστήματα Αναφοράς

προτεινόμενης λίστας). Τα αντικείμενα αυτά ωστόσο δεν σημαίνει ότι έχουν όντως βαθμολογηθεί από τον κάθε χρήστη (δεν αποτελούν δηλαδή true positives) με αποτέλεσμα η ακρίβεια να είναι σχεδόν για όλους τους χρήστες μηδενική.

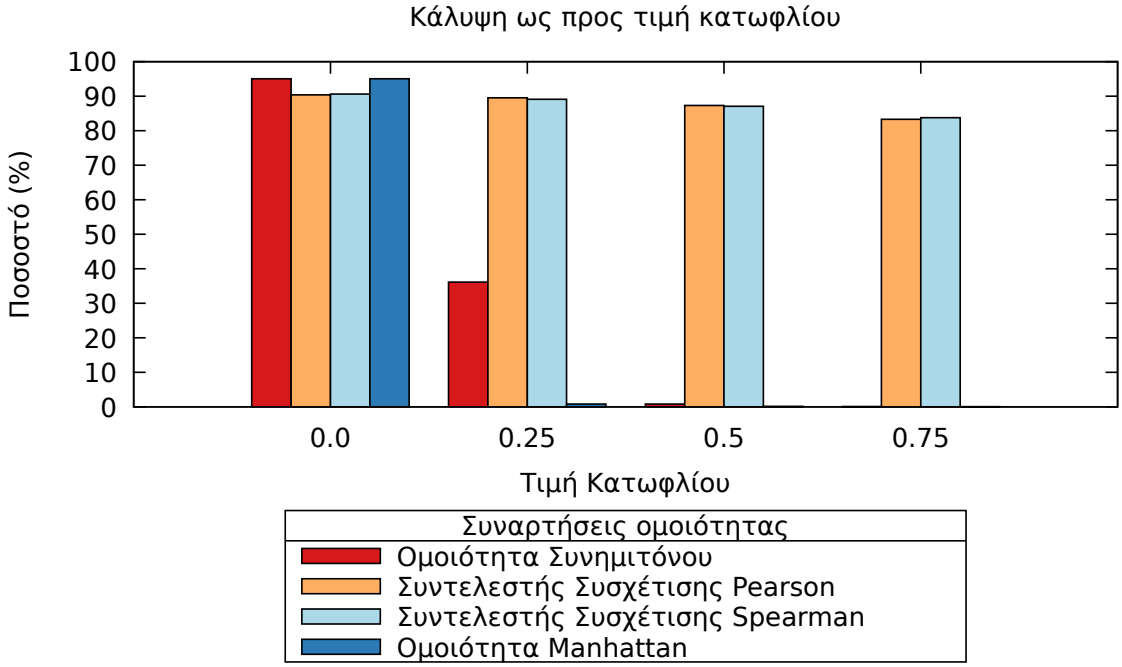
4.4.2 Συνεργατική διήθηση βασισμένη στο χρήστη (User-based CF)

Όσον αφορά τη συνεργατική διήθηση βασισμένη στο χρήστη, πραγματοποιήθηκαν πολλά πειράματα για διαφορετικές τιμές των παραμέτρων που περιλαμβάνει ένα τέτοιο σύστημα (Πίνακας 4.4)

Πίνακας 4.4: Παράμετροι συστημάτων συνεργατικής διήθησης βασιμένης στο χρήστη

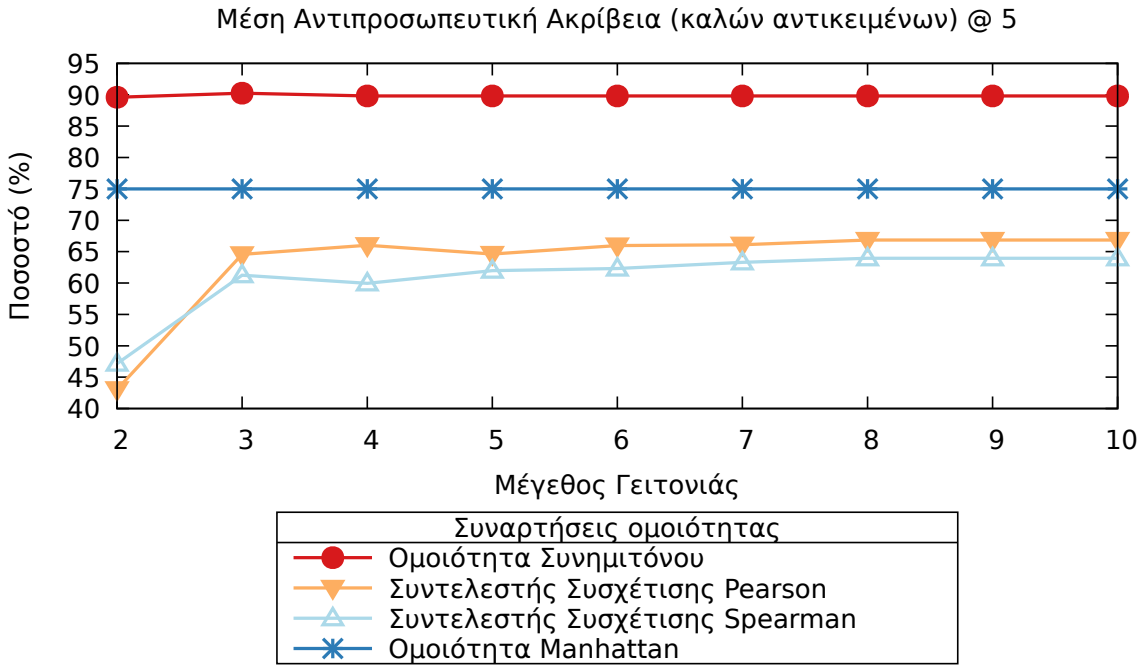
Παράμετρος	**Τιμές**
Μέγεθος γειτονιάς	1 ως 20
Συνάρτηση ομοιότητας	Συνημιτόνου, Pearson, Spearman, Manhattan
Κατώφλι ομοιότητας	$0.0, 0.25, 0.5, 0.75$
Μέγεθος λίστας	2 ως 6

Στο Σχήμα 4.7 απεικονίζεται η κάλυψη των βαθμολογιών ως προς την τιμή κατωφλίου για διαφορετικές συναρτήσεις ομοιότητας. Μια πρώτη παρατήρηση είναι πως όσο αυξάνεται η τιμή του κατωφλίου, η κάλυψη μειώνεται, πράγμα φυσιολογικό μιας και ο αριθμός των όμοιων χρηστών επίσης μειώνεται. Ωστόσο, η μείωση αυτή δεν επηρεάζει στον ίδιο βαθμό όλες τις συναρτήσεις ομοιότητας. Οι συντελεστές συσχέτισης Pearson και Spearman φαίνεται πως διατηρούν υψηλό βαθμό κάλυψης ακόμα και όταν λαμβάνεται υπόψη περιορισμένος αριθμός χρηστών, σε σχέση με τις ομοιότητες συνημιτόνου και Manhattan, η κάλυψη των οποίων πέφτει κατά πολύ ακόμα και με μικρή αύξηση της τιμής του κατωφλίου ομοιότητας. Αυτή η συμπεριφορά οφείλεται στο γεγονός πως οι συναρτήσεις ομοιότητας Pearson και Spearman χρησιμοποιούν τις κεντραρισμένες στο μέσο όρο τιμές των βαθμολογιών ενώ η ομοιότητα συνημιτόνου και Manhattan τις ακατέργαστες, όπως φαίνεται στις Εξισώσεις της Ενότητας 2.2.1. Με αυτό τον τρόπο οι πρώτες εξαλείφουν την πόλωση που εισάγει η κλίμακα στην



Σχήμα 4.7: Κάλυψη ως προς τις τιμές κατωφλίου για διάφορες συναρτήσεις ομοιότητας στα User-based CF συστήματα

