ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μοντελοποίηση και ανάλυση κοινωνικών δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΜΑΡΜΑΡΕΛΛΗ

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

Αθήνα, Ιούλιος 2014

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Μοντελοποίηση και ανάλυση κοινωνικών δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΚΩΝΣΤΑΝΤΙΝΟΥ ΜΑΡΜΑΡΕΛΛΗ

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 22 Ιουλίου 2014.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης	Στέφανος Κόλλιας	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2014

(Υπογραφή)

.....

ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΑΡΜΑΡΕΛΛΗΣ

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

Περίληψη

Η επιτυχία των online κοινωνικών δικτύων τα τελευταία χρόνια έχει καταστήσει σημαντική τη μοντελοποίηση και ανάλυση των κοινωνικών δικτύων γενικότερα. Τα online κοινωνικά δίκτυα προσφέρουν μία ασύλληπτη ποσότητα πληροφορίας, επιτρέποντας τη διεξαγωγή μελετών που ήταν αδύνατες στο παρελθόν και τον έλεγχο κοινωνικών θεωριών.

Σκοπός αυτής της διπλωματικής εργασίας είναι η παρουσίαση και ταξινόμηση μοντέλων που έχουν προταθεί για τα κοινωνικά δίκτυα και μεθόδων που χρησιμοποιούνται για την επίλυση δύο σημαντικών προβλημάτων της ανάλυσης κοινωνικών δικτύων, της πρόβλεψης συνδέσμου και της ανίχνευσης κοινοτήτων. Επίσης, η παρουσίαση και ταξινόμηση μεθόδων που χρησιμοποιούνται σε μία πιο εφαρμοσμένη περιοχή της ανάλυσης κοινωνικών δικτύων, την κοινωνική σύσταση.

Αρχικά εκθέτουμε χαρακτηριστικά των κοινωνικών δικτύων, στατικά και δυναμικά. Στη συνέχεια ταξινομούμε και επισκοπούμε μοντέλα κοινωνικών δικτύων. Για καθεμία από τις προαναφερθείσες εφαρμογές ανάλυσης κοινωνικών δικτύων, δίνουμε επίσης δείκτες αξιολόγησης των προτεινόμενων μεθόδων και συζητούμε μελλοντικές κατευθύνσεις έρευνας.

Λέξεις Κλειδιά: <<Μοντέλα Κοινωνικών Δικτύων, Ανάλυση Κοινωνικών Δικτύων, Πρόβλεψη Συνδέσμου, Ανίχνευση Κοινοτήτων, Κοινωνική Σύσταση, Συστήματα Συστάσεων>>

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

Social network modeling and analysis has become important due to the success of online social networks during the past few years. Online social networks offer an incredible amount of information, thus allowing the performance of studies that were impossible in the past and social theories testing.

The purpose of this diploma thesis is to present and classify models proposed for social networks and methods used in two important social network analysis tasks, namely link prediction and community detection, as well as in a highly applied social network analysis field, namely social recommendation.

Firstly, we quote social network features, both static and dynamic. We then classify and review social network models. For each of the aforementioned social network analysis tasks we give evaluation metrics for the methods proposed and discuss future research directions.

Keywords: <<Social Network Models, Social Network Analysis, Link Prediction, Community Detection, Social Recommendation, Recommender Systems>>

Ευχαριστίες

Ευχαριστώ τον επιβλέποντα καθηγητή της διπλωματικής εργασίας, Καθηγητή Ε.Μ.Π. κύριο Ανδρέα-Γεώργιο Σταφυλοπάτη, τόσο για την εμπιστοσύνη που μου έδειξε όσο και για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον και σύγχρονο θέμα.
Ευχαριστώ και όλους όσοι, συνειδητά ή μη, με εμψύχωσαν κατά τη διάρκεια της εκπόνησης της εργασίας.

Ευχαριστώ πολύ!

Πίνακας περιεχομένων

1

Εισαγωγή

1.1 Κοινωνικά δίκτυα – πολύπλοκα δίκτυα

Ένα κοινωνικό δίκτυο είναι μία κοινωνική δομή που αποτελείται από ένα πεπερασμένο σύνολο διακεκριμένων οντοτήτων (entities) (π. χ. ατόμων, ομάδων ή οργανισμών), οι οποίες ονομάζονται ‘κόμβοι’ (nodes) ή ‘δράστες’ (actors) και συνδέονται με μία σχέση (π. χ. εμπιστοσύνης, φιλίας, γνωριμίας, επικοινωνίας ή συγγένειας). Οι συνδέσεις μεταξύ των κόμβων ονομάζονται ‘ακμές’ (edges) ή ‘δεσμοί’ (ties) ή ‘σύνδεσμοι’ (links).

Παραδείγματα κοινωνικών δικτύων είναι τα δίκτυα φιλίας, τα δίκτυα συνεργασίας, π. χ. μεταξύ ηθοποιών ή επιστημόνων (όπως τα δίκτυα συν-συγγραφής επιστημονικών άρθρων), τα δίκτυα τηλεφωνικών κλήσεων και τα δίκτυα αποστολής μηνυμάτων ηλεκτρονικού ταχυδρομείου.

Τα κοινωνικά δίκτυα ανήκουν στην ευρύτερη κατηγορία των λεγόμενων πολύπλοκων δικτύων (complex networks), τα οποία ορίζονται ως δίκτυα που παρουσιάζουν μη τετριμμένα τοπολογικά χαρακτηριστικά. Τα πολύπλοκα δίκτυα περιλαμβάνουν επίσης τα πληροφοριακά δίκτυα (όπως τα δίκτυα παραπομπών μεταξύ επιστημονικών άρθρων και ο Παγκόσμιος Ιστός), τα τεχνολογικά δίκτυα (όπως τα ηλεκτρικά δίκτυα, τα οδικά δίκτυα, τα σιδηροδρομικά δίκτυα, τα αεροπορικά δίκτυα, τα ηλεκτρονικά κυκλώματα, τα πακέτα λογισμικού, τα δίκτυα ομοτίμων και το Διαδίκτυο) και τα βιολογικά δίκτυα (όπως τα δίκτυα

αλληλεπίδρασης πρωτεϊνών, τα μονοπάτια μετάδοσης σήματος (signal transduction pathways), τα τροφικά πλέγματα (food webs), τα νευρωνικά δίκτυα και τα μεταβολικά δίκτυα).

1.2 Online κοινωνικά δίκτυα

Τα τελευταία χρόνια οι ιστότοποι κοινωνικής δικτύωσης (όπως τα Facebook, YouTube, LinkedIn, Twitter, Google+, MySpace και Instagram) βρίσκονται μεταξύ των δημοφιλέστερων του Διαδικτύου. Οι χρήστες αυτών των ιστοτόπων σχηματίζουν κοινωνικά δίκτυα, που παρέχουν ισχυρά μέσα διαμοιρασμού, οργάνωσης και εντοπισμού περιεχομένου και επαφών. Η δημοφιλία αυτών των ιστοτόπων παρέχει μία ευκαιρία να μελετηθούν τα χαρακτηριστικά των γράφων των online κοινωνικών δικτύων σε μεγάλη κλίμακα. Η κατανόηση αυτών των γράφων είναι σημαντική τόσο για τη βελτίωση των υφιστάμενων συστημάτων όσο και για τη σχεδίαση νέων εφαρμογών online κοινωνικών δικτύων.

Για να συμμετάσχουν πλήρως σε ένα online κοινωνικό δίκτυο, οι χρήστες πρέπει να εγγραφούν στον αντίστοιχο ιστότοπο, πιθανώς με ένα ψευδώνυμο. Κάποιοι ιστότοποι επιτρέπουν την εξερεύνηση δημόσιων δεδομένων χωρίς σύνδεση. Οι χρήστες μπορούν να δώσουν πληροφορίες για τον εαυτό τους (π. χ., ημερομηνία γέννησης, τόπο διαμονής, ενδιαφέροντα), οι οποίες συνιστούν το προφίλ τους.

Το online κοινωνικό δίκτυο συντίθεται από τους λογαριασμούς των χρηστών και τους συνδέσμους μεταξύ των χρηστών. Κάποιοι ιστότοποι επιτρέπουν στους χρήστες να συνδέονται με οποιονδήποτε άλλο χρήστη, χωρίς τη συναίνεσή του. Άλλοι απαιτούν τη συναίνεση και των δύο πλευρών για τη δημιουργία ενός συνδέσμου. Ορισμένοι ιστότοποι κοινωνικής δικτύωσης χαρακτηρίζονται από κατευθυνόμενους συνδέσμους (δηλαδή, ένας σύνδεσμος από το χρήστη Α στο χρήστη Β δε συνεπάγεται την παρουσία ενός αντίστροφου συνδέσμου), ενώ άλλοι έχουν κοινωνικά δίκτυα με μη κατευθυνόμενους συνδέσμους.

Οι περισσότεροι ιστότοποι δίνουν στους χρήστες τη δυνατότητα να δημιουργήσουν και να γίνουν μέλη ομάδων ειδικού ενδιαφέροντος. Οι χρήστες μπορούν να δημοσιεύσουν μηνύματα και να αναρτήσουν περιεχόμενο στις ομάδες. Κάποιες ομάδες είναι συντονιζόμενες (moderated). Η είσοδος και οι δημοσιεύσεις σε μια τέτοια ομάδα ελέγχονται από ένα χρήστη ορισμένο σαν συντονιστή (moderator) της ομάδας. Άλλες ομάδες είναι ανοικτές, επιτρέποντας σε οποιονδήποτε χρήστη να γίνει μέλος και να δημοσιεύσει μηνύματα ή περιεχόμενο.

1.3 Αντικείμενο – οργάνωση της εργασίας

Σε αυτήν τη διπλωματική εργασία παρουσιάζουμε και ταξινομούμε μοντέλα που έχουν προταθεί για τα κοινωνικά δίκτυα και λύσεις που έχουν δοθεί σε δημοφιλή προβλήματα ανάλυσης κοινωνικών δικτύων, όπως η πρόβλεψη συνδέσμου (link prediction) και η ανίχνευση κοινοτήτων (community detection), καθώς και στο πρόβλημα της αξιοποίησης της 'κοινωνικής πληροφορίας', που προέρχεται από online κοινωνικά δίκτυα, στα συστήματα συστάσεων (recommender systems).

Συγκεκριμένα, στο δεύτερο κεφάλαιο, αφού παρουσιάσουμε χαρακτηριστικά των κοινωνικών δικτύων που χαρακτηρίζουν τη στατική τους τοπολογία και τη χρονική τους εξέλιξη, επισκοπούμε διάφορα μοντέλα που έχουν εισαχθεί για τη μοντελοποίησή τους. Στο τρίτο κεφάλαιο διατυπώνουμε το πρόβλημα της πρόβλεψης συνδέσμου, περιγράφουμε λύσεις του και δείκτες αξιολόγησής τους και αναφέρουμε πιθανές μελλοντικές κατευθύνσεις έρευνας στην περιοχή. Ακολουθώντας όμοια τακτική, στο τέταρτο κεφάλαιο ασχολούμαστε με το πρόβλημα της ανίχνευσης κοινοτήτων. Το πέμπτο κεφάλαιο είναι αφιερωμένο στην κοινωνική σύσταση (social recommendation). Αφού επισκοπήσουμε τα παραδοσιακά συστήματα συστάσεων, δίνουμε ορισμούς του προβλήματος, ταξινομούμε λύσεις του, αναφέρουμε δείκτες αξιολόγησής τους και παρουσιάζουμε μελλοντικές κατευθύνσεις έρευνας στην περιοχή. Στο έκτο κεφάλαιο συνοψίζουμε τα συμπεράσματα της εργασίας. Τέλος, στο έβδομο κεφάλαιο παραθέτουμε τη βιβλιογραφία στην οποία στηρίζεται η διπλωματική.

2

Μοντέλα κοινωνικών δικτύων

Στο κεφάλαιο αυτό αρχικά εξετάζουμε χαρακτηριστικά που έχει παρατηρηθεί ότι παρουσιάζουν τα κοινωνικά δίκτυα, στατικά και δυναμικά, και στη συνέχεια κάνουμε μία επισκόπηση μοντέλων που έχουν προταθεί για τα κοινωνικά δίκτυα.

2.1 Χαρακτηριστικά των κοινωνικών δικτύων

2.1.1 Στατικά χαρακτηριστικά

Σε αυτήν την υποενότητα παρουσιάζουμε στατικά χαρακτηριστικά των κοινωνικών δικτύων ([New03a]).

2.1.1.1 Κατανομή βαθμών κορυφών

Έχει παρατηρηθεί ότι οι βαθμοί των κόμβων ενός κοινωνικού δικτύου ακολουθούν κατά προσέγγιση κατανομή που βασίζεται στο νόμο της δύναμης (power law), δηλαδή $P(k) \propto k^{-\gamma}$, όπου $P(k)$ το κλάσμα των κόμβων με βαθμό k και γ πραγματική σταθερά (συνήθως μεταξύ 2 και 3). Έτσι, η πλειονότητα των κόμβων έχουν μικρό βαθμό και λίγοι κόμβοι έχουν σημαντικά υψηλότερο βαθμό. Τα δίκτυα με τέτοια κατανομή βαθμών κορυφών αναφέρονται και ως δίκτυα ανεξάρτητα κλίμακας (scale-free).

2.1.1.2 Κοινοτική δομή

Τα κοινωνικά δίκτυα παρουσιάζουν κοινοτική δομή (community structure), δηλαδή είναι δομημένα από ομάδες κορυφών (κοινότητες) τέτοιες ώστε η πυκνότητα των ακμών εντός των ομάδων να είναι υψηλή ενώ μεταξύ των ομάδων χαμηλή.

2.1.1.3 Συντελεστής ομαδοποίησης

Στα κοινωνικά δίκτυα έχει βρεθεί ότι αν μία κορυφή A συνδέεται με μία κορυφή B και η κορυφή B με μία κορυφή C τότε υπάρχει αυξημένη πιθανότητα η κορυφή A να συνδέεται με την κορυφή C (μεταβατικότητα – transitivity). Η μεταβατικότητα μπορεί να ποσοτικοποιηθεί με το συντελεστή ομαδοποίησης (clustering coefficient) *C*:

$$C = \frac{3 \times \#\,\text{τριγωνων στο δικτυο}}{\#\,\text{συνδεδεμενων τριαδων κορυφων}} \text{ (2.1),}$$

όπου μία συνδεδεμένη τριάδα είναι μία κορυφή με ακμές προς ένα μη διατεταγμένο ζεύγος άλλων κορυφών. Ο συντελεστής ομαδοποίησης μπορεί να ερμηνευθεί ως η μέση πιθανότητα δύο κορυφές που είναι γειτονικές μίας τρίτης να είναι μεταξύ τους γειτονικές.

Αυτός ο ορισμός του συντελεστή ομαδοποίησης έχει χρησιμοποιηθεί ευρέως στην κοινωνιολογία, όπου αναφέρεται ως το ‘κλάσμα των μεταβατικών τριάδων’. Ένας εναλλακτικός ορισμός, ο οποίος επίσης χρησιμοποιείται ευρέως, έχει δοθεί από τους Watts και Strogatz ([WS98]), οι οποίοι πρότειναν μία τοπική τιμή:

$$C_i = \frac{\#\,\text{τριγωνων που συνδεονται με την κορυφη } i}{\#\,\text{τριαδων με κεντρο την κορυφη } i} \text{ (2.2).}$$

Για κορυφές με βαθμό 0 ή 1, θεωρούμε $C_i = 0$. Ο συντελεστής ομαδοποίησης του δικτύου ορίζεται ως ο μέσος:

$$C = \frac{1}{n}\sum_i C_i \text{ (2.3).}$$

Η τοπική τιμή του συντελεστή ομαδοποίησης αναφέρεται στην κοινωνιολογία ως η ‘πυκνότητα του δικτύου’.

Ανεξάρτητα του χρησιμοποιούμενου ορισμού, η τιμή του συντελεστή ομαδοποίησης των κοινωνικών δικτύων τείνει να είναι σημαντικά υψηλότερη από αυτή ενός τυχαίου γράφου με τον ίδιο αριθμό κορυφών και ακμών.

2.1.1.4 Φαινόμενο του μικρού κόσμου

Τα δίκτυα αυτά παρουσιάζουν σχετικά μικρά μέσα μήκη συντομότερων μονοπατιών και διαμέτρους (φαινόμενο του μικρού κόσμου – small world effect). Το φαινόμενο του μικρού κόσμου είναι γνωστό και ως ‘έξι βαθμοί διαχωρισμού’ (six degrees of separation).

Η πρώτη πειραματική μελέτη της έννοιας αυτής διεξήχθη από τον Stanley Milgram και συναδέλφους του τη δεκαετία του 1960 ([Mil67], [TM69]). Ο Milgram, ζήτητσε από 296 τυχαία επιλεγμένους ανθρώπους να προσπαθήσουν να προωθήσουν ένα γράμμα προς ένα πρόσωπο-στόχο, ένα χρηματιστή που ζούσε σε ένα προάστιο της Βοστώνης. Τους έδωσε κάποιες προσωπικές πληροφορίες για το πρόσωπο-στόχο και τους ζήτησε να προωθήσουν το γράμμα σε κάποιον που γνώριζαν, με τις ίδιες οδηγίες, με στόχο να φθάσει στο πρόσωπο-στόχο το συντομότερο δυνατό. Προκάλεσε, και προκαλεί, έκπληξη το γεγονός ότι 64 από τα γράμματα έφθασαν στον τελικό προορισμό τους και μάλιστα με διάμεση τιμή μήκους μονοπατιού 6.

Το φαινόμενο του μικρού κόσμου έχει κάποιες προφανείς συνέπειες στη δυναμική των διεργασιών που λαμβάνουν χώρα στα κοινωνικά δίκτυα. Για παράδειγμα, αν θεωρήσουμε τη διάδοση πληροφορίας σε ένα κοινωνικό δίκτυο, το φαινόμενο του μικρού κόσμου συνεπάγεται ότι η διάδοση θα είναι ταχεία.

2.1.1.5 Ανθεκτικότητα σε αφαίρεση κορυφών

Ας υποθέσουμε ότι αρχίζουμε να αφαιρούμε κορυφές από ένα συνεκτικό κοινωνικό δίκτυο. Τότε οι αποστάσεις μεταξύ των ζευγών κορυφών θα αρχίσουν να αυξάνονται, ώσπου τα ζεύγη κορυφών να αποσυνδεθούν και η επικοινωνία μεταξύ τους να καταστεί αδύνατη. Έχει παρατηρηθεί ότι τα δίκτυα είναι ανθεκτικά (resilient) στην τυχαία αφαίρεση κορυφών, δηλαδή οι αποστάσεις μεταξύ των ζευγών κορυφών αυξάνονται μόνο ανεπαίσθητα όταν αρχίσουν να αφαιρούνται κορυφές με τυχαίο τρόπο. Ωστόσο, είναι ευάλωτα σε στοχευμένη αφαίρεση κορυφών (π. χ. με φθίνουσα σειρά βαθμού).

2.1.1.6 Πρότυπα ανάμιξης

Στα κοινωνικά δίκτυα συνήθως οι κορυφές μπορούν να διακριθούν σε διάφορες κατηγορίες (π. χ. με βάση την ηλικία, τη φυλή, το επάγγελμα, το εισόδημα ή τη γεωγραφική θέση) και η πιθανότητα δύο κορυφές να συνδέονται εξαρτάται από τις κατηγορίες στις οποίες ανήκουν. Έχει παρατηρηθεί ότι η πιθανότητα αυτή είναι μεγαλύτερη για κορυφές της ίδιας κατηγορίας (επιλεκτική ανάμιξη – assortative mixing – ή ομοιοφιλία – homophily).

Η επιλεκτική ανάμιξη μπορεί να ποσοτικοποιηθεί με έναν ‘συντελεστή επιλεκτικότητας’ (assortativity coefficient).

Έστω ότι οι κορυφές ενός κοινωνικού δικτύου ανήκουν σε *N* κατηγορίες και E_{ij} ο αριθμός των ακμών που συνδέουν κορυφές των κατηγοριών *i και j* με *i,j* = 1,…,*N*. Αν E είναι ο πίνακας με στοιχεία E_{ij}, ο κανονικοποιημένος πίνακας ανάμιξης ορίζεται ως:

$$\mathrm{e} = \frac{\mathrm{E}}{\left\|\mathrm{E}\right\|} \text{ (2.4),}$$

όπου $\left\|\mathrm{x}\right\|$ είναι το άθροισμα των στοιχείων του πίνακα x. Τα στοιχεία e_{ij} εκφράζουν το κλάσμα των ακμών που συνδέουν κορυφές των κατηγοριών *i και j*. Η υπό συνθήκη πιθανότητα *P*(*j*|*i*) ο γείτονας μίας κορυφής της κατηγορίας *i* να ανήκει στην κατηγορία *j* είναι $P(j \mid i) = e_{ij} \Big/ \sum_j e_{ij}$.

Οι Gupta κ. ά. ([GAM89]) πρότειναν το συντελεστή:

$$Q = \frac{\sum_i P(i \mid i) - 1}{N - 1} \text{ (2.5).}$$

Αυτή η ποσότητα έχει τιμή 1 για ένα τελείως επιλεκτικό δίκτυο (όλες οι ακμές συνδέουν κορυφές της ίδιας κατηγορίας) και 0 για τυχαία αναμεμιγμένα δίκτυα.

Ένας εναλλακτικός συντελεστής επιλεκτικότητας προτάθηκε από τον Newman ([New03b]):

$$r = \frac{\mathrm{Tre} - \left\|\mathrm{e}^2\right\|}{1 - \left\|\mathrm{e}^2\right\|} \text{ (2.6).}$$

Αυτή η ποσότητα επίσης έχει τιμή 1 για ένα τελείως επιλεκτικό δίκτυο και 0 για τυχαία αναμεμιγμένα δίκτυα.

Η ανάμιξη με βάση βαθμωτά χαρακτηριστικά (π. χ. εισόδημα ή ηλικία) μπορεί να ποσοτικοποιηθεί υπολογίζοντας ένα συντελεστή συσχέτισης για το υπό μελέτη χαρακτηριστικό.

2.1.2 Δυναμικά χαρακτηριστικά

Σε αυτήν την υποενότητα παρουσιάζουμε δυναμικά χαρακτηριστικά των κοινωνικών δικτύων ([Agg11]).

2.1.2.1 Συρρικνούμενη διάμετρος (shrinking diameter)

Οι Leskovec κ. ά. ([LKF05]) έδειξαν ότι όχι μόνο η διάμετρος των κοινωνικών δικτύων είναι μικρή αλλά και συρρικνώνεται και στη συνέχεια σταθεροποιείται με το χρόνο. Αυτό το χαρακτηριστικό μπορεί να αποδοθεί στο ‘σημείο πήξεως’ (gelling point) και τη ‘πύκνωση’ (densification) των κοινωνικών δικτύων, που περιγράφονται παρακάτω. Εν συντομία

μπορούμε να αναφέρουμε ότι στο ‘σημείο πήξεως’ πολλές μικρές συνεκτικές συνιστώσες συγχωνεύονται και σχηματίζουν τη μεγαλύτερη συνεκτική συνιστώσα του γράφου. Στη συνέχεια, με την προσθήκη νέων ακμών η διάμετρος εξακολουθεί να συρρικνώνεται ώσπου φθάνει σε ισορροπία.

2.1.2.2 Νόμος της δύναμης για την πύκνωση

Οι χρονικά εξελισσόμενοι γράφοι ακολουθούν το νόμο της δύναμης για την πύκνωση, δηλαδή $E(t) \propto N(t)^{\beta}$ για κάθε στιγμή *t*, όπου β είναι ο εκθέτης πύκνωσης και *E*(*t*) και *N*(*t*) ο αριθμός των ακμών και των κορυφών τη στιγμή *t*, αντίστοιχα.

2.1.2.3 Σημείο πήξεως

Συχνά υπάρχει ένα χρονικό σημείο (το σημείο πήξεως) κατά το οποίο η διάμετρος ελαττώνεται απότομα. Πριν από αυτό το σημείο, ο γράφος αποτελείται από πολλές μικρές συνεκτικές συνιστώσες. Μετά το σημείο αυτό διάφορες μικρές συνεκτικές συνιστώσες σχηματίζουν μία γιγάντια συνεκτική συνιστώσα (giant connected component – GCC), ο γράφος ακολουθεί το νόμο της δύναμης για την πύκνωση, η διάμετρός του ελαττώνεται και σταθεροποιείται και η γιγάντια συνεκτική συνιστώσα συνεχίζει να μεγαλώνει, απορροφώντας τη μεγάλη πλειονότητα των νεοεισερχόμενων κόμβων.

2.1.2.4 Σταθερές/ταλαντούμενες μικρές συνεκτικές συνιστώσες

Μετά το σημείο πήξεως, τα μεγέθη της δεύτερης και της τρίτης μεγαλύτερης συνεκτικής συνιστώσας ταλαντώνονται με το χρόνο. Μία απροσδόκητη ίσως παρατήρηση είναι ότι το μέγιστο μέγεθος που μπορούν να πάρουν αυτές οι συνιστώσες είναι σταθερό.

2.1.2.5 Εξέλιξη της μεγαλύτερης ιδιοτιμής του πίνακα γειτνίασης με το χρόνο

Παρατηρήθηκε ότι η μεγαλύτερη ιδιοτιμή του πίνακα γειτνίασης ακολουθεί το νόμο της δύναμης καθώς αυξάνεται ο αριθμός των ακμών. Αυτή η παρατήρηση ισχύει ιδιαίτερα μετά το σημείο πήξεως. Δηλαδή $\lambda_1 \propto E(t)^{\alpha}, \alpha \leq 0.5$.

2.2 Μοντέλα κοινωνικών δικτύων

Η μοντελοποίηση των κοινωνικών δικτύων βοηθά στην καλύτερη κατανόησή τους και παρέχει τη δυνατότητα της προσομοίωσης και ανάλυσής τους, καθώς και της πρόβλεψης της συμπεριφοράς διεργασιών που πραγματοποιούνται σε αυτά, όπως η διάχυση ή η ανάκτηση πληροφορίας.

Έχουν προταθεί διάφορα μοντέλα κοινωνικών δικτύων.

2.2.1 Το μοντέλο Erdős-Rényi

Οι Erdős και Rényi ([ER59]) πρότειναν ένα από τα πρώτα μοντέλα δικτύων, το τυχαίο γράφημα. Το μοντέλο αυτό χαρακτηρίζεται από δύο παραμέτρους: τον αριθμό των κορυφών και την πιθανότητα σύνδεσης. Κάθε ζεύγος κορυφών συνδέεται με ίση πιθανότητα, αναξάρτητα από τα άλλα ζεύγη. Παρότι τα τυχαία γραφήματα έγιναν ευρέως αποδεκτά, καθώς οι ιδιότητές τους διευκολύνουν τη μοντελοποίηση δικτύων, δεν αντανακλούν τη δομή των πραγματικών δικτύων μεγάλης κλίμακας: οι βαθμοί των κορυφών των τυχαίων γραφημάτων ακολουθούν κατανομή Poisson (αντί κατανομή που βασίζεται στο νόμο της δύναμης) και δεν αντανακλούν το φαινόμενο ομαδοποίησης (clustering). Έτσι, το μοντέλο αυτό είναι ακατάλληλο για σύγχρονες μελέτες.

2.2.2 Το μοντέλο Watts-Strogatz

Οι Watts και Strogatz ([WS98]) έδειξαν ότι η ιδιότητα του μικρού κόσμου και ο υψηλός συντελεστής ομαδοποίησης μπορούν να συνυπάρχουν στο ίδιο σύστημα. Σχεδίασαν μία κατηγορία γράφων που προκύπτουν από παρεμβολή (interpolation) μεταξύ ενός κανονικού πλέγματος (regular lattice), με υψηλό συντελεστή ομαδοποίησης, και ενός τυχαίου γράφου, με την ιδιότητα του μικρού κόσμου. Ξεκινούν από ένα πλέγμα δακτυλίων (ring lattice) στο οποίο κάθε κορυφή έχει βαθμό k, και με πιθανότητα p κάθε ακμή ξανατοποθετείται (το ένα άκρο της αλλάζει τυχαία). Όπως προκύπτει, μικρές τιμές του p αρκούν για να μειωθούν σημαντικά τα μήκη των συντομότερων μονοπατιών μεταξύ των κορυφών. Από την άλλη πλευρά, ο συντελεστής ομαδοποίησης παραμένει υψηλός. Για $p = 1$, η προκύπτουσα δομή είναι ένας τυχαίος γράφος κατά Erdős-Rényi. Το μοντέλο Watts-Strogatz έχει μελετηθεί ευρέως. Ωστόσο, δεν μπορεί να παραστήσει την κατανομή βαθμών κορυφών που βασίζεται στο νόμο της δύναμης των πραγματικών δικτύων.

2.2.3 Το μοντέλο Barabási-Albert

Το μοντέλο Barabasi-Albert ([BA99]) είναι το πιο δημοφιλές μοντέλο με κατανομή βαθμών κορυφών που βασίζεται στο νόμο της δύναμης. Βοηθάει στην κατανόηση της προέλευσης και της εξέλιξης των ιδιοτήτων των κοινωνικών δικτύων. Ο γράφος κατασκευάζεται με μία δυναμική διαδικασία όπου οι κορυφές προστίθενται μία-μία σε έναν αρχικό πυρήνα (core). Η πιθανότητα μία νέα κορυφή να συνδεθεί με μία ήδη υπάρχουσα είναι ανάλογη προς το βαθμό της τελευταίας. Με αυτόν τον τρόπο, κορυφές με υψηλό βαθμό έχουν μεγάλη πιθανότητα να επιλεγούν ως γείτονες από νέες κορυφές. Αν συμβεί αυτό, ο βαθμός τους αυξάνεται, οπότε είναι ακόμα πιθανότερο να επιλεγούν στο μέλλον. Στο όριο καθώς ο αριθμός των κορυφών

τείνει στο άπειρο, αυτή η στρατηγική δημιουργεί ένα γράφο με κατανομή βαθμών που χαρακτηρίζεται από μία ουρά που βασίζεται στο νόμο της δύναμης με εκθέτη 3 (ενώ οι αντίστοιχες τιμές που παρατηρούνται στα πραγματικά δίκτυα συνήθως κυμαίνονται μεταξύ 2 και 3). Το μοντέλο αυτό παριστάνει την κατανομή που βασίζεται στο νόμο της δύναμης, αλλά έχει άλλες ιδιότητες που ίσως δε συμφωνούν με εμπειρικά αποτελέσματα. Παρουσιάζει πολύ μικρότερο μέσο μήκος συντομότερου μονοπατιού από ένα τυχαίο γράφημα και συντελεστή ομαδοποίησης που μειώνεται με το μέγεθος του δικτύου και είναι πολύ χαμηλότερος από ό,τι στα πραγματικά κοινωνικά δίκτυα. Αποτυγχάνει να αναπαραστήσει την κοινοτική δομή των πραγματικών κοινωνικών δικτύων.

2.2.4 Μοντέλα εξέλιξης δικτύου (network evolution models – NEMs)

Στα μοντέλα εξέλιξης δικτύου ([TKK+09]) η προσθήκη νέων συνδέσμων εξαρτάται μόνο από την τοπική δομή του δικτύου. Χρησιμοποιούνται για τον έλεγχο υποθέσεων ότι συγκεκριμένοι μηχανισμοί εξέλιξης του δικτύου οδηγούν σε συγκεκριμένη δομή του δικτύου. Υποδιαιρούνται περαιτέρω σε αυξητικά (growing) μοντέλα, στα οποία προστίθενται κόμβοι και σύνδεσμοι ώσπου το δίκτυο να περιέχει τον επιθυμητό αριθμό κόμβων, και δυναμικά (dynamical) μοντέλα, στα οποία επαναλαμβάνονται βήματα προσθήκης και αφαίρεσης δεσμών σε ένα σταθερό σύνολο κόμβων ώσπου η δομή του δικτύου να μη μεταβάλλεται άλλο στατιστικά. Ορίζονται με βάση τις ακόλουθες τρεις ιδιότητες:

1. Μία πραγμάτωση G του δικτύου παράγεται με μία επαναληπτική διαδικασία που ξεκινά πάντα από μία αρχική διάταξη $G(t_0)$ που καθορίζεται από το NEM. Τα δυναμικά μοντέλα συχνά ξεκινούν με ένα κενό δίκτυο, ενώ τα αυξητικά με ένα μικρό δίκτυο-σπόρο (seed network).
2. Οι προδιαγραφές των NEMs περιλαμβάνουν ένα ρητά ορισμένο σύνολο στοχαστικών κανόνων με τους οποίους η δομή του δικτύου εξελίσσεται στο χρόνο. Αυτοί οι κανόνες αφορούν την επιλογή ενός υποσυνόλου κόμβων και συνδέσμων σε κάθε βήμα και την προσθήκη και διαγραφή κόμβων και συνδέσμων από αυτό το υποσύνολο. Τυπικά αντιστοιχούν σε αφηρημένους μηχανισμούς σχηματισμού κοινωνικών δεσμών όπως το τριαδικό κλείσιμο (triadic closure), δηλαδή ο σχηματισμός δεσμών με βάση την τάση δύο φίλοι ενός ατόμου να γνωρίζονται μεταξύ τους. Οι κανόνες πάντοτε εξαρτώνται από τη δομή του δικτύου και κάποιες φορές μπορεί να ενσωματώνουν γνωρίσματα των κόμβων. Καθορίζουν τις δυνατές μεταβάσεις από το δίκτυο $G(t_{k-1})$ στο $G(t_k)$ κατά την επαναληπτική διαδικασία η οποία θα παραγάγει την πραγμάτωση $G = G(t_{end})$.
3. Το NEM περιλαμβάνει ένα κριτήριο τερματισμού. Για ένα αυξητικό NEM, ο αλγόριθμος τερματίζει όταν το δίκτυο φθάσει ένα προκαθορισμένο μέγεθος. Για ένα

δυναμικό NEM ο αλγόριθμος τερματίζει όταν επιλεγμένες στατιστικές του δικτύου παύσουν να μεταβάλλονται.

2.2.4.1 Δυναμικά NEMs

Το μοντέλο DEB ([DEB02]): Το μοντέλο DEB χαρακτηρίζεται από 2 παραμέτρους, που συμβολίζονται με N (αριθμός των κόμβων) και p. Ακολουθεί επαναληπτικά την παρακάτω διαδικασία:

1. Επιλέγει έναν κόμβο i τυχαία. Αν ο i έχει λιγότερους από δύο δεσμούς, τον συνδέει με έναν τυχαίο κόμβο, διαφορετικά επιλέγει δύο γείτονες του i και τους συνδέει μεταξύ τους αν δε συνδέονται ήδη.
2. Επιλέγει έναν τυχαίο κόμβο και με πιθανότητα p αφαιρεί όλους τους δεσμούς του.

Το μοντέλο MVS ([MVS04]): Το μοντέλο MVS χαρακτηρίζεται από 4 παραμέτρους, που συμβολίζονται με N (αριθμός των κόμβων), ξ, η και λ. Ακολουθεί επαναληπτικά την παρακάτω διαδικασία:

1. Επιλέγει έναν κόμβο i τυχαία. Συνδέει τον i με έναν άλλο τυχαίο κόμβο με πιθανότητα η. Επιλέγει ένα φίλο φίλου του i (με ομοιόμορφα τυχαία αναζήτηση) με πιθανότητα ξ και συνδέει τον i με αυτόν αν δε συνδέονται ήδη.
2. Επιλέγει έναν τυχαίο δεσμό και τον διαγράφει με πιθανότητα λ.

Το μοντέλο KOSKK ([KOS+07]): Το μοντέλο KOSKK χαρακτηρίζεται από 6 παραμέτρους, που συμβολίζονται με N (αριθμός των κόμβων), p_Δ, p_r, w_0, p_d και δ. Ακολουθεί επαναληπτικά την παρακάτω διαδικασία:

1. Επιλέγει έναν κόμβο i τυχαία. Επιλέγει ένα φίλο φίλου του i, k, (με σταθμισμένη αναζήτηση) και τον συνδέει με τον i με πιθανότητα p_Δ (με αρχική δύναμη δεσμού w_0) αν δε συνδέονται ήδη. Αυξάνει τις δυνάμεις των δεσμών κατά δ κατά μήκος του μονοπατιού αναζήτησης, καθώς επίσης και στο σύνδεσμο l_{ik} αν υπάρχει ήδη. Επίσης, με πιθανότητα p_r (ή με πιθανότητα 1 αν ο i δεν έχει γείτονες), συνδέει τον i με έναν τυχαίο κόμβο j (με δύναμη δεσμού w_0).
2. Επιλέγει έναν τυχαίο κόμβο και με πιθανότητα p_d αφαιρεί όλους τους δεσμούς του.

2.2.4.2 Αυξητικά (growing) NEMs

Το μοντέλο TOSHK ([TOS+06]): Το μοντέλο TOSHK χαρακτηρίζεται από 3 παραμέτρους, που συμβολίζονται με N (επιθυμητός αριθμός κόμβων), p και k. Ακολουθεί επαναληπτικά την παρακάτω διαδικασία (απλοποιημένη εκδοχή):

1. Προσθέτει έναν νέο κόμβο i στο δίκτυο, συνδέοντάς τον με μία τυχαία αρχική επαφή με πιθανότητα p, ή με δύο με πιθανότητα $1 - p$.

2. Για κάθε τυχαία αρχική επαφή j, επιλέγει έναν αριθμό m_{sec} δευτερευουσών συνδέσεων από την κατανομή U[0, k] και συνδέει τον i με m_{sec} γείτονες του j αν υπάρχουν τόσοι.

Το μοντέλο του Vázquez ([Váz03]): Το μοντέλο του Vázquez χαρακτηρίζεται από 2 παραμέτρους, που συμβολίζονται με N (επιθυμητός αριθμός κόμβων), u. Ακολουθεί επαναληπτικά την παρακάτω διαδικασία:

1. Με πιθανότητα 1 – u, προσθέτει έναν νέο κόμβο n στο δίκτυο, συνδέοντάς τον με έναν τυχαίο κόμβο i. Δημιουργεί 'δυνητικές ακμές' μεταξύ του n και των γειτόνων j του i.
2. Με πιθανότητα u, μετατρέπει μία 'δυνητική ακμή' που δημιουργήθηκε σε οποιοδήποτε προηγούμενο βήμα σε ακμή.