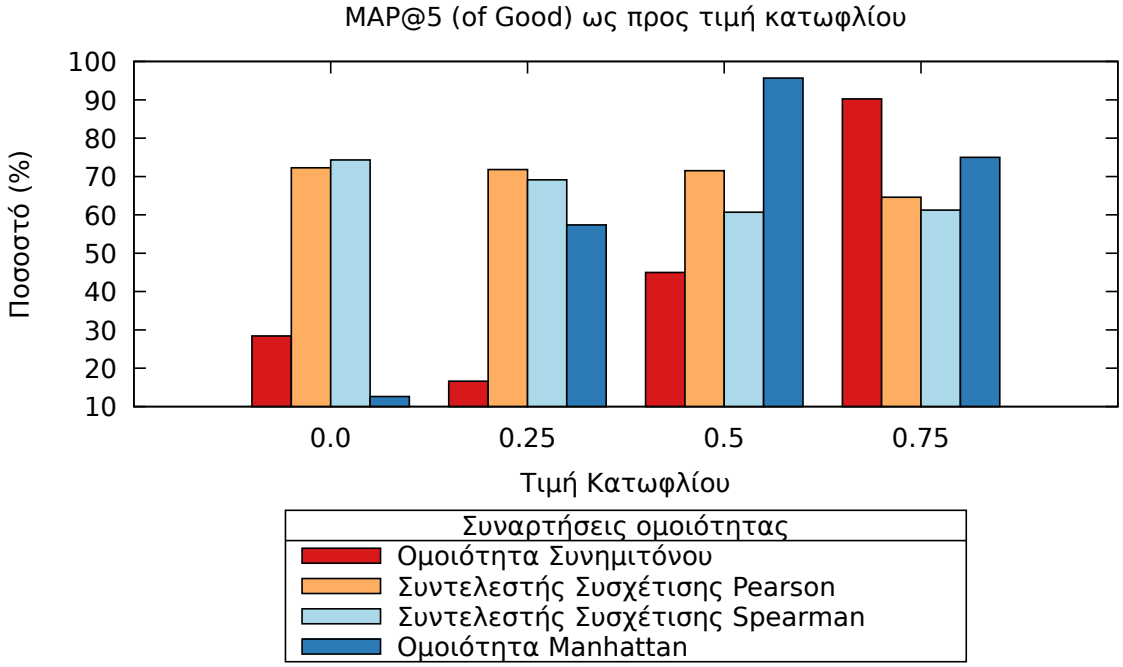
οποία βαθμολογεί ο κάθε χρήστης (δηλαδή ότι κάποιοι χρήστες βαθμολογούν γενικά με υψηλές βαθμολογίες ενώ κάποιοι άλλοι με χαμηλές) και έτσι είναι ευκολότερο να βρεθούν δύο όμοιοι χρήστες, με αποτέλεσμα να παρουσιάζουν μεγαλύτερη κάλυψη ακόμα και σε υψηλές τιμές του κατωφλίου ομοιότητας.



Σχήμα 4.8: MAP of Good @5 συναρτήσει του μεγέθους γειτονιάς για τις διαφορετικές συναρτήσεις ομοιότητας και κατώφλι ομοιότητας ίσο με 0.75

Στο Σχήμα 4.8 φαίνεται η μέση ακρίβεια (για τα "καλά" αντικείμενα) για μέγεθος λίστας ίσο με 5 (MAP of Good@5) σε συνάρτηση με το μέγεθος της γειτονιάς. Όπως γίνεται εύκολα αντιληπτό, για μεγέθη γειτονιάς μεγαλύτερα των τριών ή τεσσάρων χρηστών οι μεταβολές στην ακρίβεια είναι σχεδόν μηδενικές, ανεξαρτήτως της συνάρτησης ομοιότητας που επιλέγεται. Το φαινόμενο αυτό δείχνει ότι συνήθως δεν μπορούν να βρεθούν παραπάνω από 3 με 4 όμοιοι χρήστες για ένα χρήστη-στόχο, γεγονός το οποίο οφείλεται πρώτον στην αραιότητα του πίνακα βαθμολογιών και δεύτερον στο φαινόμενο long-tail που παρουσιάζει η συγκεκριμένη συλλογή δεδομένων, για τα οποία έγινε αναφορά στην Ενότητα 4.1.

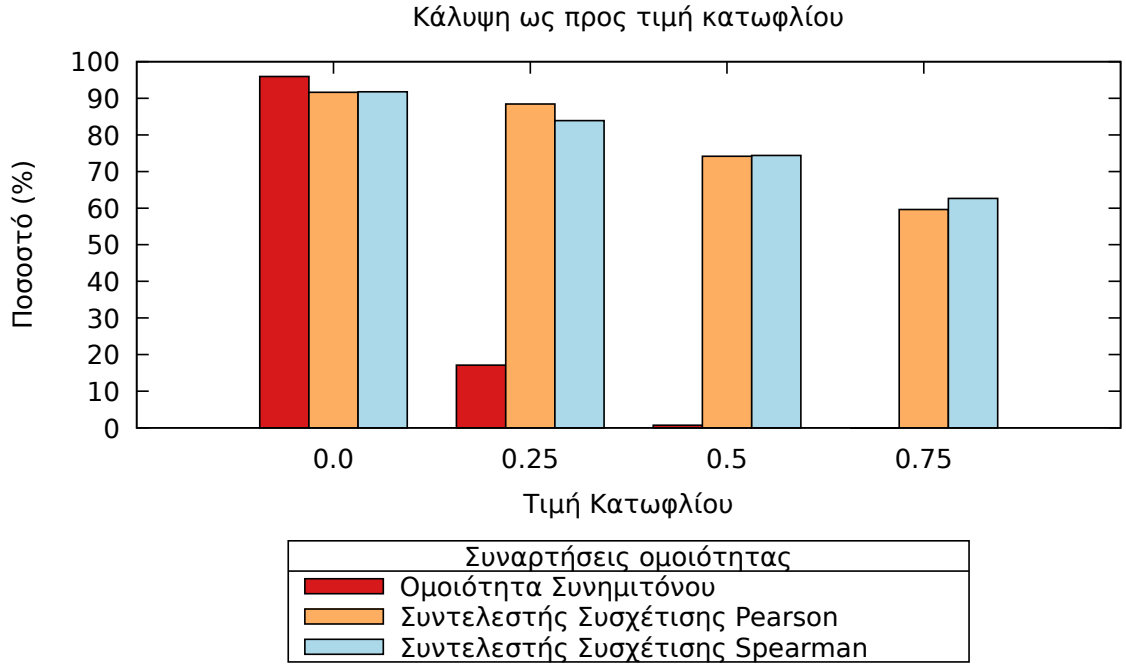
Γενικότερα, το μέγεθος γειτονιάς το οποίο εξετάστηκε ως παράμετρος σε όλα τα συστήματα που υλοποιήθηκαν, έδειξε να μην επηρεάζει τα αποτελέσματα των μετρικών για μεγέθη γειτονιάς μεγαλύτερα του 3 και για αυτό το λόγο κρίθηκε σκόπιμο να μην παρουσιαστούν περαιτέρω αναλύσεις σχετικά με την παράμετρο αυτή. Αντιθέτως, στις αναλύσεις που θα ακολουθήσουν το μέγεθος γειτονιάς είναι σταθερό και ίσο με 3.



Σχήμα 4.9: MAP of Good @5 συναρτήσει του κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας και για μέγεθος γειτονιάς ίσο με 3

Στο Σχήμα 4.9 απεικονίζεται πάλι η μέση ακρίβεια για τα "καλά αντικείμενα", αλλά αυτή τη φορά ως προς το κατώφλι ομοιότητας και για μέγεθος γειτονιάς ίσο με 3. Αυτό που παρατηρούμε στο συγκεκριμένο γράφημα είναι ότι αυξάνοντας την τιμή του κατωφλίου ομοιότητας η ακρίβεια αυξάνεται στην περίπτωση των συναρτήσεων ομοιότητας συνημιτόνου και Manhattan ενώ στην περίπτωση της ομοιότητας Pearson και Spearman παραμένει σταθερή ή μειώνεται για πολύ μεγάλη τιμή του κατωφλίου ομοιότητας.

Ωστόσο, η ανάλυση αυτή δεν είναι ανεξάρτητη αυτής του Σχήματος 4.7 αφού παρότι οι συναρτήσεις ομοιότητας συνημιτόνου και Manhattan παρουσιάζουν μεγαλύτερη ακρίβεια για υψηλές τιμές του κατωφλίου ομοιότητας, η κάλυψη τους στις περιπτώσεις αυτές είναι σχεδόν μηδενική, πράγμα που σημαίνει ότι η ακρίβεια αυτή έχει προκύψει από πολύ μικρό αριθμό επιτυχών προβλέψεων του αλγορίθμου. Υπάρχει, επομένως, ένας συμβιβασμός μεταξύ κάλυψης και ακρίβειας για τις δύο αυτές συναρτήσεις ομοιότητας στα βασισμένα στο χρήστη συστήματα.



Σχήμα 4.10: Κάλυψη συναρτήσει του κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας

4.4.3 Συνεργατική διήθηση βασισμένη στο αντικείμενο (Item-based CF)

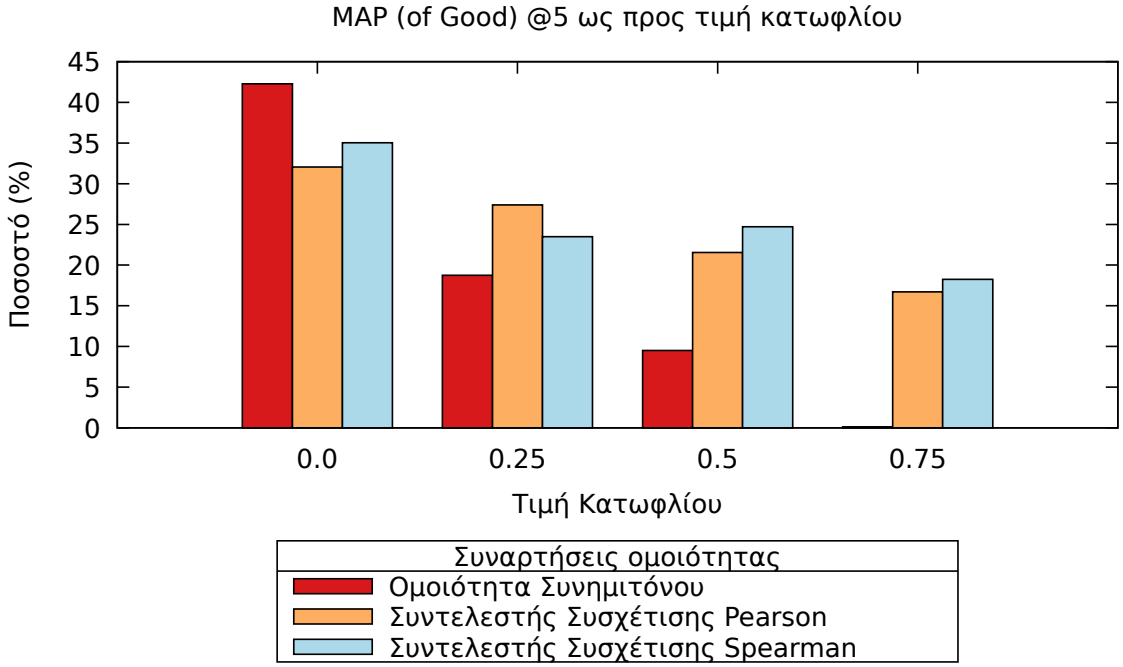
Στα συστήματα συνεργατικής διήθησης βασισμένης στο αντικείμενο (Item-based CF) πραγματοποιήθηκαν ακριβώς τα ίδια πειράματα με αυτά της βασισμένης στο χρήστη για τις ίδιες τιμές των παραμέτρων. Ωστόσο, στα item-based συστήματα δεν συμπεριλαμβάνεται στα γραφήματα η ομοιότητα Manhattan. Ο λόγος είναι ότι ελάχιστα αντικείμενα έχουν βαθμολογηθεί από κοινού από πολλούς χρήστες, οπότε σε πολλές περιπτώσεις αυτή δεν ορίζεται.

Στο Σχήμα 4.10 απεικονίζεται η κάλυψη των βαθμολογιών ως προς την τιμή κατωφλίου για τις διαφορετικές συναρτήσεις ομοιότητας. Η συμπεριφορά εδώ είναι παρόμοια με της συνεργατική διήθηση βασισμένη στο χρήστη, αφού αυξάνοντας την τιμή του κατωφλίου ομοιότητας μειώνεται η κάλυψη με τις συναρτήσεις ομοιότητας Pearson και Spearman να διατηρούν πάλι μεγαλύτερη κάλυψη από ότι η συνάρτηση συνημιτόνου, για τους λόγους που εξηγήθηκαν παραπάνω. Ωστόσο, ακόμα και οι συναρτήσεις ομοιότητας Pearson και Spearman δε διατηρούν τόσο μεγάλη κάλυψη για υψηλές τιμές του κατωφλίου ομοιότητας, όσο στην περίπτωση των user-based συστημάτων. Το γεγονός αυτό οφείλεται στο ότι είναι μικρότερος ο αριθμός των αντικειμένων τα οποία έχουν N αξιολογήσεις από ότι ο αντίστοιχος αριθμός χρηστών, όπως φαίνεται και στο Σχήμα 1.3.

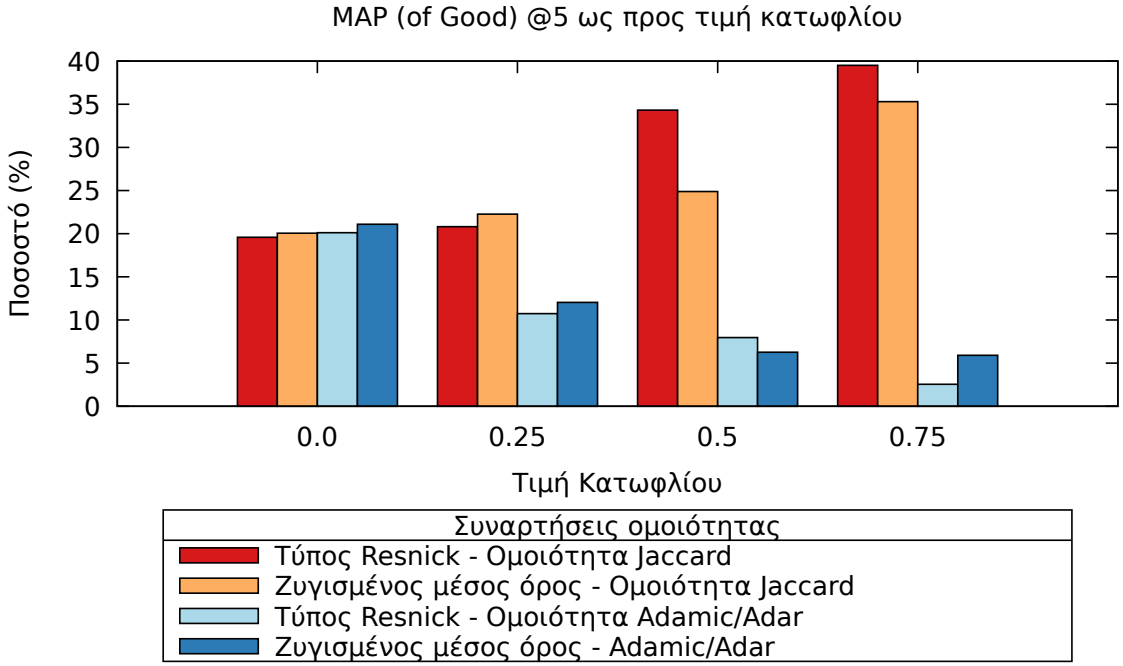
Μια άλλη γραφική παράσταση είναι αυτή της μέσης ακρίβειας (για τα "καλά" αντικείμενα) ως προς το κατώφλι ομοιότητας, η οποία φαίνεται στο Σχήμα 4.11. Στο σχήμα αυτό μπορεί κανείς να παρατηρήσει ότι η μέση ακρίβεια μειώνεται όσο αυξάνεται η τιμή το κατωφλίου ομοιότητας, σε αντίθεση με το αντίστοιχο σχήμα των user-based συστημάτων (Σχήμα 4.9). Το φαινόμενο αυτό οφείλεται στη χαμηλή κάλυψη που παρουσιάζουν τα συστήματα αυτά, όπως φάνηκε στο Σχήμα 4.10, για υψηλές τιμές του κατωφλίου ομοιότητας, το οποίο έχει ως αποτέλεσμα τη μείωση της μέσης ακρίβειας.

4.4.4 Συστήματα κοινωνικής σύστασης που βασίζονται στην ανάλυση κοινωνικών δικτύων

Εκτός από τα κλασσικά συστήματα συνεργατικής διήθησης υλοποιήθηκαν και κοινωνικά συστήματα συνεργατικής διήθησης τα οποία χρησιμοποιούν σαν συνάρτηση ομοιότητας τους συντελεστές



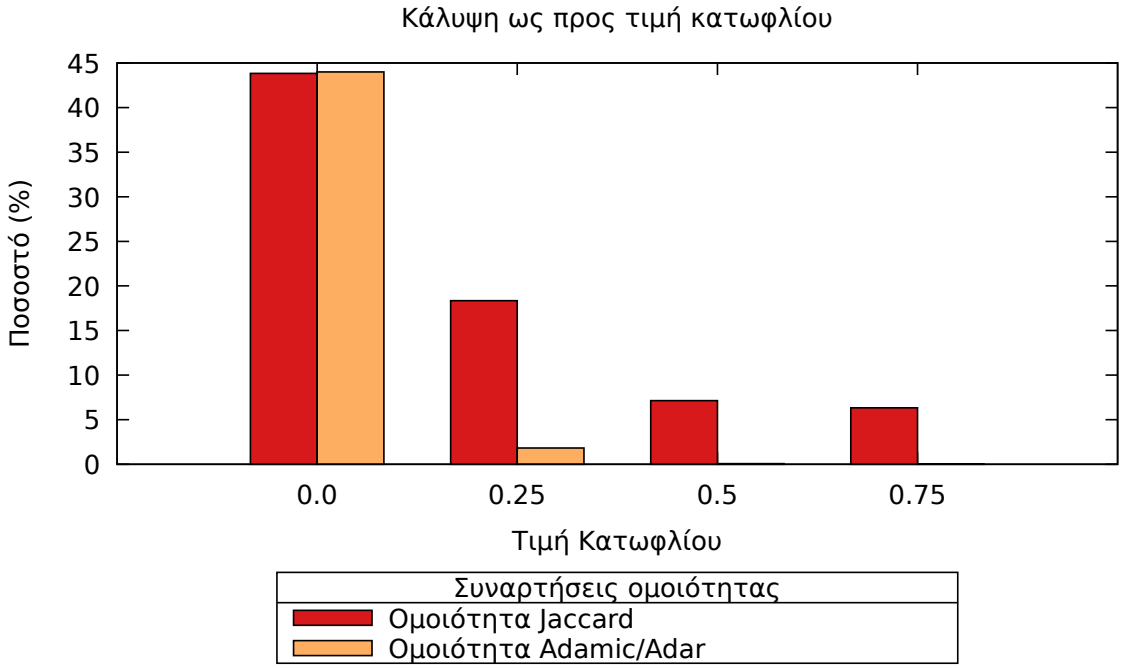
Σχήμα 4.11: MAP of Good @ 5 συναρτήσει του κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας και μέγεθος γειτονιάς ίσο με 10



Σχήμα 4.12: MAP of Good @ 5 συναρτήσει του κατωφλίου ομοιότητας για συντελεστές Jaccard και Adamic-Adar

Jaccard και Adamic-Adar, οι οποίοι παρουσιάστηκαν στη Ενότητα 3.3.3. Στο Σχήμα 4.12 απεικονίζεται η μέση ακρίβεια ως προς την τιμή κατωφλίου ομοιότητας για τις διαφορετικές συναρτήσεις ομοιότητας και πρόβλεψης. Όπως είναι φανερό, η ακρίβεια κυμαίνεται σε χαμηλές τιμές ακόμα και για υψηλές τιμές του κατωφλίου ομοιότητας φτάνοντας σε ένα μέγιστο 40%. Αυτό είναι αναμενόμενο αν αναλογιστεί κανείς ότι στα συστήματα αυτά οι όμοιοι χρήστες υπολογίζονται μόνο με βάση

το κοινωνικό δίκτυο ενώ δε λαμβάνονται καθόλου υπόψη οι βαθμολογίες τους.