2.2.5 Μοντέλα γνωρισμάτων κόμβων (nodal attribute models – NAMs)

Στα μοντέλα γνωρισμάτων κόμβων ([TKK+09]) η πιθανότητα ύπαρξης της ακμής e_{ij} μεταξύ των κόμβων i και j εξαρτάται μόνο από τα γνωρίσματα των κόμβων i και j. Συχνά βασίζονται στην έννοια της ομοιοφιλίας (homophily), δηλαδή την τάση οι όμοιοι να αλληλεπιδρούν μεταξύ τους, η οποία είναι γνωστό ότι δομεί δεσμούς διαφόρων τύπων, όπως φιλίας, εργασίας, γάμου και μεταφοράς πληροφορίας. Περιγράφονται και με τον όρο 'χωρικά μοντέλα', που αναφέρεται στο γεγονός ότι τα γνωρίσματα κάθε κόμβου καθορίζουν τη θέση του σε έναν κοινωνικό ή γεωγραφικό χώρο.

2.2.5.1 Το μοντέλο BPDA

Το μοντέλο BPDA ([BPDA04]) χαρακτηρίζεται από 3 παραμέτρους, που συμβολίζονται με N, a, και b. Κατανέμει N κόμβους με ομοιόμορφη πιθανότητα σε έναν (μονοδιάστατο) κοινωνικό χώρο (ένα τμήμα μήκους h_{max}). Συνδέει κόμβους με πιθανότητα $p = 1/(1 + (d/b)^{\alpha})$, όπου d είναι η απόστασή τους στον κοινωνικό χώρο. (Το h_{max} μπορεί να απορροφηθεί από το b) Σε περισσότερες διαστάσεις, η ομοιότητα σε μία από αυτές αρκεί για να θεωρηθούν οι κόμβοι παρόμοιοι.

2.2.5.2 Το μοντέλο WPR

Το μοντέλο WPR ([WPR06]) χαρακτηρίζεται από 4 παραμέτρους, που συμβολίζονται με N, H, p, και p_b. Κατανέμει N κόμβους σύμφωνα με μία ομοιογενή σημειακή διαδικασία Poisson (Poisson point process) σε έναν (διδιάστατο) κοινωνικό χώρο μοναδιαίου εμβαδού. Δημιουργεί ένα σύνδεσμο μεταξύ δύο κόμβων με απόσταση d με πιθανότητα $p + p_b$ αν $d < H$, και με πιθανότητα $p - p_{\Delta}$ αν $d > H$.

2.2.6 Εκθετικά μοντέλα τυχαίων γράφων (exponential random graph models – ERGMs))

Μία κατηγορία μοντέλων που έχει χρησιμοποιηθεί ευρέως στην κοινωνιολογία είναι τα εκθετικά μοντέλα τυχαίων γράφων ([Sch12]), επίσης γνωστά ως p^* μοντέλα. Αυτά τα μοντέλα είναι κατανομές επί του χώρου όλων των δυνατών δικτύων. Παρακάτω Y είναι τυχαία μεταβλητή, ένας $N \times N$ πίνακας, ο οποίος παριστάνει το δίκτυο. Η γενική μορφή των μοντέλων είναι:

$$P_\theta\{Y = y\} = \exp(\theta_1 z_1(y) + \theta_2 z_2(y) + \ldots + \theta_k z_k(y) - \psi(\theta)) \quad (2.7),$$

όπου η παράμετρος είναι $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, οι επαρκείς στατιστικές είναι $(z_1(y), z_2(y), \ldots, z_k(y))$ και ψ μία κανονικοποιητική σταθερά. Οι στατιστικές z_i μπορεί να είναι, π. χ., ο αριθμός των τριγώνων, ο αριθμός των ακμών ή ο αριθμός των k-αστέρων. Θετική τιμή της παραμέτρου θ_i δείχνει τάση προς μεγάλη τιμή της αντίστοιχης στατιστικής. Οι στατιστικές αυτές αντιστοιχούν σε απόψεις του δικτύου, όπως η αμοιβαιότητα (reciprocity) και η μεταβατικότητα (transitivity).

Ιδανικά, αν και σε μερικές περιπτώσεις είναι αδύνατο, θα μπορούσε κανείς να παρατηρήσει ένα δείγμα διάφορων δικτύων, $Y_1, \ldots, Y_n$, όλα $N \times N$ πίνακες, που μοντελοποιούνται ως ανεξάρτητες και ισόνομες παρατηρήσεις από το ίδιο πιθανοτικό μοντέλο, και να συμπεράνει τις τιμές των παραμέτρων του μοντέλου.

Στην πράξη, στη βιβλιογραφία των ERGMs, μόνο ένα δίκτυο παρατηρείται, το οποίο είναι ένα δείγμα μεγέθους 1. Από αυτό γίνονται εκτιμήσεις των παραμέτρων του πιθανοτικού μοντέλου που παρήγαγε το δίκτυο.

Η εκτίμηση των παραμέτρων από μία παρατήρηση του δικτύου για τα ERGMs στην πράξη γίνεται συνήθως είτε με μία διαδικασία εκτίμησης ψευδοπιθανοφάνειας (pseudo-likelihood) είτε με μεθόδους MCMC (Markov chain Monte Carlo). Η πρώτη επιλογή έχει απορριφθεί γιατί παράγει άπειρες τιμές ακόμη και σε περιπτώσεις όπου η ψευδοπιθανοφάνεια συγκλίνει. Οι μέθοδοι MCMC επίσης έχουν δεχθεί κριτική και αμφισβήτηση λόγω συμπερασματικού εκφυλισμού (inferential degeneracy), όπου οι αλγόριθμοι συγκλίνουν σε εκφυλισμένους γράφους, πλήρεις ή κενούς γράφους, ή ο αλγόριθμος δε συγκλίνει συνεπώς (consistently). Είναι πιθανό ο εκφυλισμός να είναι συνάρτηση της μη ύπαρξης αντίστοιχου παραγωγικού μοντέλου (generative model) για τα ERGMs (ένα παραγωγικό μοντέλο μπορεί να περιγραφεί ευρετικά ως μία αφήγηση του τρόπου με τον οποίο το δίκτυο παρήχθη ή δημιουργήθηκε). Επιπλέον, τονίζεται ότι ο συμπερασμός από ένα δείγμα μεγέθους 1 ίσως είναι προβληματικός.

2.2.7 Μοντέλα λανθάνοντος χώρου

Παρακινούμενοι από προβλήματα εκφυλισμού και αστάθειας (instability) που παρουσιάζουν τα εκθετικά μοντέλα τυχαίων γράφων και περιγράφοντας τα προβλήματα αυτά ως ελαττώματα των ίδιων των μοντέλων που δεν μπορούν να αντιμετωπισθούν με εναλλακτικές διαδικασίες εκτίμησης των παραμέτρων, οι Hoff, Raftery και Handcock εισήγαγαν τα μοντέλα λανθάνοντος χώρου (latent space models) ([HRH02], [Sch12]). Σε αυτά τα μοντέλα, τα παρατηρούμενα δεδομένα είναι ένα $n \times n$ κοινωνικό δίκτυο Y με στοιχεία y_{ij} που δηλώνουν τις σχέσεις μεταξύ των κόμβων i και j και η αντίστοιχη συμμεταβλητή (covariate) πληροφορία X. Μη παρατηρούμενες είναι οι λανθάνουσες θέσεις z_i στον κοινωνικό χώρο, για όλους τους κόμβους $i = 1,\ldots,n$, τέτοιες ώστε οι ακμές υπό τη συνθήκη αυτών των μη παρατηρούμενων θέσεων να είναι ανεξάρτητες. Τα z_i και θ αντιμετωπίζονται ως παράμετροι προς εκτίμηση στο μοντέλο $P(Y \mid z, x, \theta) = \prod_{i \neq j} P(y_{ij} \mid z_i, z_j, X_{ij}, \theta)$, όπου οι όροι στο γινόμενο μπορούν να μοντελοποιηθούν με χρήση λογιστικής παλινδρόμησης. Με μεθόδους MCMC μπορούν να βρεθούν εκτιμήσεις μέγιστης πιθανοφάνειας για τα Z και θ.

3

Πρόβλεψη συνδέσμου

3.1 Εισαγωγή

Η πρόβλεψη συνδέσμου (link prediction) ([Agg11], [LZ11])είναι μια σημαντική εφαρμογή ανάλυσης κοινωνικών δικτύων η οποία έχει εφαρμογές και σε άλλους τομείς, όπως η ανάκτηση πληροφορίας (information retrieval), η βιοπληροφορική (bioinformatics) και το ηλεκτρονικό εμπόριο (e-commerce). Πιο συγκεκριμένα, μπορεί να χρησιμοποιηθεί στην αυτόματη δημιουργία υπερσυνδέσμων και την πρόβλεψη υπερσυνδέσμων. Επίσης, στο ηλεκτρονικό εμπόριο μία από τις πιο χαρακτηριστικές της χρήσεις είναι στο κτίσιμο συστημάτων συστάσεων. Στη βιοπληροφορική έχει χρησιμοποιηθεί στην αλληλεπίδραση πρωτεϊνών (protein-protein interaction – PPI) και στο σχολιασμό (annotation) του γράφου αλληλεπίδρασης πρωτεϊνών. Στη βιβλιογραφία και τη βιβλιοθηκονομία μπορεί να χρησιμοποιηθεί για απαλοιφή διπλοτύπων (deduplication) και διασύνδεση εγγραφών (record linkage). Σε εφαρμογές ασφαλείας, μπορεί να χρησιμοποιηθεί για την αναγνώριση κρυμμένων ομάδων τρομοκρατών και εγκληματιών.

Η πρόβλεψη συνδέσμου είναι το πρόβλημα της πρόβλεψης της πιθανότητας δύο κόμβοι οι οποίοι δεν συσχετίζονται στο παρόν να συσχετιστούν στο μέλλον. Πιο τυπικά το πρόβλημα της πρόβλεψης συνδέσμου μπορεί να διατυπωθεί ως εξής ([LK07]): Δίνεται ένα κοινωνικό δίκτυο $G(V, E)$ στο οποίο μία ακμή $e = (u, v) \in E$ παριστάνει κάποια μορφή αλληλεπίδρασης μεταξύ των άκρων της κάποια συγκεκριμένη στιγμή $t(e)$. Μπορούμε να

καταγράψουμε πολλαπλές αλληλεπιδράσεις με παράλληλες ακμές ή χρησιμοποιώντας μία σύνθετη χρονοσφραγίδα (complex timestamp) για κάθε ακμή. Για τις χρονικές στιγμές t και t' με $t \le t'$, συμβολίζουμε με $G[t, t']$ τον υπογράφο του G που περιορίζεται στις ακμές με χρονοσφραγίδες μεταξύ των t και t'. Σε μια διατύπωση της πρόβλεψης συνδέσμου για επιβλεπόμενη εκπαίδευση, μπορούμε να επιλέξουμε ένα διάστημα εκπαίδευσης $[t_0, t_0']$ και ένα διάστημα δοκιμής $[t_1, t_1']$, όπου $t_0' < t_1$. Η εφαρμογή της πρόβλεψης συνδέσμου έγκειται στην εξαγωγή της λίστας των ακμών που απουσιάζουν στο γράφο $G[t_0, t_0']$ αλλά προβλέπεται να εμφανισθούν στο γράφο $G[t_1, t_1']$.

Σε αυτό το κεφάλαιο παρουσιάζουμε μία ταξινόμηση σε διάφορες ομάδες των υφιστάμενων προσεγγίσεων στο πρόβλημα της πρόβλεψης συνδέσμου σε κοινωνικά δίκτυα. Μία ομάδα αποτελείται από αλγορίθμους που υπολογίζουν ένα βαθμό (score) ομοιότητας μεταξύ δύο κόμβων, οπότε μπορεί να χρησιμοποιηθεί μία μέθοδος επιβλεπόμενης μάθησης. Σε αυτήν την ομάδα περιλαμβάνουμε επίσης μεθόδους που χρησιμοποιούν έναν πίνακα πυρήνα (kernel matrix) και στη συνέχεια έναν ταξινομητή μέγιστου περιθωρίου (maximum margin classifier). Μία άλλη ομάδα αλγορίθμων βασίζεται σε μπεϋζιανά και σχεσιακά πιθανοτικά μοντέλα. Πέραν αυτών, υπάρχουν αλγόριθμοι που βασίζονται σε μοντέλα εξέλιξης γράφων ή στη γραμμική άλγεβρα. Διάφορες μέθοδοι εμπίπτουν σε πολλές ομάδες του παραπάνω σχήματος ταξινόμησης.

3.2 Πρόβλεψη συνδέσμου βασισμένη σε χαρακτηριστικά

Μπορούμε να μοντελοποιήσουμε το πρόβλημα της πρόβλεψης συνδέσμου ως μία εφαρμογή επιβλεπόμενης ταξινόμησης όπου κάθε σημείο δεδομένων (data point) αντιστοιχεί σε ένα ζεύγος κορυφών του γράφου του κοινωνικού δικτύου. Για να εκπαιδεύσουμε το μοντέλο μάθησης μπορούμε να χρησιμοποιήσουμε τις πληροφορίες σχετικά με τους συνδέσμους από το διάστημα εκπαίδευσης $[t_0, t_0']$. Με το μοντέλο αυτό μπορούμε να κάνουμε προβλέψεις μελλοντικών συνδέσμων στο διάστημα δοκιμής $[t_1, t_1']$. Πιο τυπικά, έστω $u, v \in V$ δύο κορυφές του γράφου $G(V, E)$ και $y^{\langle u, v \rangle}$ η ετικέτα (label) του σημείου δεδομένων $\langle u, v \rangle$. Σημειώστε ότι υποθέτουμε ότι οι αλληλεπιδράσεις μεταξύ των u και v είναι συμμετρικές, οπότε τα ζεύγη $\langle u, v \rangle$ και $\langle v, u \rangle$ παριστάνουν το ίδιο σημείο δεδομένων, επομένως $y^{\langle u, v \rangle} = y^{\langle v, u \rangle}$. Θεωρούμε:

$$y^{\langle u, v \rangle} = \begin{cases} +1, \alpha \nu \langle u, v \rangle \in E \\ -1, \alpha \nu \langle u, v \rangle \notin E \end{cases} \quad (3.1).$$

Χρησιμοποιώντας την παραπάνω επισήμανση (labeling) για ένα σύνολο σημείων δεδομένων εκπαίδευσης, κτίζουμε ένα μοντέλο ταξινόμησης που μπορεί να προβλέψει την άγνωστη ετικέτα ενός ζεύγους κορυφών $\langle u, v \rangle$ όπου $\langle u, v \rangle \notin E$ στο γράφο $G[t_1, t_1']$.

Αυτή είναι μία χαρακτηριστική εφαρμογή δυαδικής (binary) ταξινόμησης για την οποία μπορεί να χρησιμοποιηθεί οποιοδήποτε από τα δημοφιλή εργαλεία επιβλεπόμενης ταξινόμησης, όπως ο απλός ταξινομητής Bayes, τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων υποστήριξης και ο αλγόριθμος των k-κοντινότερων γειτόνων. Η μεγαλύτερη πρόκληση σε αυτήν την προσέγγιση είναι η επιλογή ενός συνόλου χαρακτηριστικών για την εφαρμογή της ταξινόμησης. Στη συνέχεια παρουσιάζουμε χαρακτηριστικά που έχουν χρησιμοποιηθεί με επιτυχία σε εφαρμογές επιβλεπόμενης πρόβλεψης συνδέσμου.

3.2.1 Κατασκευή του συνόλου χαρακτηριστικών

Η επιλογή ενός κατάλληλου συνόλου χαρακτηριστικών είναι το πιο κρίσιμο μέρος κάθε αλγορίθμου μηχανικής μάθησης. Για την πρόβλεψη συνδέσμου, κάθε σημείο δεδομένων αντιστοιχεί σε ένα ζεύγος κορυφών, με την ετικέτα του να δηλώνει την κατάσταση συνδέσμου (link status), οπότε τα επιλεγμένα χαρακτηριστικά θα πρέπει να παριστάνουν κάποιας μορφής εγγύτητα μεταξύ των κορυφών του ζεύγους. Στις υπάρχουσες ερευνητικές εργασίες στην πρόβλεψη συνδέσμου, η πλειονότητα των χαρακτηριστικών εξάγονται από την τοπολογία του γράφου. Επίσης, κάποιες εργασίες κατασκευάζουν ένα σύνολο χαρακτηριστικών με βάση ένα μοντέλο εξέλιξης γράφων. Επιπρόσθετα, τα γνωρίσματα των κορυφών και των ακμών μπορεί να είναι πολύ καλά χαρακτηριστικά για πολλούς τομείς εφαρμογών.

Τα χαρακτηριστικά που βασίζονται στην τοπολογία του γράφου είναι τα πιο φυσικά στην πρόβλεψη συνδέσμου. Πολλές εργασίες ([LK07], [KA06]) επικεντρώθηκαν μόνο σε τέτοια χαρακτηριστικά. Κατά χαρακτηριστικό τρόπο, υπολογίζουν ένα μέτρο ομοιότητας με βάση τις γειτονιές των κόμβων ή τα σύνολα των μονοπατιών μεταξύ των κόμβων ενός ζεύγους. Το πλεονέκτημα αυτών των χαρακτηριστικών είναι ότι είναι γενικά (generic) και μπορούν να εφαρμοστούν σε γράφους οποιουδήποτε τομέα. Έτσι, δεν είναι αναγκαία γνώση του πεδίου (domain knowledge) για τον υπολογισμό των τιμών αυτών των χαρακτηριστικών από το κοινωνικό δίκτυο. Ωστόσο, για μεγάλα κοινωνικά δίκτυα, κάποια από αυτά τα χαρακτηριστικά μπορεί να είναι υπολογιστικά ακριβά. Παρακάτω, παρουσιάζουμε χαρακτηριστικά που βασίζονται στην τοπολογία του γράφου και τα ταξινομούμε σε δύο κατηγορίες: αυτά που βασίζονται στις γειτονιές των κόμβων και αυτά που βασίζονται σε μονοπάτια. Ακολούθως παρουσιάζουμε χαρακτηριστικά που εξάγονται από τις ιδιότητες των κορυφών και των ακμών του γράφου.

3.2.1.1 Χαρακτηριστικά που βασίζονται στις γειτονιές των κόμβων

Κοινοί γείτονες. Για δύο κόμβους x και y η τιμή του χαρακτηριστικού αυτού ορίζεται ως $|\Gamma(x) \cap \Gamma(y)|$. Η ιδέα της χρήσης του χαρακτηριστικού αυτού βασίζεται στη μεταβατική ιδιότητα των δικτύων. Με απλά λόγια, αυτό σημαίνει ότι στα κοινωνικά δίκτυα αν η κορυφή x συνδέεται με την κορυφή z και η κορυφή y συνδέεται με την κορυφή z, τότε υπάρχει αυξημένη πιθανότητα η κορυφή x να συνδεθεί επίσης με την κορυφή y. Έτσι, όσο αυξάνεται ο αριθμός των κοινών γειτόνων, η πιθανότητα οι x και y να συνδεθούν αυξάνεται.