Σχήμα 4.13: Κάλυψη συναρτήσει του κατωφλίου ομοιότητας για συντελεστές Jaccard και Adamic-Adar

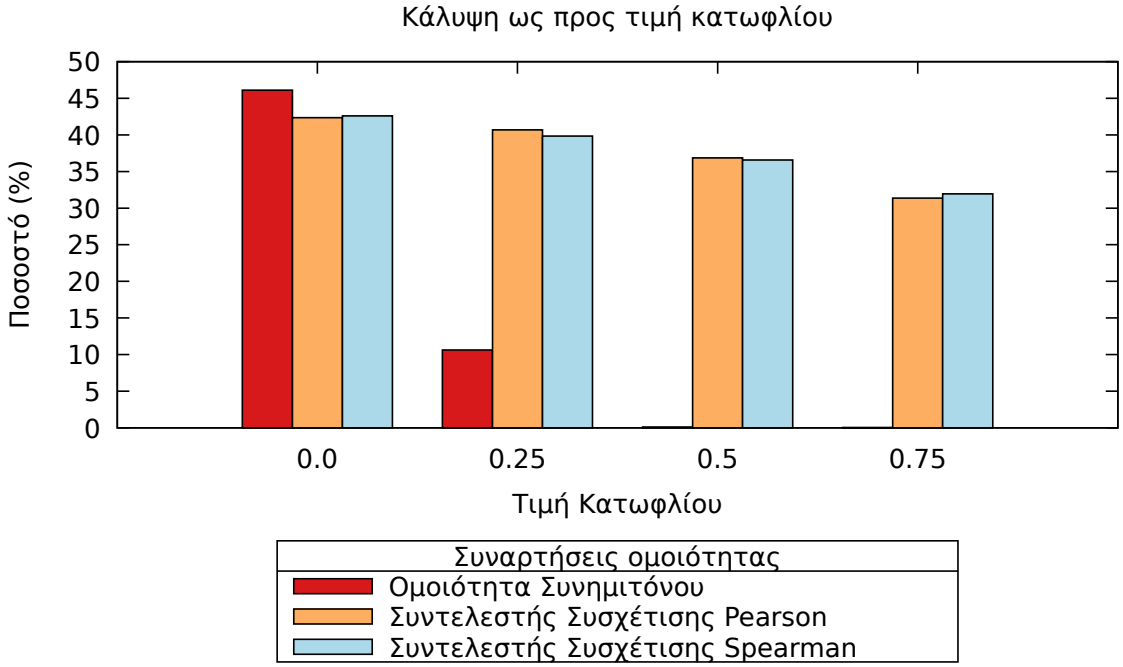
Μια άλλη παρατήρηση από το εν λόγω σχήμα είναι ότι ενώ για το συντελεστή Jaccard έχουμε αύξηση της ακρίβειας για υψηλότερες τιμές του κατωφλίου ομοιότητας, αυτό δε συμβαίνει στην περίπτωση του Adamic-Adar, όπου έχουμε μείωση. Από τη σκοπιά της ανάλυσης κοινωνικών δικτύων, μια εξήγηση που μπορεί να δοθεί είναι ότι οι βαθμολογίες των χρηστών επηρεάζονται περισσότερο από τις αξιολογήσεις των δημοφιλών χρηστών παρά από αυτές χρηστών οι οποίοι δεν έχουν πολλές ακμές και γι' αυτό το λόγο ο συντελεστής Adamic-Adar δεν παρουσιάζει καλά αποτελέσματα, αφού δίνει έμφαση στους χρήστες οι οποίο έχουν λίγες ακμές. Επίσης, η συνάρτηση πρόβλεψης δε φαίνεται να επηρεάζει την ακρίβεια του συγκεκριμένου συστήματος σε σημαντικό βαθμό.

Εξ' ορισμού, ο συντελεστής Adamic-Adar μεταξύ δύο χρηστών είναι δυσκολότερο να έχει υψηλή τιμή σε σχέση με αυτόν του Jaccard, αφού απαιτεί εκτός από κοινούς γείτονες οι χρήστες-κόμβοι να μην έχουν μεγάλο βαθμό στο δίκτυο. Αυτό έχει ως αποτέλεσμα η ομοιότητα Adamic-Adar να παρουσιάζει επίσης μικρότερη κάλυψη για υψηλότερες τιμές του κατωφλίου ομοιότητας πράγμα το οποίο επιβεβαιώνεται από το Σχήμα 4.13.

4.4.5 Το προτεινόμενο σύστημα κοινωνικής σύστασης

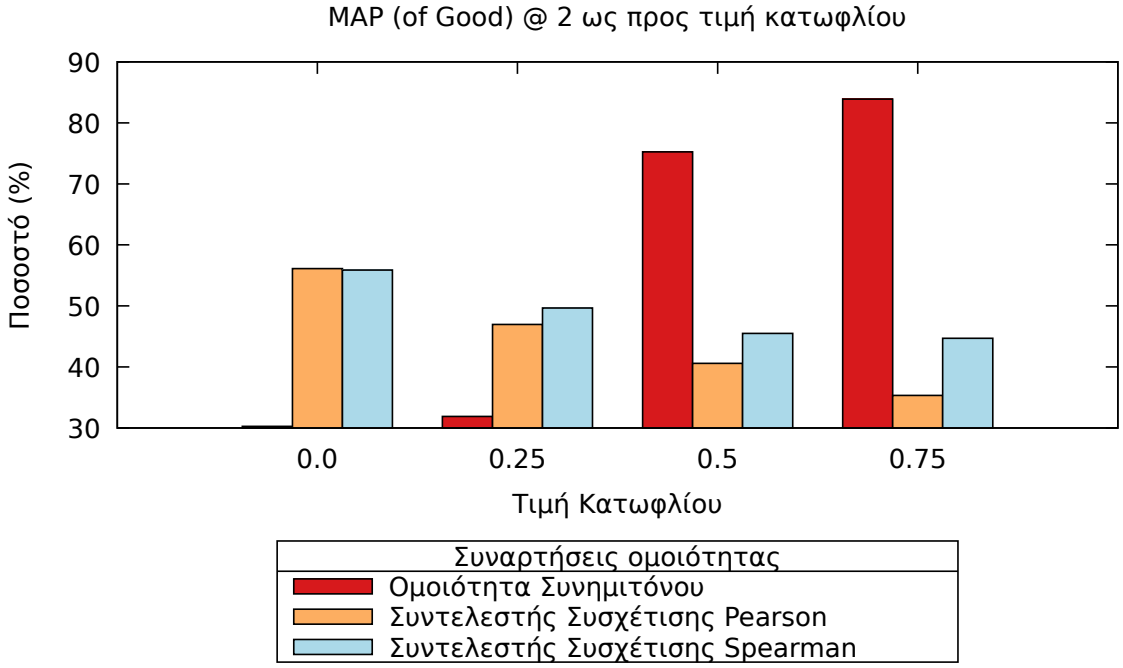
Ο τελευταίος τύπος συστήματος που υλοποιήθηκε είναι το προτεινόμενο από την παρούσα εργασία σύστημα, δηλαδή το σύστημα συνεργατικής διήθησης που βασίζεται στα αποτελέσματα του αλγορίθμου ανίχνευσης κοινοτήτων Newman-Girvan. Για το σύστημα αυτό πραγματοποιήθηκαν τα ίδια πειράματα όπως και στα συστήματα που παρουσιάστηκαν παραπάνω, για τις ίδιες τιμές των παραμέτρων.