Συντελεστής Jaccard. Το μέτρο των κοινών γειτόνων δεν είναι κανονικοποιημένο, οπότε μπορούμε να χρησιμοποιήσουμε το συντελεστή Jaccard, ο οποίος κανονικοποιεί το μέγεθος του συνόλου των κοινών γειτόνων ως εξής:

$$\textbf{Jaccard-coefficient}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (3.2).$$

Ορίζει την πιθανότητα να επιλεγεί ένας κοινός γείτονας ενός ζεύγους κορυφών x και y αν η επιλογή γίνει τυχαία από την ένωση των συνόλων των γειτόνων των x και y. Έτσι, για υψηλό αριθμό κοινών γειτόνων, η βαθμολογία (score) θα είναι υψηλότερη.

Δείκτης Adamic/Adar. Οι Adamic και Adar ([AA03]) πρότειναν αυτόν το δείκτη σαν ένα μέτρο ομοιότητας μεταξύ δύο ιστοσελίδων. Για ένα σύνολο χαρακτηριστικών z ορίζεται ως εξής:

$$\sum_{z:\,\text{χαρακτηριστικό κοινό στους } x,y} \frac{1}{\log(\mathit{frequency}(z))} \quad (3.3).$$

Για την πρόβλεψη συνδέσμου, οι ([LK07]) προσάρμοσαν το μέτρο όπως παρακάτω, όπου οι κοινοί γείτονες θεωρούνται ως χαρακτηριστικά:

$$\textbf{adamic/adar}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (3.4).$$

Με αυτόν τον τρόπο, το μέτρο αυτό σταθμίζει τους κοινούς γείτονες με μικρότερο βαθμό με μεγαλύτερο βάρος. Τα αποτελέσματα των εργασιών στην πρόβλεψη συνδέσμου δείχνουν ότι ο δείκτης Adamic/Adar δίνει καλύτερα αποτελέσματα από τα προηγούμενα δύο μέτρα.

3.2.1.2 Χαρακτηριστικά που βασίζονται σε μονοπάτια

Απόσταση συντομότερου μονοπατιού. Το γεγονός ότι ο φίλος ενός φίλου μπορεί να γίνει φίλος υποβάλλει την ιδέα ότι η απόσταση μεταξύ δύο κόμβων σε ένα κοινωνικό δίκτυο μπορεί να επηρεάσει το σχηματισμό ενός συνδέσμου μεταξύ τους. Όσο μικρότερη είναι η

απόσταση τόσο μεγαλύτερη είναι η πιθανότητα να συμβεί αυτό. Αλλά, ας σημειωθεί ότι, λόγω του φαινομένου του μικρού κόσμου ([WS98]), κατά κανόνα κάθε ζεύγος κόμβων διαχωρίζεται από μικρό αριθμό κορυφών. Έτσι, αυτό το χαρακτηριστικό κάποιες φορές δε δίνει τόσο καλά αποτελέσματα. ([LK07], [HCSZ06])

Δείκτης Katz. Ο Leo Katz πρότεινε αυτό το μέτρο στο [Kat53]. Είναι μία παραλλαγή της απόστασης, αλλά γενικά δίνει καλύτερα αποτελέσματα για την πρόβλεψη συνδέσμου. Είναι ένα άθροισμα επί όλων των μονοπατιών μεταξύ ενός ζεύγους κορυφών x και y. Αλλά, για να περιορίσει τη συνεισφορά των μακρύτερων μονοπατιών στον υπολογισμό της ομοιότητας μειώνει εκθετικά τη συνεισφορά ενός μονοπατιού κατά παράγοντα β^l, όπου l είναι το μήκος του μονοπατιού. Η ακριβής εξίσωση με την οποία υπολογίζεται η τιμή του δείκτη Katz είναι:

$$\mathbf{katz}(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |\text{paths}_{x,y}^{\langle l \rangle}| \quad (3.5),$$

όπου $\text{paths}_{x,y}^{\langle l \rangle}$ είναι το σύνολο ολων των μονοπατιών μήκους l από το x στον y. Ο δείκτης Katz γενικά αποδίδει πολύ καλύτερα από την απόσταση καθώς βασίζεται στο σύνολο όλων των μονοπατιών μεταξύ των κόμβων x και y. Η παράμετρος β (≤ 1) μπορεί να χρησιμοποιηθεί για να ομαλοποιήσει (regularize) αυτό το χρακτηριστικό. Μικρή τιμή του λαμβάνει υπόψη μόνο τα συντομότερα μονοπάτια, οπότε το χαρακτηριστικό αυτό συμπεριφέρεται περίπου όπως αυτά που βασίζονται στις γειτονιές των κόμβων. Ένα πρόβλημα που παρουσιάζει είναι ότι είναι υπολογιστικά ακριβό. Μπορεί να αποδειχθεί ότι η τιμή του δείκτη Katz σε όλα τα ζεύγη κορυφών μπορεί να βρεθεί υπολογίζοντας τον πίνακα $(I - \beta A)^{-1} - I$, όπου A είναι ο πίνακας γειτνίασης και I είναι ένας μοναδιαίος πίνακας κατάλληλου μεγέθους. Ας σημειωθεί ότι η παράμετρος β πρέπει να είναι μικρότερη από τον αντίστροφο της μεγαλύτερης ιδιοτιμής του πίνακα A, ώστε να εξασφαλισθεί η σύγκλιση της σειράς.

Χρόνος μετάβασης (hitting time). Η έννοια του χρόνου μετάβασης προέρχεται από τους τυχαίους περιπάτους σε γράφους. Για δύο κορυφές x και y ενός γράφου, ο χρόνος μετάβασης $H_{x,y}$ ορίζεται ως ο αναμενόμενος αριθμός βημάτων που απαιτούνται για να φθάσει στην κορυφή y ένας τυχαίος περίπατος που ξεκινά από την κορυφή x. Μικρότερος χρόνος μετάβασης δηλώνει ότι οι κόμβοι είναι όμοιοι μεταξύ τους, οπότε έχουν μεγαλύτερη πιθανότητα να συνδεθούν στο μέλλον. Καθώς το μέτρο αυτό δεν είναι συμμετρικό, για μη κατευθυνόμενους γράφους μπορεί να χρησιμοποιηθεί ο χρόνος επανόδου (commute time) $C_{x,y} = H_{x,y} + H_{y,x}$. Το πλεονέκτημα αυτού του μέτρου είναι ότι είναι εύκολο να υπολογισθεί

εκτελώντας κάποιους δοκιμαστικούς τυχαίους περιπάτους. Ωστόσο, η τιμή του μπορεί να παρουσιάζει υψηλή διασπορά και ως εκ τούτου η πρόβλεψη να μην είναι ακριβής ([LK07]). Για παράδειγμα, ο χρόνος μετάβασης μεταξύ των κορυφών x και y μπορεί να επηρεάζεται από μία κορυφή z, η οποία βρίσκεται μακριά από τις x και y. Αν η z έχει υψηλή στατική πιθανότητα, τότε θα είναι δύσκολο ένας τυχαίος περίπατος να ξεφύγει από τη γειτονιά της z. Για να αποφευχθεί αυτό το πρόβλημα μπορούμε να χρησιμοποιήσουμε τυχαίους περιπάτους με επανεκκίνηση, όπου περιοδικά επανεκκινούμε τον τυχαίο περίπατο επιστρέφοντας στη x με σταθερή πιθανότητα a σε κάθε βήμα. Λόγω της ανεξάρτητης κλίμακας (scale free) φύσης των κοινωνικών δικτύων, κάποιες κορυφές μπορεί να παρουσιάζουν πολύ υψηλή στατική πιθανότητα (π) σε έναν τυχαίο περίπατο. Για να αντιμετωπισθεί αυτό, ο χρόνος μετάβασης μπορεί να κανονικοποιηθεί πολλαπλασιαζόμενος με τη στατική πιθανότητα του αντίστοιχου κόμβου, όπως φαίνεται παρακάτω:

$$\textbf{normalized-hitting-time}(x, y) = H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x \quad (3.6).$$

PageRank με ρίζα (rooted PageRank). Οι Chung και Zhao ([CZ10]) έδειξαν ότι το μέτρο PageRank ([BP98]), που χρησιμοποιείται για κατάταξη (ranking) ιστοσελίδων, σχετίζεται εγγενώς με το χρόνο μετάβασης. Έτσι, η τιμή PageRank μπορεί να χρησιμοποιηθεί σαν χαρακτηριστικό για την πρόβλεψη συνδέσμου. Ωστόσο, το μέτρο PageRank είναι γνώρισμα μίας κορυφής και επομένως πρέπει να τροποποιηθεί έτσι ώστε να παριστάνει το βαθμό ομοιότητας δύο κορυφών. Σύμφωνα με τον αρχικό ορισμό, το μέτρο PageRank δείχνει τη σημαντικότητα μίας κορυφής κάτω από δύο υποθέσεις: για κάποια σταθερή πιθανότητα a, ένας επισκέπτης μίας ιστοσελίδας κάνει άλμα σε μία τυχαία ιστοσελίδα με πιθανότητα a ή ακολουθεί έναν υπερσύνδεσμο με πιθανότητα $1-a$. Σε αυτόν τον τυχαίο περίπατο, η σημαντικότητα μίας ιστοσελίδας v είναι το αναμενόμενο άθροισμα των σημαντικοτήτων όλων των ιστοσελίδων u που κατευθύνουν στη v. Στην ορολογία των τυχαίων περιπάτων, μπορεί κανείς να αντικαταστήσει τον όρο *σημαντικότητα* με τον όρο *στατική κατανομή*. Για την πρόβλεψη συνδέσμου, η υπόθεση του αρχικού PageRank σχετικά με τον τυχαίο περίπατο μπορεί να τροποποιηθεί ως εξής: ο βαθμός ομοιότητας μεταξύ δύο κορυφών x και y μπορεί να μετρηθεί ως η στατική πιθανότητα της y σε έναν τυχαίο περίπατο που επιστρέφει στη x με πιθανότητα $1-\beta$ σε κάθε βήμα ή μετακινείται σε έναν τυχαίο γείτονα με πιθανότητα β. Αυτό το μέτρο είναι μη συμμετρικό και μπορεί αντ' αυτού για δύο κορυφές να χρησιμοποιηθεί το άθροισμα των μέτρων αυτών προς τις δύο κατευθύνσεις, που είναι συμμετρικό. Αυτό το μέτρο ονομάστηκε PageRank με ρίζα (rooted PageRank). Το PageRank με ρίζα μεταξύ όλων των ζευγών κόμβων (RPR) μπορεί να προκύψει όπως περιγράφεται

παρακάτω. Έστω D ο διαγώνιος πίνακας βαθμών ($D[i, i] = \sum_j A[i, j]$). Έστω $N = D^{-1}A$ ο πίνακας γειτνίασης με τα αθροίσματα των γραμμών κανονικοποιημένα στο 1. Τότε:

$$RPR = (1-\beta)(I-\beta N)^{-1} \text{ (3.7).}$$

3.2.1.3 Χαρακτηριστικά που βασίζονται σε γνωρίσματα των κορυφών και των ακμών

Τα γνωρίσματα (attributes) των κορυφών και των ακμών παίζουν σημαντικό ρόλο στην πρόβλεψη συνδέσμου. Πολλές μελέτες ([HCSZ06], [DYTG09]) έδειξαν ότι τα γνωρίσματα των κορυφών και των ακμών ως χαρακτηριστικά εγγύτητας μπορούν να βελτιώσουν σημαντικά την επίδοση των εφαρμογών πρόβλεψης συνδέσμου. Το πλεονέκτημα ενός τέτοιου συνόλου χαρακτηριστικών είναι ότι γενικά είναι υπολογιστικά φθηνό. Όμως, τα χαρακτηριστικά αυτά είναι στενά συνδεδεμένα με το πεδίο γνώσης (domain), οπότε απαιτείται καλή γνώση του πεδίου για να προσδιορισθούν. Παρακάτω, περιγράφουμε ένα γενικό τρόπο ενσωμάτωσης των χαρακτηριστικών αυτών σε μία εφαρμογή πρόβλεψης συνδέσμου.

Συνάθροιση (aggregation) χαρακτηριστικών κορυφών. Αφού προσδιορίσουμε ένα γνώρισμα a των κόμβων ενός κοινωνικού δικτύου, πρέπει να επινοήσουμε μία συναθροιστική συνάρτηση (aggregation function) f που να έχει νόημα. Για να υπολογίσει το βαθμό ομοιότητας ανάμεσα στις κορυφές x και y, η f δέχεται τις τιμές των γνωρισμάτων των κορυφών αυτών και παράγει ένα βαθμό ομοιότητας. Η επιλογή της συνάρτησης εξαρτάται εξ ολοκλήρου από το είδος του γνωρίσματος. Παρακάτω δίνουμε δύο παραδείγματα όπου γίνεται συνάθροιση κάποιου μέτρου που χαρακτηρίζει τις κορυφές.

Βαθμός προνομιακής προσάρτησης (Preferential Attachment Score): Με λίγα λόγια, σύμφωνα με την ιδέα της προνομιακής προσάρτησης ([BJNR02]), μία κορυφή συνδέεται με άλλες κορυφές στο δίκτυο με πιθανότητα ανάλογη με το βαθμό τους. Έτσι, αν θεωρήσουμε το μέγεθος της γειτονιάς (βαθμό) σαν χαρακτηριστικό, η πράξη του πολλαπλασιασμού θα μπορούσε να είναι μία συναθροιστική συνάρτηση. Η τιμή της ονομάζεται βαθμός προνομιακής προσάρτησης:

$$\textbf{preferential-attachment-score}(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| \text{ (3.8).}$$

Στην πραγματικότητα, θα μπορούσε να χρησιμοποιηθεί επίσης αντί του πολλαπλασιασμού η πρόσθεση, η οποία έχει χρησιμοποιηθεί δίνοντας πολύ καλά αποτελέσματα ([HCSZ06]).

Συντελεστής ομαδοποίησης (Clustering Coefficient): Ο συντελεστής ομαδοποίησης μίας κορυφής v ορίζεται ως εξής:

$$\textbf{clustering-coefficient}(v) = \frac{3 \times \#\,\tau\rho\iota\gamma\acute{\omega}\nu\omega\nu\ \gamma\epsilon\iota\tau o\nu\iota\kappa\acute{\omega}\nu\ \sigma\tau\eta\ v}{\#\,\text{τριαδων γειτονικων στη}\ v} \text{ (3.9).}$$

Και σε αυτήν την περίπτωση, μπορεί να χρησιμοποιηθεί ο πολλαπλασιασμός ή η πρόσθεση.

3.3 Μπεϋζιανά πιθανοτικά μοντέλα

Σε αυτήν την ενότητα παρουσιάζουμε επιβλεπόμενα μοντέλα που χρησιμοποιούν μπεϋζιανές έννοιες. Η κύρια ιδέα είναι να εξαγάγουμε μία ύστερη (posterior) πιθανότητα που δηλώνει την πιθανότητα ταυτόχρονης εμφάνισης των ζευγών κορυφών που μας ενδιαφέρουν. Ένα πλεονέκτημα ενός τέτοιου μοντέλου είναι ότι η πιθανότητα αυτή μπορεί η ίδια να χρησιμοποιηθεί σαν χαρακτηριστικό στην ταξινόμηση.

3.3.1 Πρόβλεψη συνδέσμου με ένα τοπικό πιθανοτικό μοντέλο

Οι Wang κ. ά. ([WSP07]) πρότειναν ένα τοπικό πιθανοτικό μοντέλο για την πρόβλεψη συνδέσμου που χρησιμοποιεί ένα τυχαίο πεδίο Markov (Markov Random Field – MRF), ένα μη κατευθυνόμενο μοντέλο γράφου. Για να προβλέψει αν δύο κόμβοι x και y θα συνδεθούν, εισάγει την έννοια του συνόλου κεντρικής γειτονιάς (central neighborhood set), που αποτελείται από άλλους κόμβους που βρίσκονται στην τοπική γειτονιά των x και y. Έστω $\{w, x, y, z\}$ ένα τέτοιο σύνολο. Ο κύριος στόχος του μοντέλου είναι να υπολογίσει την από κοινού πιθανότητα $P(\{w, x, y, z\})$, η οποία παριστάνει την πιθανότητα της ταυτόχρονης εμφάνισης των αντικειμένων σε αυτό το σύνολο. Από αυτήν την πιθανότητα μπορεί να υπολογισθεί η περιθώρια πιθανότητα $p(x, y)$ (για όλα τα δυνατά w και z) της ταυτόχρονης εμφάνισης των x και y. Μπορεί να υπάρχουν πολλά σύνολα κεντρικής γειτονιάς (ποικίλου μεγέθους) για το ζεύγος των x και y, πράγμα το οποίο καθιστά τη εκμάθηση της περιθώριας πιθανότητας $p(x, y)$ δύσκολη. Οι συγγραφείς χρησιμοποίησαν τα MRF για να λύσουν το πρόβλημα της μάθησης. Η προσέγγισή τους αποτελείται από τρία βήματα, που περιγράφονται παρακάτω.

Το πρώτο βήμα είναι να βρεθεί μία συλλογή από σύνολα κεντρικής γειτονιάς. Τα σύνολα κεντρικής γειτονιάς δύο κόμβων x και y μπορούν να βρεθούν με πολλούς τρόπους. Ο πιο φυσικός είναι να βρεθεί ένα συντομότερο μονοπάτι μεταξύ των x και y και όλοι οι κόμβοι σε αυτό το μονοπάτι να περιληφθούν σε ένα σύνολο κεντρικής γειτονιάς. Αν υπάρχουν πολλά συντομότερα μονοπάτια, όλα μπορούν να περιληφθούν στη συλλογή. Συμβολίζουμε με Q το σύνολο όλων των αντικειμένων που εμφανίζονται στα σύνολα κεντρικής γειτονιάς.

Το δεύτερο βήμα είναι να εξαγάγουμε τα δεδομένα εκπαίδετσης για το μοντέλο MRF, τα οποία λαμβάνονται από το αρχείο καταγραφής γεγονότων (event log) του κοινωνικού δικτύου. Ένα κοινωνικό δίκτυο σχηματίζεται από ένα χρονολογικό σύνολο γεγονότων (chronological set of events) όπου συμμετέχουν δύο ή περισσότεροι δράστες (actors) στο

δίκτυο. Δεδομένης μίας λίστας γεγονότων (event-list), οι συγγραφείς σχηματίζουν ένα σύνολο δεδομένων συναλλαγών (transaction dataset), όπου κάθε συναλλαγή περιλαμβάνει το σύνολο των δραστών που συμμετέχουν σε αυτό το γεγονός. Σε αυτό το σύνολο δεδομένων (dataset), εκτελούν μία παραλλαγή εξόρυξης συνόλων αντικειμένων (itemset mining), που ονομάζεται nonderivable itemset mining, η οποία εξάγει όλα τα μη πλεονάζοντα σύνολα αντικειμένων (μαζί με τις συχνότητές τους) στα δεδομένα συναλλαγών (transaction data). Αυτή η συλλογή περαιτέρω εκλεπτύνεται ώστε να περιλάβει μόνο τα σύνολα αντικειμένων που περιέχουν μόνο τα αντικείμενα που ανήκουν στο σύνολο Q. Αυτή η συλλογή συμβολίζεται με $\mathcal{V}_Q$.