Στο Σχήμα 4.14 απεικονίζεται η κάλυψη ως προς την τιμή του κατωφλίου ομοιότητας για τις διάφορες συναρτήσεις ομοιότητας. Η πρώτη παρατήρηση είναι ότι όσο αυξάνεται το κατώφλι μειώνεται το ποσοστό κάλυψης, σε μικρότερο βαθμό για τους συντελεστές συσχέτισης Pearson και Spearman και σε πολύ μεγαλύτερο για την ομοιότητα συνημιτόνου. Η συμπεριφορά αυτή είναι παρόμοια με τα βασισμένα στο χρήστη συστήματα συνεργατικής διήθησης και οφείλεται στους λόγους οι οποίοι αναφέρθηκαν στην Ενότητα 4.4.2. Η δεύτερη παρατήρηση είναι πως το ποσοστό κάλυψης είναι, ακόμα



Σχήμα 4.14: Κάλυψη συναρτήσει του κατωφλίου ομοιότητας για το προτεινόμενο σύστημα

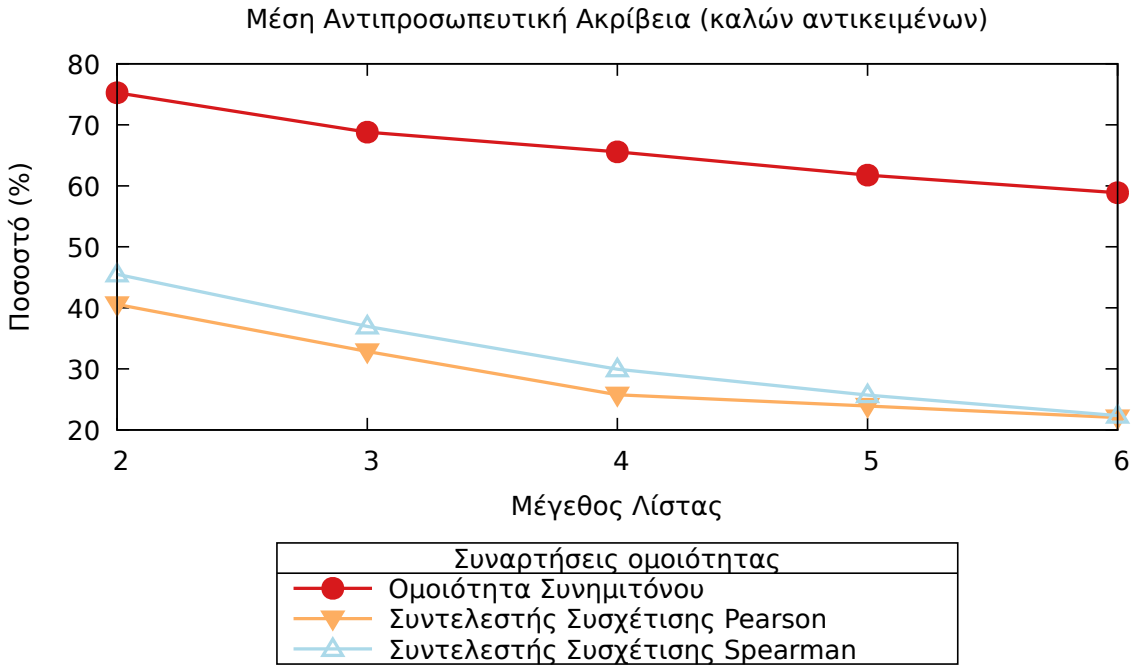
και για μηδενικό κατώφλι ομοιότητας, μειωμένο σε σχέση με το αντίστοιχο του Σχήματος 4.7. Αυτό οφείλεται στο γεγονός ότι περιορίζοντας το δίκτυο κάθε χρήστη στην κοινότητα όπου τοποθετήθηκε με βάση τα αποτελέσματα της διαδικασίας ανίχνευσης κοινοτήτων, γίνεται δυσκολότερο να βρεθούν χρήστες οι οποίοι να έχουν βαθμολογήσει από κοινού κάποιο αντικείμενο (απ' ότι θα ήταν αν για την εύρεση της γειτονιάς κάθε χρήστη λαμβανόταν υπόψη το συνολικό δίκτυο των χρηστών).



Σχήμα 4.15: MAP of Good @ 2 συναρτήσει του κατωφλίου ομοιότητας για το προτεινόμενο σύστημα

Εν συνεχεία, στο Σχήμα 4.15 απεικονίζεται η μέση ακρίβεια (των "καλών" αντικειμένων) για

μέγεθος προτεινόμενης λίστας ίσο με 2. Αυτό που παρατηρούμε είναι ότι αυξάνοντας την τιμή του κατωφλίου ομοιότητας, η ακρίβεια αυξάνεται για την ομοιότητα συνημιτόνου ενώ μειώνεται για τους συντελεστές συσχέτισης Pearson και Spearman. Ωστόσο, όπως και στην περίπτωση των συστημάτων συνεργατικής διήθησης βασισμένης στο χρήστη, οι πολύ υψηλές τιμές ακρίβειας για τη συνάρτηση συνημιτόνου, εξετάζοντας και το Σχήμα 4.14, συνδυάζονται με πολύ μικρή κάλυψη, επομένως και σε αυτή την περίπτωση υπάρχει ένας συμβιβασμός μεταξύ ακρίβειας και κάλυψης.



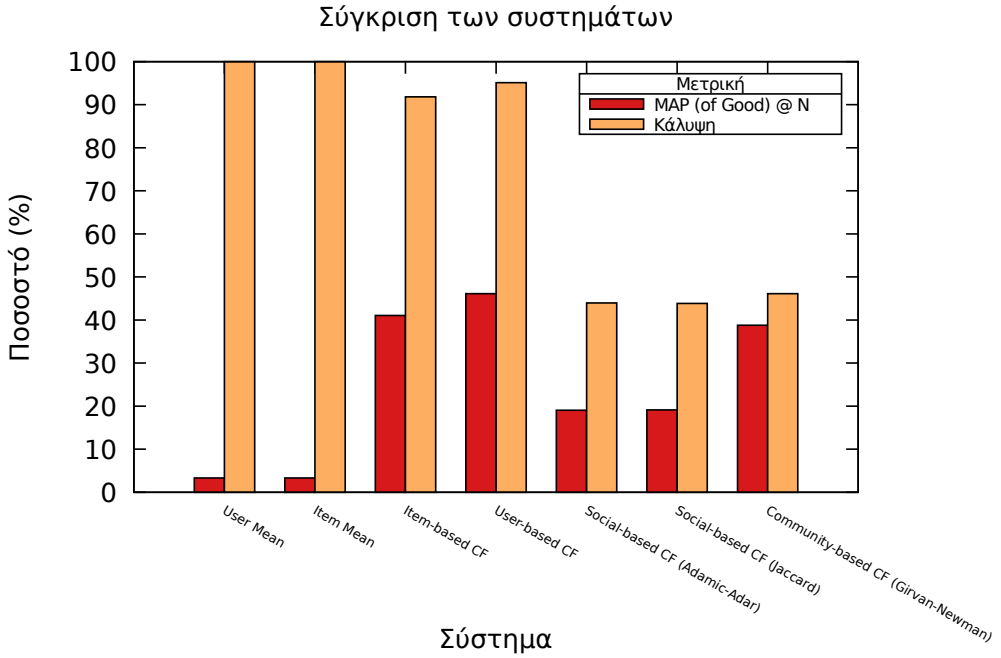
Σχήμα 4.16: Μέση Αντιπροσωπευτική Ακρίβεια (καλών αντικειμένων) συναρτήσει του μεγέθους λίστας για το προτεινόμενο σύστημα

Τέλος, στο Σχήμα 4.16 φαίνεται πάλι η μέση ακρίβεια (για τα "καλά" αντικείμενα), αυτή τη φορά συναρτήσει του μεγέθους προτεινόμενης λίστας, για τιμή κατωφλίου ομοιότητας ίση με 0.75. Η πρώτη παρατήρηση εδώ είναι ότι, όπως και στο Σχήμα 4.15, η συνάρτηση συνημιτόνου παρουσιάζει μεγαλύτερη μέση ακρίβεια από τις άλλες δύο. Η δεύτερη παρατήρηση είναι ότι όσο αυξάνεται το μέγεθος της προτεινόμενης λίστας μειώνεται η μέση ακρίβεια. Το φαινόμενο αυτό είναι λογικό από τη στιγμή που αυξάνοντας το μέγεθος της λίστας αυξάνεται ο αριθμός των αντικειμένων τα οποία προτείνονται αλλά δεν έχουν όντως βαθμολογηθεί από το χρήστη (false positives) με αποτέλεσμα ο παρανομαστής της ακρίβειας (Εξίσωση 4.3) να αυξάνεται και αυτή να μειώνεται.

4.4.6 Σύγκριση των συστημάτων που υλοποιήθηκαν

Οι καλύτερες διαμορφώσεις για κάθε ένα από τα παραπάνω συστήματα επιλέχθηκαν προκειμένου να πραγματοποιηθεί η σύγκριση μεταξύ αυτών και να εξαχθούν συμπεράσματα. Έτσι, στο Σχήμα 4.17 απεικονίζεται η κάλυψη και η μέση αντιπροσωπευτική ακρίβεια για κάθε ένα από τα παραπάνω συστήματα, για τον καλύτερο συνδυασμό των δύο μετρικών που αυτά πέτυχαν.

Το πρώτο συμπέρασμα που προκύπτει από τη γραφική αυτή είναι ότι τα συστήματα αναφοράς (user mean και item mean) παρότι παρουσιάζουν 100% κάλυψη διότι μπορούν πάντα να προτείνουν κάτι, υστερούν πολύ σε ακρίβεια έναντι των άλλων συστημάτων. Αυτό συμβαίνει γιατί, όπως περιγράψαμε και στην Ενότητα 4.4.1, προτείνουν τα ίδια top-N καλύτερα αντικείμενα σε όλους τους χρήστες, τα οποία όμως προφανώς δεν αντιπροσωπεύουν τις προτιμήσεις του καθενός. Γι' αυτό το λόγο, τα συστήματα αυτά χρησιμοποιούνται ως συστήματα βάσης σε πρακτικά υλοποιήσιμους αλγορίθμους, ώστε να μπορεί το σύστημα πάντα να παράγει κάποια πρόβλεψη. Επίσης, η συμπεριφορά αυτή είναι



Σχήμα 4.17: Σύγκριση των συστημάτων που εξετάστηκαν ως προς την Κάλυψη και την Μέση Αντιπροσωπευτική Ακρίβεια

δείγμα ότι οφείλει κάποιος να αξιολογεί τα συστήματα συστάσεων σε μια σειρά μετρικών και όχι με μια, γιατί έτσι μπορεί να οδηγηθεί σε λάθος συμπεράσματα, όπως για παράδειγμα λαμβάνοντας υπόψη στην προκειμένη περίπτωση μόνο την κάλυψη.