Στο τελικό βήμα, ένα μοντέλο MRF, έστω M, εκπαιδεύεται με τα δεδομένα εκπαίδευσης. Αυτή η διαδικασία εκπαίδευσης μετασχηματίζεται σε ένα πρόβλημα βελτιστοποίησης μέγιστης εντροπίας, το οποίο λύνεται με έναν αλγόριθμο επαναληπτικής κλιμάκωσης (iterative scaling). Αν $P_M(Q)$ είναι η κατανομή πιθανότητας επί του δυναμοσυνόλου του συνόλου Q, έχουμε $\sum_{q\in\wp(Q)} P_M(q) = 1$, όπου $\wp(Q)$ είναι το δυναμοσύνολο του Q. Κάθε σύνολο αντικειμένων με την αντίστοιχη συχνότητα στο σύνολο $\mathcal{V}_Q$ επιβάλλει έναν περιορισμό σε αυτήν την κατανομή ορίζοντας μία τιμή για αυτό το συγκεκριμένο υποσύνολο του Q. Όλες οι συχνότητες μαζί περιορίζουν την κατανομή σε ένα εφικτό (feasible) σύνολο κατανομών πιθανότητας, έστω $\mathcal{P}$. Καθώς οι συχνότητες των συνόλων αντικειμένων προέρχονται από εμπειρικά δεδομένα, το σύνολο $\mathcal{P}$ είναι μη κενό. Όμως, το σύνολο των περιορισμών που επιβάλλονται από το $\mathcal{V}_Q$ υπο-περιορίζει (under-constrains) την κατανομή-στόχο, για την οποία υιοθετούμε την αρχή μέγιστης εντροπίας (maximum entropy principle), ώστε να μπορεί να ληφθεί μία μοναδική (και αμερόληπτη) εκτίμηση της $P_M(Q)$ από το σύνολο $\mathcal{P}$. Έτσι, προσπαθούμε να λύσουμε το ακόλουθο πρόβλημα βελτιστοποίησης:

$$P_M(Q) = \arg\max_{p\in\mathcal{P}} H(p) \quad (3.10),$$

όπου $H(p) = -\sum_x p(x)\log p(x)$. Το πρόβλημα βελτιστοποίησης είναι εφικτό και υπάρχει μία μοναδική κατανομή-στόχος αν οι περιορισμοί είναι συνεπείς (consistent) (σε αυτήν την περίπτωση οι περιορισμοί είναι συνεπείς καθώς ελήφθησαν από την τιμή υποστήριξης του συνόλου αντικειμένων). Η λύση έχει την ακόλουθη μορφή γινομένου:

$$P_M(Q) = \mu_0 \prod_{j:V_j\in\mathcal{V}_Q} \mu_j^{I(\text{ο περιορισμός } V_j \text{ ικανοποιείται})} \quad (3.11).$$

Εδώ, $\mu_j : j \in \{1...|\mathcal{V}_Q|\}$ είναι παράμετροι σχετικές με κάθε περιορισμό, I είναι μία δείκτρια συνάρτηση (indicator function) που εξασφαλίζει ότι ο περιορισμός λαμβάνεται

υπόψη μόνο αν ικανοποιείται και μ_0 είναι μία σταθερά κανονικοποίησης ώστε να εξασφαλισθεί ότι $\sum_{q\in\wp(Q)} P_M(q)=1$. Η τιμή των παραμέτρων μπορεί να υπολογισθεί με έναν αλγόριθμο επαναληπτικής κλιμάκωσης.

Αφού κτισθεί το μοντέλο $P_M(Q)$, μπορεί να χρησιμοποιηθεί συμπερασμός (inference) για την εκτίμηση της από κοινού πιθανότητας των κορυφών x και y. Το πλεονέκτημα ενός τοπικού μοντέλου είναι ότι ο αριθμός των μεταβλητών στο σύνολο ν_Q είναι μικρός, επομένως είναι εφικτός ο ακριβής συμπερασμός (exact inference).

3.3.2 Πιθανοτικό μοντέλο βασισμένο στην εξέλιξη του δικτύου

Οι Kashima κ. ά. ([KA06]) πρότειναν ένα ενδιαφέρον πιθανοτικό μοντέλο εξέλιξης δικτύου που μπορεί να χρησιμοποιηθεί για την πρόβλεψη συνδέσμου. Παρακάτω παρουσιάζουμε το μοντέλο και στη συνάχεια δείχνουμε πώς χρησιμοποιήθηκε στην πρόβλεψη συνδέσμου.

Το μοντέλο που προτάθηκε λαμβάνει υπόψη μόνο τις τοπολογικές (δομικές) ιδιότητες του δικτύου. Για ένα γράφο $G(V,\phi)$, όπου V είναι το σύνολο των κόμβων και $\phi: V\times V\to[0,1]$ μία συνάρτηση που αποδίδει ετικέτες στις ακμές, με $\phi(x,y)$ δηλώνεται η πιθανότητα να υπάρχει μία ακμή μεταξύ των κόμβων x και y στο γράφο G. Συγκεκριμένα, $\phi(x,y)=1$ αν υπάρχει ακμή και $\phi(x,y)=0$ αν δεν υπάρχει. Με $\phi^{(t)}$ συμβολίζουμε τη συνάρτηση ϕ τη στιγμή t (η συνάρτηση ϕ αλλάζει με το χρόνο). Επιπλέον, το μοντέλο είναι μαρκοβιανό, δηλαδή η $\phi^{(t+1)}$ εξαρτάται μόνο από τη $\phi^{(t)}$. Το σύνολο V θεωρείται σταθερό. Το μοντέλο εξελίσσεται με το χρόνο ως εξής: Μία ετικέτα αντιγράφεται από τον κόμβο l στον κόμβο m τυχαία με πιθανότητα w_{lm}. Αρχικά, το μοντέλο επιλέγει τους l και m. Στη συνέχεια, επιλέγει μία ετικέτα ομοιόμορφα από τις $|V|-1$ ετικέτες του l (αποκλείεται η $\phi(l,m)$) και την αντιγράφει σαν ετικέτα του m. Το μοντέλο ικανοποιεί τους ακόλουθους περιορισμούς:

$$\sum_{l,m} w_{lm}=1,\ w_{lm}>0,\ w_{ll}=0 \quad (3.12).$$

Η παραπάνω ιδέα μοιάζει με τη μεταβατική ιδιότητα των κοινωνικών δικτύων. Μέσω της διαδικασίας αντιγραφής ετικετών, ο l μπορεί να γίνει φίλος ενός φίλου του m. Η εφαρμογή της μάθησης στο παραπάνω μοντέλο είναι ο υπολογισμός των βαρών w_{lm} και των ετικετών $\phi^{(t+1)}$ δεδομένων των ετικετών $\phi^{(t)}$ από το σύνολο των δεδομένων εκπαίδευσης.

Μπορούμε να διακρίνουμε δύο τρόπους με τους οποίους η ετικέτα $\phi^{(t+1)}(i,j)$ μπορεί να πάρει μια συγκεκριμένη τιμή. Ο ένας είναι ο κόμβος k να αντέγραψε μία ετικέτα του σε κάποιον από τους i και j. Ο άλλος είναι η αντιγραφή να συνέβη αλλού και να ισχύει $\phi^{(t+1)}(i,j)=\phi^{(t)}(i,j)$. Σύμφωνα με την παρατήρηση αυτή, έχουμε:

$$\begin{aligned}\phi^{(t+1)}(i,j)=&\frac{1}{|V|-1}\sum_{k\neq i,j} w_{kj}\phi^{(t)}(k,i)I(\phi^{(t)}(k,j)=1)\\&+\frac{1}{|V|-1}\sum_{k\neq i,j} w_{ki}\phi^{(t)}(k,j)I(\phi^{(t)}(k,i)=1)\\&+\left(1-\frac{1}{|V|-1}\sum_{k\neq i,j}\left(w_{kj}+w_{ki}\right)\right)\phi^{(t)}(i,j)\end{aligned} \quad (3.13).$$

Ας σημειωθεί ότι, στην περίπτωση που συμβαίνει αντιγραφή, αν ο κόμβος k αντιγράφει την ετικέτα του στον κόμβο i, τότε πρέπει να υπάρχει ήδη μία ακμή μεταξύ των k και j, ενώ αν ο κόμβος k αντιγράφει την ετικέτα του στον κόμβο j, τότε πρέπει να υπάρχει ήδη μία ακμή μεταξύ των k και i. Αυτή η απαίτηση δηλώνεται με τη δείκτρια συνάρτηση I, η οποία παίρνει την τιμή 0 αν η συνθήκη μέσα στην παρένθεση δεν ικανοποιείται. Με επαναληπτική εφαρμογή της εξίσωσης στις ετικέτες, η δομή του δικτύου εξελίσσεται με το χρόνο.

Για την εφαρμογή της πρόβλεψης συνδέσμου, το μοντέλο θεωρεί ότι την παρούσα στιγμή το δίκτυο βρίσκεται στη στάσιμη κατάσταση, δηλαδή $\phi^{(\infty)}(k,i)=\phi^{(t+1)}(k,i)=\phi^{(t)}(k,i)$. Εισάγοντας την υπόθεση αυτή στην εξίσωση (3.13), έχουμε την ακόλουθη εξίσωση:

$$\phi^{(\infty)}(i,j)=\frac{\sum_{k\neq i,j}\left(w_{kj}\phi^{(\infty)}(k,i)+w_{ki}\phi^{(\infty)}(k,j)\right)}{\sum_{k\neq i,j}(w_{kj}+w_{ki})} \quad (3.14).$$

Η λογαριθμική πιθανοφάνεια (log-likelihood) της ετικέτας $\phi(i,j)$ μπορεί να γραφεί:

$$\begin{aligned}L_{ij}=&\phi^{(\infty)}(i,j)\log\frac{\sum_{k\neq i,j}\left(w_{kj}\phi^{(\infty)}(k,i)+w_{ki}\phi^{(\infty)}(k,j)\right)}{\sum_{k\neq i,j}(w_{kj}+w_{ki})}\\&\left(1-\phi^{(\infty)}(i,j)\right)\log\left(1-\frac{\sum_{k\neq i,j}\left(w_{kj}\phi^{(\infty)}(k,i)+w_{ki}\phi^{(\infty)}(k,j)\right)}{\sum_{k\neq i,j}(w_{kj}+w_{ki})}\right)\end{aligned} \quad (3.15).$$

Η συνολική λογαριθμική πιθανοφάνεια (total log-likelihood) για τις γνωστές ετικέτες ορίζεται ως:

$$L(W)=\sum_{(i,j)\in E^{\text{train}}} L_{ij} \quad (3.16).$$

Η διαδικασία εκτίμησης των παραμέτρων ανάγεται στην επίλυση του παρακάτω προβλήματος βελτιστοποίησης:

$$\text{Maximize}_{w,\phi^{(\infty)}(i,j)\ \text{για}\ (i,j)\in E^{\text{train}}} L(W)$$

τ. ώ.

$$\phi^{(\infty)}(i,j) = \frac{\sum_{k\neq i,j}\left(w_{kj}\phi^{(\infty)}(k,i) + w_{ki}\phi^{(\infty)}(k,j)\right)}{\sum_{k\neq i,j}(w_{kj}+w_{ki})}, \forall (i,j)\in E^{\text{train}} \quad (3.17).$$

$$\text{και} \sum_{l,m} w_{lm} = 1, w_{lm} \geq 0$$

Το παραπάνω πρόβλημα βελτιστοποίησης μπορεί να λυθεί με μεταβιβαστική μάθηση (transductive learning) τύπου μεγιστοποίησης της προσδοκίας.

Το πλεονέκτημα αυτού του μοντέλου είναι ότι είναι πολύ γενικό και μπορεί να εφαρμοστεί σε οποιοδήποτε κοινωνικό δίκτυο. Επιπλέον, η μάθηση που βασίζεται στη μεγιστοποίηση της προσδοκίας δίνει έναν αποδοτικό αλγόριθμο. Ωστόσο, η επίδοση του αλγορίθμου εξαρτάται εξολοκλήρου από το βαθμό στον οποίο το δίκτυο συμφωνεί με το μοντέλο εξέλιξης γράφου που προτείνεται.

3.3.3 Ιεραρχικό πιθανοτικό μοντέλο

Οι Clauset κ. ά. ([CMN08]) πρότειναν ένα πιθανοτικό μοντέλο το οποίο λαμβάνει υπόψη την ιεραρχική οργάνωση στο δίκτυο, όπου οι κορυφές διαχωρίζονται σε ομάδες που περαιτέρω διαχωρίζονται σε υποομάδες κ.ο.κ. Το μοντέλο συμπεραίνει (infers) την ιεραρχική δομή από δεδομένα του δικτύου και μπορεί να χρησιμοποιηθεί για πρόβλεψη απόντων (missing) συνδέσμων. Προτείνεται ως πιθανοτικό μοντέλο για ιεραρχικούς τυχαίους γράφους. Η εφαρμογή της μάθησης έγκειται στη χρήση των παρατηρούμενων δεδομένων του δικτύου για την προσαρμογή της πιο πιθανής ιεραρχικής δομής μέσω στατιστικού συμπερασμού (statistical inference) — ένας συνδυασμός της προσέγγισης μέγιστης πιθανοφάνειας και ενός αλγορίθμου δειγματοληψίας Monte Carlo.

Έστω G ένας γράφος με n κορυφές. Ένα δενδρόγραμμα (dendrogram) D είναι ένα δυαδικό δένδρο με n φύλλα που αντιστοιχούν στις κορυφές του G. Καθένας από τους $n-1$ εσωτερικούς κόμβους του D αντιστοιχεί στην ομάδα των κορυφών που είναι απόγονοί του. Με κάθε εσωτερικό κόμβο r συσχετίζεται μία πιθανότητα p_r. Αν i, j είναι δύο κορυφές του G, η πιθανότητα p_{ij} να συνδεθούν με ακμή είναι $p_{ij} = p_r$, όπου r είναι ο χαμηλότερος κοινός πρόγονος (lowest common ancestor) τους στο D. Ο συνδυασμός $(D,\{p_r\})$ του δενδρογράμματος και του συνόλου πιθανοτήτων ορίζει έναν ιεραρχικό τυχαίο γράφο (hierarchical random graph).

Η εφαρμογή της μάθησης έγκειται στην εύρεση του ιεραρχικού τυχαίου γράφου (ή γράφων) που προσαρμόζεται καλύτερα στα παρατηρούμενα δεδομένα του δικτύου. Υποθέτοντας ότι όλοι οι ιεραρχικοί τυχαίοι γράφοι είναι εκ των προτέρων εξίσου πιθανοί, συμπεραίνουμε ότι

η πιθανότητα ένα δεδομένο μοντέλο $(D,\{p_r\})$ να είναι η σωστή επεξήγηση των δεδομένων είναι, από το θεώρημα του Bayes, ανάλογη προς την ύστερη (posterior) πιθανότητα ή πιθανοφάνεια $\mathcal{L}$ με την οποία το μοντέλο παράγει το παρατηρούμενο δίκτυο. Ο στόχος είναι να μεγιστοποιήσουμε την $\mathcal{L}$.

Έστω E_r ο αριθμός των ακμών του G των οποίων τα άκρα έχουν χαμηλότερο κοινό πρόγονο στο D τον r. Επίσης, έστω L_r και R_r ο αριθμός των φύλλων στο αριστερό και το δεξιό υποδένδρο με ρίζα τον r, αντίστοιχα. Τότε, η πιθανοφάνεια του ιεραρχικού τυχαίου γράφου είναι $\mathcal{L}(D,\{p_r\}) = \prod_{r\in D} p_r^{E_r}(1-p_r)^{L_rR_r-E_r}$ (με τη σύμβαση $0^0=1$). Αν σταθεροποιήσουμε το δενδρόγραμμα D είναι εύκολο να βρούμε τις πιθανότητες $\{\overline{p}_r\}$ που μεγιστοποιούν την $\mathcal{L}(D,\{p_r\})$, οι οποίες είναι:

$$\overline{p}_r = \frac{E_r}{L_rR_r} \quad (3.18),$$

το κλάσμα των δυνατών ακμών μεταξύ των δύο υποδένδρων του r που εμφανίζονται στο γράφο G. Ο λογάριθμος της πιθανοφάνειας είναι:

$$\log \mathcal{L}(D) = -\sum_{r\in D} L_rR_rh(\overline{p}_r) \quad (3.19),$$

όπου $h(p) = -p\log p - (1-p)\log(1-p)$. Ας σημειωθεί ότι κάθε όρος $-L_rR_rh(\overline{p}_r)$ μεγιστοποιείται όταν το $\overline{p}_r$ είναι κοντά στο 0 ή κοντά στο 1. Με άλλα λόγια, τα δενδρογράμματα με υψηλή πιθανοφάνεια είναι αυτά τα οποία διαμερίζουν το σύνολο των κορυφών σε ομάδες οι συνδέσεις μεταξύ των οποίων είναι είτε πολύ πυκνές είτε πολύ αραιές.

Η επιλογή μεταξύ των δενδρογραμμάτων γίνεται με μία μέθοδο δειγματοληψίας MCMC (Markov chain Monte Carlo) με πιθανότητες ανάλογες προς τις πιθανοφάνειές τους. Για τη δημιουργία της αλυσίδας Markov, η μέθοδος αρχικά δημιουργεί ένα σύνολο μεταβάσεων μεταξύ των δυνατών δενδρογραμμάτων μέσω αναδιάταξης (rearrangement). Για την αναδιάταξη, η μέθοδος επιλέγει έναν εσωτερικό κόμβο ενός δενδρογράμματος και έπειτα επιλέγει ομοιόμορφα μεταξύ των διάφορων διατάξεων (configurations) του υποδένδρου αυτού του κόμβου. Μόλις τα κριτήρια μετάβασης γίνουν γνωστά η διαδικασία δειγματοληψίας ξεκινά έναν τυχαίο περίπατο. Μία νέα αναδιάταξη γίνεται αποδεκτή σύμφωνα με τον κανόνα δειγματοληψίας Metropolis-Hastings, δηλαδή μία μετάβαση από ένα δενδρόγραμμα D σε ένα άλλο αναδιατεταγμένο δενδρόγραμμα D' γίνεται αποδεκτή αν η ποσότητα $\Delta\log\mathcal{L} = \log\mathcal{L}(D') - \log\mathcal{L}(D)$ είναι μη αρνητική, διαφορετικά γίνεται αποδεκτή με πιθανότητα $\mathcal{L}(D')/\mathcal{L}(D)$. Οι συγγραφείς απέδειξαν ότι ο τυχαίος περίπατος είναι

εργοδικός και στη στατική κατανομή τα δενδρογράμματα δειγματοληπτούνται σύμφωνα με την πιθανοφάνειά τους.

Για την εφαρμογή της πρόβλεψης συνδέσμου, ένα σύνολο δενδρογραμμάτων-δειγμάτων λαμβάνονται κατά τακτά διαστήματα αφού ο τυχαίος περίπατος MCMC φθάσει σε ισορροπία. Τότε, για ένα ζεύγος κορυφών x και y μεταξύ των οποίων δεν υπάρχει σύνδεση, το μοντέλο υπολογίζει μία μέση πιθανότητα p_{xy} να συνδεθούν θεωρώντας το μέσο όλων των αντίστοιχων πιθανοτήτων p_{xy} σε όλα τα δειγματοληπτημένα δενδρογράμματα. Για μία δυαδική (binary) απόφαση, μπορεί να γίνει βαθμονόμηση μοντέλου (model calibration) μέσω ενός συνόλου δεδομένων βαθμονόμησης (calibration dataset). Το μοντέλο έχει την ιδιαιτερότητα ότι επιτρέπει ιεραρχία. Επίσης, επιτρέπει τη λήψη δειγμάτων από το σύνολο των ιεραρχικών δομών για τη λήψη μίας πιθανότητας συναίνεσης (consensus probability). Από την άλλη πλευρά, μπορεί να μην είναι πολύ ακριβές αν η μέθοδος MCMC δε συγκλίνει στη στατική κατανομή σε εύλογο αριθμό βημάτων. Επίσης, για μεγάλους γράφους η όλη διαδικασία μπορεί να είναι πολύ δαπανηρή.