Συγκρίνοντας τα συστήματα συνεργατικής διήθησης που βασίζονται στην ανάλυση κοινωνικών δικτύων (social-based CF με ομοιότητας Jaccard ή Adamic-Adar) με τα υπόλοιπα συστήματα, αυτό που παρατηρούμε είναι ότι αυτά υστερούν και ως προς τις δύο μετρικές έναντι των υπολοίπων. Όσον αφορά την κάλυψη αυτό οφείλεται στο γεγονός ότι τα συστήματα αυτά ουσιαστικά δεν μπορούν να κάνουν προβλέψεις για χρήστες οι οποίοι δε συνδέονται με άλλους χρήστες στο κοινωνικό δίκτυο ή οι χρήστες με τους οποίους συνδέονται δεν έχουν αξιολογήσει τα ίδια αντικείμενα με αυτούς. Αναφορικά με τη μέση αντιπροσωπευτική ακρίβεια, αυτή όπως εξηγήσαμε και στη Ενότητα 4.4.4, είναι χαμηλή διότι η σύνδεση στο κοινωνικό δίκτυο αποτελεί απλά ένδειξη συνάφειας στις προτιμήσεις μεταξύ χρηστών και όχι απόδειξη. Επομένως, είναι αναμενόμενο τα συστήματα αυτά να υστερούν στην ακρίβεια σε σχέση με τα υπόλοιπα από τη στιγμή που δε λαμβάνουν καθόλου υπόψη τις βαθμολογίες κατά τον υπολογισμό των όμοιων χρηστών.

Οι καλύτεροι συνδυασμοί κάλυψης και μέσης αντιπροσωπευτικής ακρίβειας εντοπίζονται στα κλασσικά συστήματα συνεργατικής διήθησης, με λίγο υψηλότερες τιμές και στις δύο μετρικές αξιολόγησης να εμφανίζονται στα user-based συστήματα. Προφανώς, τα συστήματα αυτά παρουσιάζουν υψηλότερη κάλυψη από τα κοινωνικά συστήματα αφού λαμβάνουν υπόψη το σύνολο των χρηστών στον υπολογισμό των όμοιων χρηστών ενός χρήστη-στόχου. Η ελάχιστα υψηλότερη κάλυψη οφείλεται, όπως αναφέρθηκε και στην Ενότητα 4.4.3, στο γεγονός ότι το φαινόμενο long-tail παρατηρείται πιο έντονα στις αξιολογήσεις ανά αντικείμενο απ' ότι ανά χρήστη (Σχήμα 1.3). Βεβαίως, τα δύο αυτά συστήματα παρουσιάζουν και υψηλή ακρίβεια, εκμεταλλευόμενα όλα τα πλεονεκτήματα των συστημάτων συνεργατικής διήθησης που βασίζονται στη μνήμη, τα οποία αναφέρθηκαν στην Ενότητα 2.2.1.

Όσον αφορά το προτεινόμενο σύστημα κοινωνικής σύστασης (comunity-based CF), αυτό που παρατηρούμε στο Σχήμα 4.17 είναι ότι παρουσιάζει εξίσου υψηλή τιμή μέσης ακρίβειας με τα κλασικά συστήματα συνεργατικής διήθησης. Το σύστημα αυτό λαμβάνει υπόψη και τη θέση του χρήστη-στόχου στο κοινωνικό δίκτυο αλλά και τις βαθμολογίες των χρηστών και σε αυτό το συνδυασμό

φαίνεται να οφείλεται η εξίσου καλή ακρίβεια που επιτυγχάνεται. Από την άλλη, το σύστημα αυτό υστερεί στην μετρική της κάλυψης έναντι των κλασικών συστημάτων συνεργατικής διήθησης, γεγονός όμως το οποίο οφείλεται κατά κύριο λόγο στα χαρακτηριστικά της συγκεκριμένης συλλογής δεδομένων που χρησιμοποιήθηκε, αφού όπως είδαμε και στην Ενότητα 4.2.2, πολλοί από τους χρήστες δε συμμετέχουν στο κοινωνικό δίκτυο.

Κεφάλαιο 5

Συμπεράσματα και μελλοντικές κατευθύνσεις

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία ερευνήθηκε κατά πόσο η ανίχνευση κοινοτήτων σε ένα κοινωνικό δίκτυο όπως το Twitter μπορεί να συμβάλλει στη βελτίωση της ποιότητας ενός συστήματος συστάσεων συνεργατικής διήθησης. Από τα αποτελέσματα που παρουσιάστηκαν στο Κεφάλαιο 4 προκύπτουν αρκετά συμπεράσματα τόσο για το κοινωνικό δίκτυο και την ανάλυσή του αλλά και για τους διάφορους τύπους των συστημάτων συστάσεων που υλοποιήθηκαν, ανάμεσά τους και το κοινωνικό σύστημα που προτείνουμε.

Το πρώτο συμπέρασμα στο οποίο οδηγούμαστε από τα αποτελέσματα του αλγορίθμου ανίχνευσης κοινοτήτων Girvan-Newman είναι ότι το κοινωνικό δίκτυο Twitter δεν έχει κοινοτική δομή. Το Twitter, σε αντίθεση με δίκτυα που βασίζονται σε σχέσεις φιλίας όπως είναι το Facebook, είναι ένα δίκτυο βασισμένο σε σχέσεις ακολούθων. Αυτό έχει ως αποτέλεσμα να υπάρχουν χρήστες οι οποίοι συγκεντρώνουν μεγάλο πλήθος ακολούθων, λ.χ. επειδή είναι διάσημοι. Η πραγματικότητα αυτή δεν ευνοεί ιδιαίτερα τη δημιουργία κοινοτήτων. Εστιάζοντας στο χρησιμοποιηθέν σύνολο δεδομένων της παρούσας εργασίας, το οποίο περιλάμβανε χρήστες του Twitter οι οποίοι απλώς έχουν βαθμολογήσει ταινίες μέσω του IMDb χωρίς όμως να προϋποτίθεται κάποια άλλη σχέση μεταξύ τους, ήταν φυσικό επόμενο να υπάρχουν πολλοί χρήστες οι οποίοι συνδέονται ελάχιστα ή καθόλου με άλλους χρήστες δημιουργώντας έτσι ένα πολύ αραιό δίκτυο το οποίο έκανε πολύ δύσκολη τη διαδικασία ανίχνευσης κοινοτήτων.

Όσον αφορά το κομμάτι της σύστασης, πάλι η σημαντικότερη δυσκολία είναι η αραιότητα του πίνακα βαθμολογιών (πυκνότητα βαθμολογίας $0,048\%$) αλλά και το φαινόμενο long-tail (Σχήμα 1.3). Επομένως εξ' αρχής το πρόβλημα της σύστασης ήταν αρκετά δύσκολο με αυτά τα δύο δεδομένα. Σε αυτές τις δύο δυσκολίες οφείλονται τα χαμηλά ποσοστά κάλυψης των αλγορίθμων συνεργατικής διήθησης που παρουσιάστηκαν στην Ενότητα 4.4, για περιπτώσεις όπου δεν υπήρχε σύστημα αναφοράς. Για τον ίδιο λόγο επίσης το μέγεθος της γειτονιάς των όμοιων χρηστών ενός χρήστη δεν έδειξε να επηρεάζει τα αποτελέσματα στη βασισμένη στο χρήστη συνεργατική διήθηση, αφού για ελάχιστους χρήστες μπορούσαν να βρεθούν πάνω από 2-3 παρόμοιοι χρήστες για ένα μεσαίας τιμής κατώφλι ομοιότητας.

Παρά τις παραπάνω δυσκολίες, παρατηρώντας τα αποτελέσματα της σύγκρισης του προτεινόμενου συστήματος κοινωνικής σύστασης σε σχέση με την απλή συνεργατική διήθηση βασισμένη στο χρήστη, είναι ξεκάθαρο ότι η μέση ακρίβεια για μικρές προτεινόμενες λίστες αντικειμένων εξίσου υψηλή όταν οι όμοιοι χρήστες προέρχονται αποκλειστικά από την κοινότητα στην οποία ανήκει ο χρήστης. Αυτό επιβεβαιώνεται και από το Σχήμα 4.17, αν παρατηρήσει κανείς την απόδοση των σχετικών μετρικών. Γενικότερα, η μέση αντιπροσωπευτική ακρίβεια θεωρείται από τις πιο σημαντικές μετρικές στα συστήματα συστάσεων αφού πρώτον είναι πιο κοντά στα πραγματικά συστήματα τα οποία συνήθως προτείνουν λίστες αντικειμένων και όχι μεμονωμένα αντικείμενα και δεύτερον δίνει έμφαση στη σειρά κατάταξης των αντικειμένων [Agga16].