3.4 Πιθανοτικά σχεσιακά μοντέλα

Τα πιθανοτικά σχεσιακά μοντέλα (probabilistic relational models – PRMs) παριστάνουν μία από κοινού κατανομή πιθανότητας επί των γνωρισμάτων ενός σχεσιακού συνόλου δεδομένων. Επιτρέπουν οι ιδιότητες ενός αντικειμένου να εξαρτώνται πιθανοτικά τόσο από άλλες ιδιότητές του όσο και από ιδιότητες σχετιζόμενων αντικειμένων. Σε αντίθεση με τα παραδοσιακά μοντέλα γράφων, που χρησιμοποιούν ένα γράφο για να μοντελοποιήσουν τη σχέση μεταξύ των γνωρισμάτων ομοιογενών οντοτήτων, τα PRMs αποτελούνται από τρεις γράφους: το γράφο δεδομένων G_D, το γράφο του μοντέλου G_M και το γράφο συμπερασμού G_I.

Ο γράφος δεδομένων $G_D = (V_D, E_D)$ παριστάνει το δίκτυο εισόδου, όπου οι κόμβοι είναι τα αντικείμενα και οι ακμές παριστάνουν τις σχέσεις μεταξύ των αντικειμένων. Κάθε κόμβος $v_i \in V_D$ και ακμή $e_j \in E_D$ σχετίζονται με έναν τύπο $T(v_i) = t_{v_i}$, $T(e_j) = t_{e_j}$. Κάθε τύπος στοιχείου (αντικειμένου ή ακμής) $t \in T$ έχει έναν αριθμό σχετιζόμενων γνωρισμάτων X^t. Συνεπώς, κάθε αντικείμενο v_i και σύνδεσμος e_j σχετίζονται με ένα σύνολο τιμών γνωρισμάτων, $\mathrm{x}_{v_i}^{t_{v_i}}$ και $\mathrm{x}_{e_j}^{t_{e_j}}$, που καθορίζεται από τους τύπους τους, t_v και t_e, αντίστοιχα. Ένα PRM παριστάνει μία από κοινού κατανομή πιθανότητας επί των τιμών όλων των γνωρισμάτων του γράφου δεδομένων $\mathrm{x} = \{\mathrm{x}_{v_i}^{t_{v_i}} : v_i \in V_D, T(v_i) = t_{v_i}\} \bigcup \{\mathrm{x}_{e_j}^{t_{e_j}} : e_j \in E_D, T(e_j) = t_{e_j}\}$.

Ο γράφος του μοντέλου $G_M = (V_M, E_M)$ παριστάνει τις εξαρτήσεις μεταξύ των γνωρισμάτων στο επίπεδο των τύπων των στοιχείων. Τα γνωρίσματα ενός στοιχείου μπορεί να εξαρτώνται πιθανοτικά από άλλα γνωρίσματα του ίδιου στοιχείου, καθώς επίσης και από γνωρίσματα άλλων σχετιζόμενων αντικειμένων ή συνδέσμων στον G_D. Κάθε κόμβος στο σύνολο V_M αντιστοιχεί σε ένα γνώρισμα $X_i^t \in \mathrm{X}^t$ όπου $t \in T$. Ο G_M αποτελείται από δύο μέρη: τη δομή εξάρτησης μεταξύ όλων των γνωρισμάτων των τύπων και την υπό συνθήκη κατανομή πιθανότητας που σχετίζεται με τους κόμβους στον G_M.

Ο γράφος συμπερασμού $G_I = (V_I, E_I)$ παριστάνει τις πιθανοτικές εξαρτήσεις μεταξύ όλων των μεταβλητών σε ένα σύνολο δοκιμής. Μπορεί να λάβει υπόσταση (instantiated) με μία διαδικασία roll-out των G_D και G_M . Κάθε ζεύγος στοιχείου-γνωρίσματος στον G_D παίρνει ένα ξεχωριστό αντίγραφο της αντίστοιχης υπό συνθήκη κατανομής πιθανότητας από τον G_M. Η δομή του G_I καθορίζεται τόσο από τον G_D, όσο και από τον G_M.

Υπάρχουν δύο καινοτόμες προσεγγίσεις των PRMs, από τις οποίες η μία βασίζεται σε μπεϋζιανά δίκτυα και θεωρεί τους συνδέσμους μη κατευθυνόμενους ([GFKT02]) και η άλλη βασίζεται σε σχεσιακά μαρκοβιανά δίκτυα και θεωρεί τους συνδέσμους κατευθυνόμενους ([TWAK03]). Αν και αμφότερες είναι κατάλληλες για την εφαρμογή της πρόβλεψης συνδέσμου, για τα περισσότερα δίκτυα ένα μη κατευθυνόμενο μοντέλο φαίνεται πιο κατάλληλο λόγω της ευελιξίας του.

3.5 Μέθοδοι γραμμικής άλγεβρας

Οι Kunegis κ. ά. ([KL09]) πρότειναν μία πολύ γενική μέθοδο που γενικεύει μεθόδους πυρήνων γράφων (graph kernels) και μείωσης της διαστατικότητας (dimensionality reduction) για την επίλυση του προβλήματος της πρόβλεψης συνδέσμου. Αυτή η μέθοδος έχει την ιδιαιτερότητα ότι προτείνει τη μάθηση μίας συνάρτησης F η οποία δρα κατευθείαν στον πίνακα γειτνίασης ή το Λαπλασιανό (Laplacian) πίνακα του γράφου.

Έστω $\mathbf{A}$ και $\mathbf{B}$ δύο πίνακες γειτνίασης των συνόλων εκπαίδευσης και δοκιμής για την πρόβλεψη συνδέσμου. Υποθέτουμε ότι αναφέρονται στο ίδιο σύνολο κορυφών. Θεωρούμε ένα φασματικό μετασχηματισμό (spectral transformation function) F που απεικονίζει τον $\mathbf{A}$ στο $\mathbf{B}$ με ελάχιστο σφάλμα και δίνεται από τη λύση του ακόλουθου προβλήματος βελτιστοποίησης:

$$\min_F \left\| F(\mathbf{A}) - \mathbf{B} \right\|_F \quad (3.20),$$
$$\tau.\acute{\omega}.\ F \in \mathcal{S}$$

όπου $\left\| \cdot \right\|_F$ η νόρμα του Frobenius. Ο περιορισμός εξασφαλίζει ότι η συνάρτηση F ανήκει στην οικογένεια των φασματικών μετασχηματισμών ($\mathcal{S}$). Δεδομένου ενός συμμετρικού

πίνακα $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, για μία τέτοια συνάρτηση F, έχουμε $F(\mathbf{A}) = \mathbf{U}F(\mathbf{\Lambda})\mathbf{U}^T$, όπου η F εφαρμόζει την αντίστοιχη συνάρτηση πραγματικής μεταβλητής σε κάθε ιδιοτιμή ξεχωριστά.

Αυτό το πρόβλημα βελτιστοποίησης μπορεί να λυθεί με παραγοντοποίηση με βάση τις ιδιοτιμές (eigenvalue decomposition) $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ και χρησιμοποιώντας το γεγονός ότι η νόρμα του Frobenius παραμένει αναλλοίωτη στον πολλαπλασιασμό με ορθογώνιο πίνακα:

$$\begin{aligned} &\left\| F(\mathbf{A}) - \mathbf{B} \right\|_F \\ &= \left\| \mathbf{U}F(\mathbf{\Lambda})\mathbf{U}^T - \mathbf{B} \right\|_F \quad (3.21). \\ &= \left\| F(\mathbf{\Lambda}) - \mathbf{U}^T\mathbf{B}\mathbf{U} \right\|_F \end{aligned}$$

Αφού τα μη διαγώνια στοιχεία στην παραπάνω έκφραση δεν εξαρτώνται από τη συνάρτηση F, η ζητούμενη συνάρτηση μπορεί να βρεθεί με λύση του ακόλουθου προβλήματος:

$$\min_f \sum_i (f(\mathbf{\Lambda}_{ii}) - \mathbf{U}_{.i}^T\mathbf{B}\mathbf{U}_{.i})^2 \quad (3.22).$$

Έτσι, το πρόβλημα της πρόβλεψης συνδέσμου ανάγεται σε ένα μονοδιάστατο πρόβλημα προσαρμογής καμπύλης (curve fitting) ελαχίστων τετραγώνων.

Η παραπάνω γενική μέθοδος μπορεί να χρησιμοποιηθεί για προσαρμογή πολλών δυνατών φασματικών μετασχηματισμών. Συγκεκριμένα, αναζητούμε μία συνάρτηση F η οποία δέχεται έναν πίνακα και επιστρέφει έναν άλλο πίνακα ο οποίος είναι κατάλληλος για την πρόβλεψη συνδέσμου, δηλαδή τα στοιχεία του κωδικοποιούν την ομοιότητα των αντίστοιχων ζευγών κορυφών. Υπάρχουν πολλοί πυρήνες γράφων (graph kernels) που μπορούν να χρησιμοποιηθούν για τη συνάρτηση F.

Εκθετικός πυρήνας. Για έναν πίνακα γειτνίασης $\mathbf{A}$ ενός γράφου χωρίς βάρη, οι δυνάμεις $\mathbf{A}^n$ δηλώνουν τον αριθμό των μονοπατιών μήκους n που συνδέουν τα ζεύγη των κόμβων. Με βάση την ιδέα ότι οι κόμβοι που συνδέονται με πολλά μονοπάτια πρέπει να θεωρούνται πλησιέστεροι από ό,τι οι κόμβοι που συνδέονται με λίγα μονοπάτια, μπορούμε να θεωρήσουμε μία συνάρτηση F για την πρόβλεψη συνδέσμου της ακόλουθης μορφής:

$$F_P(\mathbf{A}) = \sum_{i=0}^{d} \alpha_i \mathbf{A}^i \quad (3.23).$$

Οι σταθερές α_i πρέπει να μειώνονται καθώς αυξάνεται το i, για να περιορισθεί η συνεισφορά των μακρύτερων μονοπατιών. Ένας εκθετικός πυρήνας μπορεί να εκφρασθεί όπως παρακάτω:

$$\exp(\alpha\mathbf{A}) = \sum_{i=0}^{\infty} \frac{\alpha^i}{i!} \mathbf{A}^i \quad (3.24).$$

Πυρήνας von Neumann. Ορίζεται παρόμοια με τον εκθετικό πυρήνα:

$$(\mathbf{I}-\alpha\mathbf{A})^{-1}=\sum_{i=0}^{\infty}\alpha^{i}\mathbf{A}^{i} \qquad (3.25)$$

Λαπλασιανοί πυρήνες. Η γενική ιδέα που προτείνεται σε αυτήν τη μέθοδο δε μας περιορίζει να χρησιμοποιήσουμε συναρτήσεις που δρουν στον πίνακα γειτνίασης, Μπορεί κανείς να χρησιμοποιήσει συναρτήσεις που εφαρμόζονται στο Λαπλασιανό πίνακα $\mathbf{L}$, ο οποίος ορίζεται ως $\mathbf{L}=\mathbf{D}-\mathbf{A}$, όπου $\mathbf{D}$ είναι ο διαγώνιος πίνακας βαθμών. Ο κανονικοποιημένος Λαπλασιανός πίνακας $\mathcal{L}$ ορίζεται ως $\mathcal{L}=\mathbf{I}-\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$. Πολλοί πυρήνες γράφων ορίζονται επί του Λαπλασιανού πίνακα. Για παράδειγμα, παίρνοντας τον ψευδοαντίστροφο κατά Moore-Penrose[1] του Λαπλασιανού πίνακα έχουμε τον πυρήνα χρόνου επανόδου:

$$\begin{aligned}F_{COM}(\mathbf{L})&=\mathbf{L}^{+}\\ F_{COM}(\mathcal{L})&=\mathcal{L}^{+}\end{aligned} \quad (3.26).$$

Εφαρμόζοντας ομαλοποίηση (regularization) έχουμε τους ομαλοποιημένους πυρήνες χρόνου επανόδου:

$$\begin{aligned}F_{COMR}(\mathbf{L})&=(\mathbf{I}+\alpha\mathbf{L})^{-1}\\ F_{COMR}(\mathcal{L})&=(\mathbf{I}+\alpha\mathcal{L})^{-1}\end{aligned} \quad (3.27).$$

Μπορούμε επίσης να πάρουμε τους πυρήνες διάχυσης θερμότητας:

$$\begin{aligned}f_{HEAT}(\mathbf{L})&=\exp(-\alpha\mathbf{L})\\ f_{HEAT}(\mathcal{L})&=\exp(-\alpha\mathcal{L})\end{aligned} \quad (3.28).$$

Τα πλεονεκτήματα αυτής της μεθόδου είναι η γενικότητα και η απλότητά της. Ο αριθμός των παραμέτρων προς μάθηση στο μοντέλο αυτό είναι πολύ μικρότερος σε σχέση με πολλά άλλα μοντέλα που παρουσιάσαμε. Όμως, αυτό το μοντέλο δεν μπορεί να ενσωματώσει γνωρίσματα των κορυφών. Επιπλέον το υπολογιστικό του κόστος εξαρτάται ως επί το πλείστον από την παραγοντοποίηση με βάση τις ιδιοτιμές (eigen-decomposition) του πίνακα $\mathbf{A}$, που για

---

[1] Ψευδοαντίστροφος κατά Moore-Penrose ενός μιγαδικού πίνακα A τύπου $m\times n$ ονομάζεται ένας μιγαδικός πίνακας B τύπου $n\times m$ τέτοιος ώστε:

1. $ABA=A$.
2. $BAB=B$.
3. Ο πίνακας AB είναι ερμιτιανός (δηλαδή $(AB)^{*}=AB$).
4. Ο πίνακας BA είναι ερμιτιανός (δηλαδή $(BA)^{*}=BA$).

Αποδεικνύεται ότι για κάθε πίνακα A υπάρχει ένας μοναδικός τέτοιος πίνακας B, ο οποίος συμβολίζεται με A^{+}. Στην περίπτωση, μάλιστα, που ο A είναι τετραγωνικός και αντιστρέψιμος, ο ψευδοαντίστροφός του κατά Moore-Penrose ταυτίζεται με τον αντίστροφό του, δηλαδή $A^{+}=A^{-1}$.

μεγάλους πίνακες είναι δαπανηρή. Ωστόσο, έχουν προταθεί αποδοτικές μέθοδοι για αυτήν την εφαρμογή ([FKV98]).

3.6 Δείκτες αξιολόγησης

Θεωρούμε ένα μη κατευθυνόμενο δίκτυο $G(V, E)$ όπου V είναι το σύνολο των κόμβων και E το σύνολο των συνδέσμων. Με U συμβολίζουμε το καθολικό (universal) σύνολο των $\frac{|V|\cdot(|V|-1)}{2}$ δυνατών συνδέσμων. Τότε, το σύνολο των ανύπαρκτων συνδέσμων είναι $U-E$. Υποθέτουμε ότι το σύνολο $U-E$ περιέχει κάποιους απόντες συνδέσμους (ή τους συνδέσμους που θα εμφανισθούν στο μέλλον).

Γενικά, δε γνωρίζουμε ποιοι σύνδεσμοι είναι οι απόντες. Έτσι, για να ελέγξουμε την ορθότητα (accuracy) ενός αλγορίθμου, διαμερίζουμε το σύνολο των παρατηρούμενων συνδέσμων E τυχαία σε δύο μέρη: το σύνολο εκπαίδευσης E^T, το οποίο θεωρούμε ως γνωστή πληροφορία, και το σύνολο δοκιμής E^P, το οποίο χρησιμοποιούμε για δοκιμή. Με αυτήν τη μέθοδο κάποιοι σύνδεσμοι μπορεί να μην επιλεγούν ποτέ στο σύνολο δοκιμής, ενώ άλλοι μπορεί να επιλεγούν περισσότερες από μία φορές, με αποτέλεσμα στατιστική πόλωση (statistical bias). Ο περιορισμός αυτός μπορεί να ξεπεραστεί χρησιμοποιώντας την K-πλή διασταυρωμένη επικύρωση (K-fold cross-validation), στην οποία το σύνολο των παρατηρούμενων συνδέσμων διαμερίζεται τυχαία σε K υποσύνολα. Κάθε φορά ένα υποσύνολο επιλέγεται ως σύνολο δοκιμής και τα υπόλοιπα $K-1$ υποσύνολα συνιστούν το σύνολο εκπαίδευσης. Η διαδικασία διασταυρωμένης επικύρωσης επαναλαμβάνεται K φορές, με κάθε ένα από τα K υποσύνολα να χρησιμοποιείται ακριβώς μία φορά ως το σύνολο δοκιμής. Με αυτήν τη μέθοδο, όλοι οι σύνδεσμοι χρησιμοποιούνται τόσο για εκπαίδευση, όσο και για επικύρωση και κάθε σύνδεσμος χρησιμοποιείται για πρόβλεψη ακριβώς μία φορά. Είναι προφανές ότι μεγαλύτερη τιμή του K οδηγεί σε μικρότερη στατιστική πόλωση αλλά απαιτεί περισσότερο υπολογισμό.

Δύο καθιερωμένα μέτρα χρησιμοποιούνται για την ποσοτικοποίηση της ορθότητας των αλγορίθμων πρόβλεψης συνδέσμου: το εμβαδό κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη (area under the receiver operating characteristic curve – AUC) ([HM82]) και η ακρίβεια (precision) ([Gei93], [HKTR04]). Γενικά, ένας αλγόριθμος πρόβλεψης συνδέσμου δίνει μία διατεταγμένη λίστα όλων των μη παρατηρούμενων συνδέσμων (δηλαδή των συνδέσμων του συνόλου $U-E^T$) ή, ισοδύναμα, δίνει σε κάθε μη παρατηρούμενο σύνδεσμο $(x, y) \in U-E^T$ μία βαθμολογία s_{xy} για να ποσοτικοποιήσει την πιθανότητα ύπαρξής του. Το AUC αξιολογεί την επίδοση του αλγορίθμου με βάση ολόκληρη

τη λίστα, ενώ η ακρίβεια βασίζεται μόνο στους L συνδέσμους με τις μεγαλύτερες βαθμολογίες.

3.6.1 AUC

Με δεδομένη τη σειρά κατάταξης (rank) όλων των μη παρατηρούμενων συνδέσμων, η τιμή του AUC μπορεί να ερμηνευθεί ως η πιθανότητα ένας τυχαία επιλεγμένος απών σύνδεσμος (δηλαδή ένας σύνδεσμος του συνόλου E^P) να έχει λάβει μεγαλύτερη βαθμολογία από έναν τυχαία επιλεγμένο ανύπαρκτο σύνδεσμο (δηλαδή ένα σύνδεσμο του συνόλου $U - E$). Συνήθως υπολογίζουμε τη βαθμολογία κάθε μη παρατηρούμενου συνδέσμου αντί να δίνουμε τη διατεταγμένη λίστα. Έπειτα, κάθε φορά επιλέγουμε τυχαία έναν απόντα σύνδεσμο και έναν ανύπαρκτο σύνδεσμο για να συγκρίνουμε τις βαθμολογίες τους. Αν σε n ανεξάρτητες συγκρίσεις n' φορές ο απών σύνδεσμος έχει μεγαλύτερη βαθμολογία και n'' φορές οι δύο σύνδεσμοι έχουν το ίδιο, τότε η τιμή του AUC είναι:

$$\text{AUC} = \frac{n' + 0.5n''}{n} \text{ (3.29).}$$

Αν όλες οι βαθμολογίες παράγονται από ανεξάρτητες και ισόνομες κατανομές (independent and identical distributions) η τιμή του AUC θα πρέπει να είναι περίπου 0.5. Έτσι, ο βαθμός στον οποίο η τιμή υπερβαίνει το 0.5 δείχνει πόσο καλύτερα αποδίδει ο αλγόριθμος από την καθαρή τύχη.

3.6.2 Ακρίβεια

Δεδομένης της κατάταξης των μη παρατηρούμενων συνδέσμων, η ακρίβεια ορίζεται ως ο λόγος των σχετικών αντικειμένων που επελέγησαν προς τον αριθμό των αντικειμένων που επελέγησαν. Δηλαδή, αν πάρουμε τους L συνδέσμους με τις μεγαλύτερες βαθμολογίες ως τους προβλεπόμενους και μεταξύ αυτών L_r είναι σωστοί (δηλαδή ανήκουν στο σύνολο E^P), τότε η ακρίβεια είναι L_r / L. Προφανώς, μεγαλύτερη ακρίβεια σημαίνει μεγαλύτερη ορθότητα πρόβλεψης.

3.7 Κατευθύνσεις έρευνας στην πρόβλεψη συνδέσμου

3.7.1 Κατευθυνόμενα δίκτυα – δίκτυα με βάρη

Η μελέτη του προβλήματος της πρόβλεψης συνδέσμου μέχρι τώρα έχει εστιασθεί στην περίπτωση των μη κατευθυνόμενων δικτύων χωρίς βάρη. Στα κατευθυνόμενα δίκτυα οι απλοί δείκτες ομοιότητας που βασίζονται στους κοινούς γείτονες πρέπει να τροποποιηθούν. Ακόμη

και όταν μπορούμε να προβλέψουμε την ύπαρξη μίας ακμής μεταξύ δύο κόμβων, δεν μπορούμε να καθορίσουμε την κατεύθυνσή της. Επίσης, οι δείκτες ομοιότητας που βασίζονται σε μονοπάτια πρέπει να επεκταθούν ώστε να λαμβάνουν υπόψη την κατεύθυνση του συνδέσμου. Στα δίκτυα με βάρη η εφαρμογή της πρόβλεψης συνδέσμου έχει μελετηθεί από τους Murata κ. ά. ([MM07]) και Lu κ. ά. ([LZ10]). Οι πρώτοι πρότειναν ότι οι σύνδεσμοι με μεγαλύτερα βάρη είναι σημαντικότεροι στην πρόβλεψη απόντων συνδέσμων, ενώ οι δεύτεροι κατέληξαν στο αντίθετο συμπέρασμα, δηλαδή ότι οι ασθενέστεροι σύνδεσμοι παίζουν σημαντικότερο ρόλο. Η σωστή αξιοποίηση της πληροφορίας των βαρών ώστε να βελτιωθεί η ακρίβεια της πρόβλεψης είναι ένα άλυτο πρόβλημα. Ένα δυσκολότερο πρόβλημα είναι η πρόβλεψη του βάρους των συνδέσμων, που σχετίζεται με την πρόβλεψη της κίνησης στα αστικά και αεροπορικά συστήματα μεταφορών.

3.7.2 Χρονική πληροφορία

Οι περισσότερες σημερινές προσεγγίσεις λαμβάνουν υπόψη ένα στιγμιότυπο του δικτύου για να προβλέψουν τους μελλοντικούς συνδέσμους. Η στατική αναπαράσταση του γράφου, ωστόσο, δυσκολεύει την πρόβλεψη επανειλημμένων εμφανίσεων συνδέσμων. Για παράδειγμα, είναι αδύνατο να προβλεφθεί αν και πότε δύο συγγραφείς θα ξανασυνεργασθούν σε ένα δίκτυο συν-συγγραφής (co-authorship network). Οι Huang και Lin ([HL09]), αντιμετωπίζοντας αυτό το πρόβλημα, πρότειναν μία προσέγγιση που λαμβάνει υπόψη τη χρονική εξέλιξη των εμφανίσεων συνδέσμων. Ένας άλλος τρόπος για την χρησιμοποίηση χρονικής πληροφορίας προκύπτει από το γεγονός ότι παλαιότερα γεγονότα είναι λιγότερο πιθανό να επηρεάσουν την εμφάνιση μελλοντικών συνδέσμων από ό,τι πρόσφατα γεγονότα. Για παράδειγμα, τα ενδιαφέροντα ενός συγγραφέα ίσως αλλάζουν με το χρόνο και, επομένως, παλιές δημοσιεύσεις ίσως είναι λιγότερο σχετικές με την τρέχουσα περιοχή έρευνάς του. Οι Tylenda κ. ά. ([TAB09]) ανέπτυξαν μία μέθοδο που ενσωματώνει χρονική πληροφορία σε εξελισσόμενα δίκτυα. Διαπίστωσαν ότι η επίδοση μπορεί να βελτιωθεί με απόδοση βαρών στις ακμές είτε με βάση το χρόνο (δίνοντας μικρότερα βάρη στα παλαιότερα γεγονότα ή αγνοώντας τα) ή σύμφωνα με τη δύναμη της σύνδεσης.

3.7.3 Πολυδιάστατα δίκτυα

Στα πολυδιάστατα δίκτυα οι ακμές ή οι κόμβοι είναι διαφόρων τύπων. Για παράδειγμα, ένα κοινωνικό δίκτυο μπορεί να αποτελείται από θετικούς και αρνητικούς συνδέσμους οι οποίοι κατευθύνονται αντίστοιχα προς φίλους και εχθρούς ή έμπιστους και μη έμπιστους ομοτίμους. Οι Leskovec κ. ά. ([LHK10]) πρότειναν μία μέθοδο για την πρόβλεψη του προσήμου των συνδέσμων, ωστόσο το πρόβλημα της πρόβλεψης τόσο της ύπαρξης ενός συνδέσμου όσο και του προσήμου δεν έχει μελετηθεί επαρκώς.

Ένας πιο περίπλοκος τύπος πολυδιάστατων δικτύων περιλαμβάνει τα δίκτυα που αποτελούνται από διάφορες κατηγορίες κόμβων. Για παράδειγμα, ένα online σύστημα διαμοιρασμού πόρων (resource-sharing), όπως το Del.icio.us, μπορεί να παρασταθεί ως ένα δίκτυο που αποτελείται από τρία είδη κόμβων: χρήστες, URLs και ετικέτες (tags). Σε αντίθεση με τους τριμερείς γράφους, κόμβοι της ίδιας κατηγορίας μπορεί να συνδέονται, όπως ένας χρήστης με τον ακόλουθό του. Αγνοώντας τις συνδέσεις εντός των κατηγοριών, η πρόβλεψη των συνδέσμων μεταξύ χρηστών και αντικειμένων έχει ήδη μελετηθεί ([ZZZ10]). Όμως, δεν υπάρχει μελέτη που να λαμβάνει υπόψη τους συνδέσμους τόσο εντός όσο και μεταξύ των κατηγοριών.

3.7.4 Εξωτερική (μη δομική) πληροφορία

Η επίδοση των αλγορίθμων μπορεί να βελτιωθεί λαμβάνοντας υπόψη εξωτερική πληροφορία, όπως γνωρίσματα των κόμβων ([Lin98]). Η κοινή λογική υπαγορεύει την ιδέα ότι δύο άνθρωποι που έχουν περισσότερα κοινά χαρακτηριστικά (όπως ηλικία, φύλο ή επάγγελμα) θα έχουν και περισσότερα κοινά ενδιαφέροντα και προτιμήσεις (και, επομένως, μεγαλύτερη πιθανότητα να συνδέονται σε ένα κοινωνικό δίκτυο). Η πληροφορία των γνωρισμάτων μπορεί να χρησιμοποιηθεί για την πρόβλεψη συνδέσμου χωρίς να ληφθεί υπόψη η δομή του δικτύου. Έτσι, όταν οι υπάρχοντες σύνδεσμοι είναι αναξιόπιστοι, οι μέθοδοι που βασίζονται σε γνωρίσματα είναι προτιμότερες και μπορούν να λύσουν το λεγόμενο πρόβλημα της ψυχρής έναρξης (cold start problem).

4

Ανίχνευση κοινοτήτων

4.1 Εισαγωγή

Η σύγχρονη επιστήμη των δικτύων έχει βελτιώσει σημαντικά την κατανόηση των πολύπλοκων συστημάτων (complex systems). Ένα από τα πιο σημαντικά χαρακτηριστικά των γράφων που παριστάνουν πραγματικά συστήματα είναι η κοινοτική δομή (community structure) ή ομαδοποίηση (clustering), δηλαδή η οργάνωση των κορυφών σε ομάδες (clusters) ώστε πολλές ακμές να ενώνουν κορυφές της ίδιας ομάδας και συγκριτικά λίγες ακμές να ενώνουν κορυφές διαφορετικών ομάδων. Αυτές οι ομάδες ή κοινότητες μπορούν να θεωρηθούν σε κάποιο βαθμό ως ανεξάρτητα μέρη του γράφου, τα οποία παίζουν ένα ρόλο παρόμοιο, για παράδειγμα, με αυτόν των ιστών ή των οργάνων στο ανθρώπινο σώμα. Η ανίχνευση κοινοτήτων ([For10], [Agg11]) έχει μεγάλη σημασία στην κοινωνιολογία, τη βιολογία και την πληροφορική, τομείς όπου τα συστήματα συχνά παριστάνονται ως γράφοι. Το πρόβλημα της ανίχνευσης κοινοτήτων είναι πολύ δύσκολο και δεν έχει βρεθεί ακόμη ικανοποιητική λύση, παρά την πολύ μεγάλη προσπάθεια μίας μεγάλης διεπιστημονικής κοινότητας που εργάζεται για τη λύση του τα τελευταία χρόνια.

4.2 Στοιχεία ανίχνευσης κοινοτήτων

Το πρόβλημα της ομαδοποίησης γράφων (graph clustering), διαισθητικό εκ πρώτης όψεως, στην πραγματικότητα δεν είναι καλώς ορισμένο. Τα κύρια στοιχεία του, δηλαδή οι έννοιες της κοινότητας και της διαμέρισης, δεν είναι αυστηρά ορισμένα.

Είναι σημαντικό να τονισθεί ότι ο προσδιορισμός των ομάδων είναι δυνατός μόνο αν ο εξεταζόμενος γράφος είναι αραιός, δηλαδή αν το πλήθος των ακμών είναι της τάξης του πλήθους των κορυφών του γράφου. Στην αντίθετη περίπτωση, οι κοινότητες δεν έχουν νόημα και το πρόβλημα μετατρέπεται σε κάτι κάπως διαφορετικό, παρόμοιο με το πρόβλημα της ομαδοποίησης δεδομένων (data clustering), το οποίο απαιτεί έννοιες και μεθόδους διαφορετικής φύσης.

4.2.1 Η έννοια της κοινότητας

4.2.1.1 Βασικά

Το πρώτο πρόβλημα στην ομαδοποίηση γράφων (graph clustering) είναι η αναζήτηση ενός ποσοτικού ορισμού της έννοιας της κοινότητας. Κανένας ορισμός δεν είναι καθολικά αποδεκτός. Ο ορισμός της έννοιας της κοινότητας συχνά εξαρτάται από το συγκεκριμένο σύστημα και την εφαρμογή. Η διαίσθηση υποβάλλει την ιδέα ότι πρέπει να υπάρχουν περισσότερες ακμές εντός της κοινότητας από ό,τι μεταξύ κορυφών της κοινότητας και του υπόλοιπου γράφου. Αυτή είναι η κατευθυντήρια γραμμή στη βάση των περισσότερων ορισμών. Ωστόσο, στις περισσότερες περιπτώσεις, οι κοινότητες ορίζονται αλγοριθμικά, δηλαδή είναι απλώς η έξοδος του αλγορίθμου, χωρίς να υπάρχει ένας ακριβής εκ των προτέρων ορισμός.

Ας θεωρήσουμε έναν υπογράφο C ενός γράφου G, με $|C| = n_c$ και $|G| = n$ κορυφές, αντίστοιχα. Ορίζουμε τον εσωτερικό και εξωτερικό βαθμό 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. Αν $k_v^{ext} = 0$, η κορυφή έχει γείτονες μόνο μέσα στο C, ο οποίος είναι πιθανώς μία 'καλή' ομάδα για τη v. Αν $k_v^{int} = 0$, η κορυφή θα έπρεπε καλύτερα να εκχωρηθεί σε διαφορετική ομάδα. Ο εσωτερικός βαθμός k_{int}^C του C ορίζεται ως το άθροισμα των εσωτερικών βαθμών των κορυφών του. Ομοίως, ο εξωτερικός βαθμός k_{ext}^C του C ορίζεται ως το άθροισμα των εξωτερικών βαθμών των κορυφών του. Ο συνολικός βαθμός k^C του C ορίζεται ως το άθροισμα των βαθμών των κορυφών του C. Προφανώς, $k^C = k_{int}^C + k_{ext}^C$.

Ορίζουμε την ενδοομαδική 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:

$$\delta_{int}(C) = \frac{\#\,\text{εσωτερικών ακμών του } C}{n_C(n_C - 1)/2} \quad (4.1).$$

Ομοίως, η διομαδική 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:

$$\delta_{ext}(C) = \frac{\#\,\text{δια-ομαδικών ακμών του } C}{n_C(n - n_C)} \quad (4.2).$$

Για να είναι ο C μία κοινότητα, αναμένουμε η $\delta_{int}(C)$ 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 $n(n\text{-}1)/2$. Από την άλλη πλευρά, η $\delta_{ext}(C)$ πρέπει να είναι πολύ μικρότερη από τη $\delta(G)$. Η αναζήτηση της καλύτερης αντιστάθμισης μεταξύ μίας μεγάλης $\delta_{int}(C)$ και μίας μικρής $\delta_{ext}(C)$ είναι ρητά ή σιωπηρά ο στόχος των περισσότερων αλγορίθμων ομαδοποίησης.

Μία ιδιότητα που απαιτείται να έχει μία κοινότητα είναι η συνεκτικότητα: για να είναι ο C μία κοινότητα πρέπει να υπάρχει για κάθε ζεύγος κορυφών του ένα μονοπάτι που τις συνδέει και περνά μόνο από κορυφές του C.

Παρακάτω θα εισαγάγουμε τους κύριους ορισμούς της έννοιας της κοινότητας. Διάφοροι ορισμοί έχουν δοθεί, από αναλυτές κοινωνικών δικτύων, επιστήμονες πληροφορικής και φυσικούς. Διακρίνουμε τρεις κατηγορίες ορισμών: τοπικούς, καθολικούς και βασισμένους στην ομοιότητα κορυφών.

4.2.1.2 Τοπικοί ορισμοί

Οι τοπικοί ορισμοί εστιάζουν στον υπό μελέτη υπογράφο και, πιθανώς, την άμεση γειτονιά του, αλλά αγνοούν τον υπόλοιπο γράφο. Παραθέτουμε τους κύριους ορισμούς που έχουν υιοθετηθεί στην ανάλυση κοινωνικών δικτύων, ακολυθώντας τους Wasserman και Faust ([WF94]), οι οποίοι καθορίζουν τέσσερα κριτήρια: την πλήρη αμοιβαιότητα (complete mutuality), την προσβασιμότητα (reachability), το βαθμό κορυφής και τη σύγκριση μεταξύ εσωτερικής και εξωτερικής συνοχής (cohesion).

Οι κοινότητες μπορούν να ορισθούν ως ομάδες καθένα από τα μέλη των οποίων είναι ‘φίλος’ με όλα τα υπόλοιπα (πλήρης αμοιβαιότητα) ([LP49]). Σε γραφοθεωρητικούς όρους, αυτό αντιστοιχεί σε μία κλίκα, δηλαδή ένα υποσύνολο του οποίου όλες οι κορυφές είναι γειτονικές η μία με την άλλη. Στην ανάλυση κοινωνικών δικτύων, μία κλίκα είναι ένας μέγιστος

υπογράφος, ενώ στη θεωρία γραφημάτων συνήθως κλίκες καλούνται επίσης και μη μέγιστοι υπογράφοι. Τα τρίγωνα είναι οι απλούστερες κλίκες, και είναι συχνά στα πραγματικά δίκτυα. Όμως, μεγαλύτερες κλίκες είναι λιγότερο συχνές.

Μία *n*-κλίκα (*n*-clique) είναι ένας μέγιστος υπογράφος τέτοιος ώστε η απόσταση κάθε ζεύγους κορυφών του δεν υπερβαίνει το *n* ([Alb73], [Luc50]). Μία *n*-οικογένεια (*n*-clan) είναι μία *n*-κλίκα της οποίας η διάμετρος δεν υπερβαίνει το *n* ([Mok79]). Επίσης, έχει προταθεί η έννοια του μέγιστου υπογράφου διαμέτρου *n* (*n*-club) ([Mok79]).

Ακόμη, έχει χρησιμοποιηθεί και η έννοια του μέγιστου υπογράφου στον οποίο κάθε κορυφή είναι γειτονική με όλες τις άλλες κορυφές του υπογράφου εκτός το πολύ *k* από αυτές (*k*-plex) ([SF78]). Παρομοίως, ένας *k*-πυρήνας (*k*-core) είναι ένας μέγιστος υπογράφος στον οποίο κάθε κορυφή είναι γειτονική με τουλάχιστον *k* άλλες κορυφές του υπογράφου ([Sei83]).

Ένα *LS*-σύνολο ([LS69]) ή ισχυρή κοινότητα (strong community) ([RCC+04]) είναι ένας υπογράφος ο εσωτερικός βαθμός κάθε κορυφής του οποίου είναι μεγαλύτερος από τον εξωτερικό της βαθμό. Μία ασθενής κοινότητα (weak community) είναι ένας υπογράφος του οποίου ο εσωτερικός βαθμός υπερβαίνει τον εξωτερικό βαθμό. Ένα *LS*-σύνολο είναι επίσης ασθενής κοινότητα, ενώ το αντίστροφο δεν ισχύει πάντα. Οι Hu κ.ά. ([HCZ+08]) εισήγαγαν εναλλακτικούς ορισμούς των ισχυρών και ασθενών κοινοτήτων: μία κοινότητα είναι ισχυρή αν ο εσωτερικός βαθμός κάθε κορυφής της υπερβαίνει το πλήθος των ακμών που τη συνδέουν με κάθε άλλη κοινότητα, ενώ μία κοινότητα είναι ασθενής αν ο συνολικός εσωτερικός της βαθμός υπερβαίνει το πλήθος των ακμών που τη συνδέουν με κάθε άλλη κοινότητα. Ένα *LS*-σύνολο είναι επίσης ισχυρή κοινότητα κατά την έννοια των Hu κ.ά.. Ομοίως, μία ασθενής κοινότητα κατά Radicchi κ.ά. είναι ασθενής και κατά Hu κ.ά. Και στις δύο περιπτώσεις το αντίστροφο δεν ισχύει. Ένας άλλος ορισμός χρησιμοποιεί την έννοια της ακμικής συνδετικότητας (edge connectivity). Η ακμική συνδετικότητα ενός ζεύγους κορυφών ενός γράφου είναι ο ελάχιστος αριθμός ακμών που πρέπει να απομακρυνθούν για να αποσυνδεθούν οι δύο κορυφές, δηλαδή να μην υπάρχει μονοπάτι που να τις συνδέει. Ένα λάμδα σύνολο είναι ένας υπογράφος κάθε ζεύγος κορυφών του οποίου έχει μεγαλύτερη ακμική συνδετικότητα από κάθε ζεύγος κορυφών που αποτελείται από μία κορυφή του υπογράφου και μία κορυφή από τον υπόλοιπο γράφο ([BES90]).