Ωστόσο, το σημείο στο οποίο υστερεί το προτεινόμενο σύστημα σε σχέση με την κλασική συνεργατική διήθηση βασισμένη στο χρήστη είναι η κάλυψη, σε περιπτώσεις πάντα όπου δεν υπάρχει σύστημα αναφοράς. Αυτό οφείλεται στο γεγονός ότι η ήδη περιορισμένη κάλυψη λόγω των φαινομένων αραιότητας του πίνακα βαθμολογιών και long-tail, επηρεάζεται ακόμα περισσότερο αν, χωρίζο-

ντας το δίκτυο σε κοινότητες, ορισμένοι χρήστες οι οποίοι για παράδειγμα συνδέονταν μόνο με 2-3 άλλους χρήστες βρεθούν σε διαφορετική κοινότητα από αυτούς. Βέβαια, δεδομένου ότι όλα τα συστήματα χρησιμοποιούν ένα σύστημα αναφοράς (όπως ο μέσος όρος βαθμολογιών του χρήστη ή του αντικειμένου) για τις περιπτώσεις όπου αποτυγχάνει ο βασικός αλγόριθμος, η μείωση αυτή δεν είναι καταδικαστική για το προτεινόμενο σύστημα, εφόσον αυτό φαίνεται να λειτουργεί καλά σε συνδυασμό με το σύστημα αναφοράς, όπως φαίνεται χρησιμοποιώντας τον τύπο του Resnick σαν συνάρτηση πρόβλεψης.

Όσον αφορά τα συστήματα κοινωνικής σύστασης τα οποία χρησιμοποιούν την ομοιότητα Jaccard και Adamic-Adar αυτά είναι λογικό να παρουσιάζουν χαμηλή κάλυψη αλλά και ακρίβεια εφόσον δε λαμβάνουν καθόλου υπόψη τις προσδιορισμένες βαθμολογίες των χρηστών αλλά μόνο τις συνδέσεις τους μέσα στο δίκτυο. Έτσι είναι λογικό αφενός να μην είναι εύκολο να βρεθούν πάντα οι όμοιοι χρήστες, πόσο μάλλον σε έναν αραιό γράφο όπως αυτός που εξετάστηκε, και αφετέρου οι προβλεπόμενες βαθμολογίες να παρουσιάζουν υψηλό σφάλμα σε πολλές περιπτώσεις. Από τη μεταξύ τους σύγκριση επίσης, λαμβάνοντας υπόψη τη μειωμένη ακρίβεια που παρουσιάζει το σύστημα με ομοιότητα Adamic-Adar, καταλήγουμε στο συμπέρασμα ότι ο συντελεστής αυτός δεν είναι κατάλληλος για τον υπολογισμό ομοιότητας μεταξύ χρηστών σε κοινωνικά συστήματα συστάσεων αφού δίνει έμφαση στις βαθμολογίες χρηστών με λίγες συνδέσεις ενώ στην πραγματικότητα οι περισσότεροι χρήστες δείχνουν να επηρεάζονται περισσότερο από δημοφιλείς χρήστες.

Τέλος, η ενσωμάτωση της ανίχνευσης κοινοτήτων στα συστήματα συστάσεων, με τον τρόπο που παρουσιάστηκε στην παρούσα εργασία (αλγόριθμος Girvan-Newman), έδειξε ότι μπορεί να προσφέρει στην ακρίβεια ενός συστήματος σύστασης. Πράγματι, ο αλγόριθμος αυτός αποτελεί έναν από τους πλέον διαδεδομένους και αποτελεσματικούς αλγορίθμους ανίχνευσης κοινοτήτων [Fort10] και παρά την αραιότητα του υπό εξέταση δικτύου, οι τελικές κοινότητες που προέκυψαν από τον αλγόριθμο εκ του αποτελέσματος φάνηκε να περιέχουν σημαντική πληροφορία σχετικά με τις προτιμήσεις των χρηστών.

5.2 Μελλοντικές Κατευθύνσεις

Η παρούσα διπλωματική εργασία ασχολείται με τα πεδία της ανάλυσης κοινωνικών δικτύων και των συστημάτων συστάσεων τα οποία είναι σε συνεχή εξέλιξη τα τελευταία χρόνια, επομένως υπάρχουν πολλές κατευθύνσεις επέκτασης της εργασίας αυτής. Αρχικά, όσον αφορά τα συστήματα συστάσεων, η συνεργατική διήθηση βασισμένη στη μνήμη μπορεί να είναι από τις πρώτες και πιο χρησιμοποιημένες τεχνικές σύστασης, αλλά έχει αποδειχθεί ότι οι μέθοδοι που χρησιμοποιούν μοντέλα αποτελούν την αιχμή των συστημάτων συστάσεων. Επομένως, μια δυνατότητα που πιθανόν να βελτίωνε την ποιότητα των συστάσεων είναι να δημιουργηθεί ένα σύστημα κοινωνικής σύστασης το οποίο θα αξιοποιεί την πληροφορία της ανίχνευσης κοινοτήτων σε ένα σύστημα βασισμένο σε κάποιο μοντέλο όπως για παράδειγμα ένα μοντέλο λανθανουσών παραγόντων.

Ένα πρόβλημα που παρατηρήθηκε στο προτεινόμενο σύστημα κοινωνικής σύστασης είναι ότι η επιλογή της γειτονιάς του χρήστη μόνο από τους χρήστες οι οποίοι ανήκουν στην ίδια με αυτόν κοινότητα περιορίζει την κάλυψη του συστήματος αφού για πολλούς χρήστες δεν μπορούν να βρεθούν άλλοι, οι οποίοι να έχουν βαθμολογήσει τα ίδια αντικείμενα. Για την αντιμετώπιση αυτού του φαινομένου, μια ιδέα θα ήταν ο υπολογισμός των γειτονικών να χρηστών να γίνεται από το σύνολο των χρηστών του δικτύου αλλά να δίνεται κάποια μεγαλύτερη βαρύτητα αν οι δύο χρήστες ανήκουν στην ίδια κοινότητα. Με αυτό τον τρόπο θα μπορούσε να επιτευχθεί σίγουρα μεγαλύτερο ποσοστό κάλυψης του αλγορίθμου, ωστόσο θα αναιρούταν η επιτάχυνση η οποία εξασφαλίζεται στον αλγόριθμο όταν το δίκτυο διαιρείται σε κοινότητες.

Μια άλλη κατεύθυνση έρευνας που μπορεί να βελτιώσει την ποιότητα ενός συστήματος κοινωνικής σύσταση όπως αυτό που προτείνεται είναι η επιπλέον μελέτη και κατάταξη του τύπου των σχέσεων μεταξύ των χρηστών του δικτύου. Τα περισσότερα κοινωνικά συστήματα συστάσεων όπως και αυτό που υλοποιήθηκε, υποθέτουν ότι σύνδεση ενός χρήστη με έναν άλλο σημαίνει ότι οι χρήστες αυτοί πιθανόν να μοιράζονται τα ίδια ενδιαφέροντα και να αρέσκονται στα ίδια αντικείμενα. Ωστόσο,

οι σχέσεις σε ένα κοινωνικό δίκτυο μπορεί να είναι πολλών τύπων [Tang12]. Για παράδειγμα, ένας χρήστης μπορεί να εμπιστεύεται έναν άλλο χρήστη σε έναν τομέα αλλά όχι σε ένα δεύτερο. Στο σύστημα που πραγματεύεται η παρούσα εργασία π.χ. θα ήταν πολύ χρήσιμο αν με κάποιο τρόπο ήταν γνωστό ότι ο χρήστης Α εμπιστεύεται περισσότερο το χρήστη Β στην επιλογή ταινιών από ότι το χρήστη Γ παρότι συνδέεται και με τους δύο. Επομένως, η μελέτη και ο προσδιορισμός του τύπου των σχέσεων μεταξύ των χρηστών ενός κοινωνικού δικτύου θα μπορούσε να βελτιώσει την ποιότητα των συστάσεων ενός κοινωνικού συστήματος.

Τέλος, η παραπάνω κατεύθυνση θα μπορούσε να μελετηθεί και να εφαρμοστεί και στο κομμάτι της ανίχνευσης κοινοτήτων. Αντί δηλαδή οι κοινότητες να είναι απλά μια ομαδοποίηση των κόμβων του γράφου χωρίς περαιτέρω περιεχόμενο, να αποτελούν εξ' αρχής κοινότητες εμπιστοσύνης σχετικά με την επιλογή ταινιών, αν θεωρήσουμε το σύστημα που πραγματεύεται η παρούσα εργασία. Έτσι, το κοινωνικό σύστημα σύστασης θα έχει μια πιο ολοκληρωμένη κοινωνική πληροφορία να αξιοποιήσει για τις συστάσεις του με αποτέλεσμα να είναι πιο ακριβές στις προβλέψεις του.

Βιβλιογραφία

[Adam03] Lada A Adamic and Eytan Adar, "Friends and neighbors on the web", *Social networks*, vol. 25, no. 3, pp. 211–230, 2003.

[Agga01] Charu C Aggarwal and Srinivasan Parthasarathy, "Mining massively incomplete data sets by conceptual reconstruction", in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 227–232, ACM, 2001.

[Agga11] Charu Aggarwal, "Social Network Data Analytics", 2011.

[Agga15] Charu C Aggarwal, "Data mining: the textbook", 2015.

[Agga16] Charu C Aggarwal, *Recommender Systems: The Textbook*, Springer, 2016.

[Andr76] George E Andrews, "The Theory of Partitions, volume 2 of Encyclopedia of Mathematics and its Applications", *Addison-Wesley*, vol. 19, no. 76, p. 18, 1976.

[Barn54] John Arundel Barnes, "Class and committees in a Norwegian island parish", *Human relations*, vol. 7, no. 1, pp. 39–58, 1954.

[boyd07] danah boyd and Nicole B. Ellison, "Social network sites: Definition, history, and scholarship", *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[Bras98] Daniel J Brass, Kenneth D Butterfield and Bruce C Skaggs, "Relationships and unethical behavior: A social network perspective", *Academy of Management Review*, vol. 23, no. 1, pp. 14–31, 1998.

[Chun97] Fan RK Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.

[Clau04] Aaron Clauset, Mark EJ Newman and Cristopher Moore, "Finding community structure in very large networks", *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[Clau09] Aaron Clauset, Cosma Rohilla Shalizi and Mark EJ Newman, "Power-law distributions in empirical data", *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[Dijk59] Edsger W Dijkstra, "A note on two problems in connexion with graphs", *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[Doom13] Simon Dooms, Toon De Pessemier and Luc Martens, "MovieTweetings: a Movie Rating Dataset Collected From Twitter", in *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013*, 2013.

[Erdo59] Paul Erdős and Alfréd Rényi, "On random graphs", *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[Fled07] Daniel M Fleder and Kartik Hosanagar, "Recommender systems and their impact on sales diversity", in *Proceedings of the 8th ACM conference on Electronic commerce*, pp. 192–199, ACM, 2007.

[Fort10] Santo Fortunato, "Community detection in graphs", *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[Free77] Linton C Freeman, "A set of measures of centrality based on betweenness", *Sociometry*, pp. 35–41, 1977.

[Free04] Linton Freeman, "The development of social network analysis", *A Study in the Sociology of Science*, 2004.

[Frie01] Jerome Friedman, Trevor Hastie and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics Springer, Berlin, 2001.

[Golb06] Jennifer Golbeck, "Generating predictive movie recommendations from trust in social networks", in *International Conference on Trust Management*, pp. 93–104, Springer, 2006.

[Golb09] Jennifer Golbeck, "Trust and nuanced profile similarity in online social networks", *ACM Transactions on the Web (TWEB)*, vol. 3, no. 4, p. 12, 2009.

[Golb13] Jennifer Golbeck, *Analyzing the social web*, Newnes, 2013.

[Good99] Nathaniel Good, J Ben Schafer, Joseph A Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, John Riedl et al., "Combining collaborative filtering with personal agents for better recommendations", in *AAAI/IAAI*, pp. 439–446, 1999.

[Good10] Benjamin H Good, Yves-Alexandre de Montjoye and Aaron Clauset, "Performance of modularity maximization in practical contexts", *Physical Review E*, vol. 81, no. 4, p. 046106, 2010.

[Gori07] Marco Gori, Augusto Pucci, V Roma and I Siena, "ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines.", in *IJCAI*, vol. 7, pp. 2766–2771, 2007.

[Guha04] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan and Andrew Tomkins, "Propagation of trust and distrust", in *Proceedings of the 13th international conference on World Wide Web*, pp. 403–412, ACM, 2004.

[Guy15] Ido Guy, "Social recommender systems", in *Recommender Systems Handbook*, pp. 511–543, Springer, 2015.

[Herl99] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers and John Riedl, "An Algorithmic Framework for Performing Collaborative Filtering", in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 230–237, New York, NY, USA, 1999, ACM.

[Howe08] Adele E. Howe and Ryan D. Forbes, "Re-considering Neighborhood-based Collaborative Filtering Parameters in the Context of New Data", in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pp. 1481–1482, New York, NY, USA, 2008, ACM.

[Jacc01] Paul Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[Jama09] Mohsen Jamali and Martin Ester, "Trustwalker: a random walk model for combining trust-based and item-based recommendation", in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 397–406, ACM, 2009.

[Katz53] Leo Katz, "A new status index derived from sociometric analysis", *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[Kaus15] Shaivya Kaushik and Pradeep Tomar, "Evaluation of Similarity Functions by using User based Collaborative Filtering approach in Recommendation Systems", 2015.

[Kaut97] Henry Kautz, Bart Selman and Mehul Shah, "Referral Web: combining social networks and collaborative filtering", *Communications of the ACM*, vol. 40, no. 3, pp. 63–65, 1997.

[Kern70] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs", *The Bell System Technical Journal*, vol. 49, no. 2, pp. 291–307, Feb 1970.

[Koha95] Ron Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection", in *Ijcai*, vol. 14, pp. 1137–1145, Stanford, CA, 1995.

[Lova93] László Lovász, "Combinatorial problems and exercises", 1993.

[Mass04] Paolo Massa and Paolo Avesani, "Trust-aware collaborative filtering for recommender systems", in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 492–508, Springer, 2004.

[McLa05] Geoffrey McLachlan, Kim-Anh Do and Christophe Ambroise, *Analyzing microarray gene expression data*, vol. 422, John Wiley & Sons, 2005.

[McPh01] Miller McPherson, Lynn Smith-Lovin and James M Cook, "Birds of a feather: Homophily in social networks", *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[Mood03] James Moody and Douglas R White, "Structural cohesion and embeddedness: A hierarchical concept of social groups", *American Sociological Review*, pp. 103–127, 2003.

[Newm04a] Mark EJ Newman, "Detecting community structure in networks", *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[Newm04b] Mark EJ Newman, "Fast algorithm for detecting community structure in networks", *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[Newm04c] Mark EJ Newman and Michelle Girvan, "Finding and evaluating community structure in networks", *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[Newm05] Mark EJ Newman, "Power laws, Pareto distributions and Zipf's law", *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[Otte02] Evelien Otte and Ronald Rousseau, "Social network analysis: a powerful strategy, also for the information sciences", *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[Podo97] Joel M Podolny and James N Baron, "Resources and relationships: Social networks and mobility in the workplace", *American sociological review*, pp. 673–693, 1997.

[Powe11] David Martin Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", 2011.

[Radi04] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto and Domenico Parisi, "Defining and identifying communities in networks", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[Ravi96] Si Pi Ravikumār, *Parallel methods for VLSI layout design*, Greenwood Publishing Group, 1996.

[Resn94] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom and John Riedl, "GroupLens: an open architecture for collaborative filtering of netnews", in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, ACM, 1994.

[Ricc11] Francesco Ricci, Lior Rokach and Bracha Shapira, *Introduction to recommender systems handbook*, Springer, 2011.

[Scha01] J Ben Schafer, Joseph A Konstan and John Riedl, "E-commerce recommendation applications", in *Applications of Data Mining to Electronic Commerce*, pp. 115–153, Springer, 2001.

[Shan11] Guy Shani and Asela Gunawardana, *Evaluating Recommendation Systems*, pp. 257–297, Springer US, Boston, MA, 2011.

[Stra03] G. Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press, 2003.

[Suar88] Peter R Suaris and Gershon Kedem, "An algorithm for quadrisection and its application to standard cell placement", *IEEE Transactions on Circuits and Systems*, vol. 35, no. 3, pp. 294–303, 1988.

[Tang12] Jiliang Tang, Huiji Gao and Huan Liu, "mTrust: discerning multi-faceted trust in a connected world", in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 93–102, ACM, 2012.

[Trav67] Jeffrey Travers and Stanley Milgram, "The small world problem", *Phychology Today*, vol. 1, pp. 61–67, 1967.

[Tyle05] Joshua R Tyler, Dennis M Wilkinson and Bernardo A Huberman, "E-mail as spectroscopy: Automated discovery of community structure within organizations", *The Information Society*, vol. 21, no. 2, pp. 143–153, 2005.

[VonL06] Ulrike Von Luxburg, "A tutorial on spectral clustering. Max Planck Institute for Biological Cybernetics", *Tech Rep*, 2006.

[Wass94] Stanley Wasserman and Katherine Faust, "Social Network Analysis in the Social and Behavioral Sciences", in *Social Network Analysis: Methods and Applications*, pp. 3–27, Cambridge University Press, Cambridge, 11 1994.

[Watt98] Duncan J Watts and Steven H Strogatz, "Collective dynamics of 'small-world'networks", *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[Weng10] Jianshu Weng, Ee-Peng Lim, Jing Jiang and Qi He, "Twitterrank: finding topic-sensitive influential twitterers", in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270, ACM, 2010.

[Zach77] Wayne W Zachary, "An information flow model for conflict and fission in small groups", *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[Zhan12] Peng Zhang and Wanhua Su, "Statistical inference on recall, precision and average precision under random selection", in *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pp. 1348–1352, IEEE, 2012.

[Zieg07] Cai-Nicolas Ziegler and Jennifer Golbeck, "Investigating interactions of trust and interest similarity", *Decision support systems*, vol. 43, no. 2, pp. 460–475, 2007.