4.2.1.3 Καθολικοί ορισμοί

Έχουν προταθεί πολλά καθολικά κριτήρια για τον προσδιορισμό κοινοτήτων. Στις περισσότερες περιπτώσεις είναι πλάγιοι (indirect) ορισμοί, όπου κάποια καθολική ιδιότητα του γράφου χρησιμοποιείται σε έναν αλγόριθμο που στο τέλος δίνει κοινότητες. Ωστόσο, υπάρχουν και ευθείς (proper) ορισμοί, οι οποίοι βασίζονται στην ιδέα ότι ένας γράφος έχει κοινοτική δομή αν διαφέρει από έναν τυχαίο γράφο. Ένας τυχαίος γράφος κατά Erdös-Rényi,

για παράδειγμα, δεν αναμένεται να έχει κοινοτική δομή, καθώς οποιεσδήποτε δύο κορυφές του έχουν την ίδια πιθανότητα να είναι γειτονικές. Επομένως, μπορεί κανείς να ορίσει ένα μηδενικό μοντέλο (null model), δηλαδή ένα γράφο που ταιριάζει με τον αρχικό σε κάποια από τα δομικά του χαρακτηριστικά αλλά κατά τα άλλα είναι τυχαίος γράφος. Το μηδενικό μοντέλο χρησιμοποιείται για να επιβεβαιωθεί αν ο υπό μελέτη γράφος παρουσιάζει κοινοτική δομή ή όχι. Το δημοφιλέστερο μηδενικό μοντέλο είναι αυτό που πρότειναν οι Newman και Girvan, το οποίο αποτελείται από μία τυχαιοποιημένη (randomized) εκδοχή του αρχικού γράφου, όπου οι ακμές ξανατοποθετούνται, υπό τον περιορισμό ότι ο αναμενόμενος βαθμός κάθε κορυφής ταιριάζει με το βαθμό της στον αρχικό γράφο ([NG04]). Αυτό το μηδενικό μοντέλο είναι η βασική έννοια πίσω από τον ορισμό της τμηματικότητας (modularity), μίας συνάρτησης η οποία αξιολογεί την ποιότητα (goodness) των διαμερίσεων του γράφου σε ομάδες. Στην καθιερωμένη διατύπωση της τμηματικότητας, ένας υπογράφος είναι μία κοινότητα αν το πλήθος των ακμών εντός του υπογράφου υπερβαίνει το αναμενόμενο πλήθος των εσωτερικών ακμών του υπογράφου στο μηδενικό μοντέλο. Αυτό το αναμενόμενο πλήθος είναι ένας μέσος επί όλων των δυνατών πραγματώσεων (realizations) του μηδενικού μοντέλου. Έχουν προταθεί διάφορες τροποποιήσεις της τμηματικότητας.

4.2.1.4 Ορισμοί βασισμένοι στην ομοιότητα κορυφών

Είναι φυσικό να υποθέσει κανείς ότι οι κοινότητες είναι ομάδες παρόμοιων κορυφών. Η ομοιότητα ενός ζεύγους κορυφών μπορεί να υπολογισθεί ως προς κάποια ιδιότητα αναφοράς, τοπική ή καθολική, αδιάφορο αν συνδέονται με ακμή ή όχι. Κάθε κορυφή καταλήγει στην ομάδα της οποίας οι κορυφές είναι πιο παρόμοιες με αυτήν. Τα μέτρα ομοιότητας βρίσκονται στη βάση των παραδοσιακών μεθόδων ανίχνευσης κοινοτήτων. Εδώ αναφέρονται κάποια δημοφιλή τέτοια μέτρα.

Αν είναι δυνατό να ενσωματώσουμε (embed) τις κορυφές του γράφου σε ένα *n*-διάστατο Ευκλείδιο χώρο, μπορεί κανείς να χρησιμοποιήσει την απόσταση μεταξύ δύο κορυφών ως ένα μέτρο της ομοιότητάς τους (για την ακρίβεια, της ανομοιότητάς τους). Δεδομένων δύο σημείων δεδομένων $A = (a_1, a_2, \ldots, a_n)$ και $B = (b_1, b_2, \ldots, b_n)$, μπορεί κανείς να χρησιμοποιήσει οποιαδήποτε νόρμα L_m, όπως την Ευκλείδια απόσταση (L_2-νόρμα),

$$d_{AB}^{E} = \sum_{k=1}^{n} \sqrt{(a_k - b_k)^2} \quad (4.3),$$

η απόσταση Manhattan (L_1-νόρμα),

$$d_{AB}^{M} = \sum_{k=1}^{n} \left| a_k - b_k \right| \quad (4.4),$$

και η L_∞ -νόρμα

$$d_{AB}^{\infty} = \max_{k\in[1,n]} \left| a_k - b_k \right| \text{ (4.5).}$$

Ένα άλλο δημοφιλές μέτρο είναι η ομοιότητα συνημιτόνου, η οποία ορίζεται ως

$$\rho_{AB} = \arccos \frac{\mathrm{a}\cdot\mathrm{b}}{\sqrt{\sum_{k=1}^{n} a_k^2}\sqrt{\sum_{k=1}^{n} b_k^2}} \text{ (4.6),}$$

όπου $\mathrm{a}\cdot\mathrm{b}$ είναι το εσωτερικό γινόμενο των διανυσμάτων a και b. Η μεταβλητή ρ_{AB} παίρνει τιμές στο διάστημα [0, π).

Αν ο γράφος δεν μπορεί να ενσωματωθεί στο χώρο, η ομοιότητα πρέπει κατ' ανάγκη να εξαχθεί από τις σχέσεις γειτνίασης μεταξύ των κορυφών. Μία δυνατότητα είναι να ορισθεί μία απόσταση ([Bur76], [WF94]) μεταξύ κορυφών ως

$$d_{ij} = \sqrt{\sum_{k\neq i,j} (A_{ik} - A_{jk})^2} \text{ (4.7),}$$

όπου **A** είναι ο πίνακας γειτνίασης. Εναλλακτικά, κανείς μπορεί να μετρήσει την επικάλυψη μεταξύ των γειτονιών Γ(*i*) και Γ(*j*) των κορυφών *I* και *j,* η οποία δίνεται από το λόγο του μεγέθους της τομής προς το μέγεθος της ένωσης των γειτονιών, δηλαδή

$$\omega_{ij} = \frac{\left|\Gamma(i)\cap\Gamma(j)\right|}{\left|\Gamma(i)\cup\Gamma(j)\right|} \text{ (4.8).}$$

Ένα άλλο μέτρο είναι η συσχέτιση Pearson μεταξύ στηλών ή γραμμών του πίνακα γειτνίασης,

$$C_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i\sigma_j} \text{ (4.9),}$$

όπου $\mu_i = \left(\sum_j A_{ij}\right)\Big/ n$ (μέσοι) και $\sigma_i = \sqrt{\sum_j (A_{ij} - \mu_i)^2 \Big/ n}$ (διασπορές).

Ένα άλλο μέτρο είναι το πλήθος των ανεξάρτητων ως προς τις ακμές (ή τις κορυφές) (edge- (vertex-) independent) μονοπατιών μεταξύ των δύο κορυφών. Ομοίως, μπορεί κανείς να θεωρήσει όλα τα μονοπάτια μεταξύ των δύο κορυφών. Σε αυτήν την περίπτωση, υπάρχει το πρόβλημα ότι το συνολικό πλήθος των μονοπατιών είναι άπειρο, το οποίο μπορεί να ξεπεραστεί παίρνοντας ένα σταθμισμένο άθροισμα του πλήθους των μονοπατιών. Για παράδειγμα, τα μονοπάτια μήκους *l* μπορούν να σταθμισθούν με παράγοντα a^l, με $a<1$, ή με παράγοντα $1/l!$ ([EH08], [EH09]), ώστε η συνεισφορά των μακριών μονοπατιών να περιορισθεί και το άθροισμα να συγκλίνει.

Μία άλλη σημαντική κατηγορία μέτρων ομοιότητας κορυφών βασίζεται στις ιδιότητες των τυχαίων περιπάτων στους γράφους. Μία από αυτές είναι ο χρόνος επανόδου (commute-time) μεταξύ δύο κορυφών, ο οποίος είναι ο μέσος αριθμός βημάτων που χρειάζονται για να φθάσει ένας τυχαίος περιπατητής (random walker) που ξεκινά από οποιαδήποτε από τις δύο κορυφές

στην άλλη για πρώτη φορά και να επιστρέψει στην αρχική. Ο χρόνος επανόδου συνδέεται στενά ([CRRS89]) με την απόσταση αντίστασης (resistance distance), που εισήχθη από τους Klein και Randic ([KR93]) και εκφράζει την ηλεκτρική αντίσταση μεταξύ των δύο κορυφών αν ο γράφος θεωρηθεί ως δίκτυο αντιστάσεων. Οι White και Smyth ([WS03]) και ο Zhou ([Zho03]) χρησιμοποίησαν το μέσο χρόνο πρώτης διέλευσης (first passage time), δηλαδή το μέσο αριθμό βημάτων που χρειάζονται για να φθάσουμε για πρώτη φορά στην κορυφή-στόχο από την αφετηρία. Οι Harel και Koren ([HK01]) πρότειναν τη χρήση μέτρων που βασίζονται σε ποσότητες όπως η πιθανότητα επίσκεψης μίας κορυφής-στόχου σε όχι περισσότερο από ένα δεδομένο αριθμό βημάτων ξεκινώντας από μία κορυφή-αφετηρία ή η πιθανότητα ένας τυχαίος περιπατητής που ξεκινά από μία κορυφή-αφετηρία να επισκεφθεί ακριβώς μία φορά την κορυφή-στόχο πριν επιστρέψει στην αφετηρία. Μία άλλη ποσότητα που χρησιμοποιείται για τον ορισμό μέτρων ομοιότητας είναι η πιθανότητα διαφυγής (escape probability), η οποία ορίζεται ως η πιθανότητα ο τυχαίος περιπατητής να φθάσει την κορυφή-στόχο πριν επιστρέψει στην αφετηρία ([PF03], [TFP08]). Η πιθανότητα διαφυγής σχετίζεται με την ενεργό αγωγιμότητα (effective conductance) μεταξύ των δύο κορυφών στο ισοδύναμο δίκτυο αντιστάσεων. Άλλοι συγγραφείς αξιοποίησαν ιδιότητες τροποποιημένων τυχαίων περιπάτων.

4.2.2 Η έννοια της διαμέρισης

4.2.2.1 Βασικά

Μία διαμέριση είναι μία διαίρεση ενός γράφου σε ομάδες, έτσι ώστε κάθε κορυφή να ανήκει σε μία ομάδα. Στα πραγματικά συστήματα μία κορυφή μπορεί να ανήκει σε διαφορετικές κοινότητες. Μία διαίρεση του γράφου σε επικαλυπτόμενες (ή ασαφείς) κοινότητες ονομάζεται κάλυμμα (cover).

Το πλήθος των δυνατών διαμερίσεων ενός γράφου αυξάνεται ταχύτερα από εκθετικά με το μέγεθος του γράφου. Αυτό σημαίνει ότι η απαρίθμηση ή η αξιολόγηση όλων των διαμερίσεων ενός γράφου δεν είναι δυνατή παρά μόνο για πολύ μικρό μέγεθος του γράφου.

Οι διαμερίσεις μπορεί να είναι ιεραρχικά διατεταγμένες, αν ο γράφος έχει διαφορετικά επίπεδα οργάνωσης/δομής σε διαφορετικές κλίμακες. Σε αυτήν την περίπτωση, οι ομάδες παρουσιάζουν και οι ίδιες κοινοτική δομή, περιέχοντας μικρότερες κοινότητες, οι οποίες με τη σειρά τους μπορεί να περιέχουν μικρότερες κοινότητες κ.ο.κ. Η ιεραρχική οργάνωση είναι σύνηθες χαρακτηριστικό πολλών πραγματικών δικτύων. Ένας φυσικός τρόπος αναπαράστασης της ιεραρχικής δομής ενός γράφου είναι η σχεδίαση ενός δενδρογράμματος .

4.2.2.2 Συναρτήσεις ποιότητας: τμηματικότητα

Πολλοί αλγόριθμοι καθορίζουν ένα σύνολο διαμερίσεων που έχουν νόημα, ιδανικά μία ή λίγες, ενώ κάποιοι άλλοι δίνουν ένα μεγάλο αριθμό διαμερίσεων. Αυτό δε σημαίνει ότι οι διαμερίσεις που βρίσκουν είναι εξίσου 'καλές'. Επομένως, είναι χρήσιμο να έχουμε ένα ποσοτικό κριτήριο για να εκτιμάμε την ποιότητα (goodness) μίας διαμέρισης. Μία συνάρτηση ποιότητας (quality function) είναι μία συνάρτηση που αντιστοιχίζει έναν αριθμό σε κάθε διαμέριση ενός γράφου. Οι διαμερίσεις με υψηλή βαθμολογία είναι 'καλές', οπότε αυτή με την υψηλότερη είναι η καλύτερη. Ωστόσο, πρέπει να έχει κανείς κατά νου ότι το ερώτημα αν μία διαμέριση είναι καλύτερη από μία άλλη είναι κακώς τεθειμένο και ότι η απάντηση εξαρτάται από τη συγκεκριμένη έννοια της κοινότητας και τη χρησιμοποιούμενη συνάρτηση ποιότητας.

Ένα παράδειγμα συνάρτησης ποιότητας είναι η επίδοση (performance) *P*, η οποία μετρά το πλήθος των ορθώς 'ερμηνευμένων' ζευγών κορυφών, δηλαδή των ζευγών κορυφών που ανήκουν στην ίδια κοινότητα και συνδέονται με ακμή και των ζευγών κορυφών που ανήκουν σε διαφορετικές κοινότητες και δε συνδέονται με ακμή. Ο ορισμός της επίδοσης, για μία διαμέριση $\mathcal{P}$, είναι

$$P(\mathcal{P}) = \frac{\left|\{(i,j) \in E, C_i = C_j\}\right| + \left|\{(i,j) \notin E, C_i \neq C_j\}\right|}{n(n-1)/2} \text{ (4.10).}$$

Ένα άλλο παράδειγμα είναι η κάλυψη (coverage), δηλαδή ο λόγος του πλήθους των ενδοκοινοτικών ακμών προς το συνολικό πλήθος των ακμών.

Η πιο δημοφιλής συνάρτηση ποιότητας είναι η τμηματικότητα (modularity) των Newman και Girvan ([NG04]). Βασίζεται στην ιδέα ότι ένας τυχαίος γράφος δεν αναμένεται να έχει κοινοτική δομή, οπότε η πιθανή ύπαρξη ομάδων αποκαλύπτεται με τη σύγκριση μεταξύ της πραγματικής πυκνότητας των ακμών σε έναν υπογράφο και της αναμενόμενης πυκνότητας ακμών στον υπογράφο αν οι κορυφές συνδέονταν ασχέτως κοινοτικής δομής. Αυτή η αναμενόμενη πυκνότητα εξαρτάται από το επιλεγμένο μηδενικό μοντέλο, δηλαδή το αντίγραφο του αρχικού γράφου που διατηρεί κάποιες από τις δομικές του ιδιότητες αλλά στερείται κοινοτικής δομής. Το καθιερωμένο μηδενικό μοντέλο της τμηματικότητας επιβάλλει η αναμενόμενη ακολουθία βαθμών να ταιριάζει με την πραγματική ακολουθία βαθμών του γράφου. Η τμηματικότητα γράφεται

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j) \text{ (4.11),}$$

όπου το άθροισμα διατρέχει όλα τα ζεύγη κορυφών, *A* ο πίνακας γειτνίασης του γράφου, *m* ο συνολικός αριθμός των ακμών του γράφου, k_i ο βαθμός της κορυφής *I* και η συνάρτηση δ

παίρνει την τιμή 1 αν οι κορυφές *I* και *j* ανήκουν στην ίδια κοινότητα (C_i=C_j) και την τιμή 0 διαφορετικά. Η τμηματικότητα μπορεί να γραφεί και ως

$$Q = \sum_{c=1}^{n_c} \left[\frac{l_c}{m} - \left(\frac{d_c}{2m} \right)^2 \right] \quad (4.12),$$

όπου n_c ο αριθμός των ομάδων, l_c ο συνολικός αριθμός των ακμών που συνδέουν κορυφές της ομάδας c και d_c το άθροισμα των βαθμών των κορυφών της ομάδας c. Στην τελευταία εξίσωση ο πρώτος όρος κάθε προσθετέου είναι το κλάσμα των ακμών του γράφου εντός της ομάδας, ενώ ο δεύτερος παριστάνει το αναμενόμενο κλάσμα των ακμών εντός της ομάδας αν ο γράφος ήταν τυχαίος, με τον ίδιο αναμενόμενο βαθμό για κάθε κορυφή. Επομένως, ένας υπογράφος θεωρείται κοινότητα αν η συνεισφορά του στην τμηματικότητα είναι θετική. Μεγάλες θετικές τιμές της τμηματικότητας εν γένει δείχνουν 'καλές' διαμερίσεις.

4.3 Παραδοσιακές μέθοδοι

4.3.1 Διαμερισμός γράφου

Το πρόβλημα του διαμερισμού γράφου συνίσταται στο διαχωρισμό του συνόλου των κορυφών σε *g* ομάδες προκαθορισμένου μεγέθους, έτσι ώστε ο αριθμός των ακμών μεταξύ των ομάδων να είναι ελάχιστος. Ο αριθμός των ακμών μεταξύ των ομάδων ονομάζεται μέγεθος της τομής (cut size). Πολλοί αλγόριθμοι που έχουν προταθεί διχοτομούν το γράφο. Διαμερίσεις που αποτελούνται από περισσότερες από δύο ομάδες λαμβάνονται συνήθως με επαναληπτική διχοτόμηση. Επιπλέον, στις περισσότερες περιπτώσεις επιβάλλεται ο περιορισμός οι ομάδες να έχουν ίσα μεγέθη. Το πρόβλημα αυτό καλείται ελάχιστη διχοτόμηση (minimal bisection).

Οι αλγόριθμοι διαμερισμού γράφου δεν είναι 'καλοί' για ανίχνευση κοινοτήτων γιατί απαιτούν ως είσοδο τον αριθμό των ομάδων και, σε κάποιες περιπτώσεις, τα μεγέθη τους, για τα οποία γενικά δε γνωρίζουμε τίποτα. Αντ' αυτού, θα ήταν επιθυμητός ένας αλγόριθμος ο οποίος θα παρήγε αυτήν την πληροφορία σαν έξοδο. Πέραν αυτού, η επαναληπτική διχοτόμηση για το διαχωρισμό του γράφου σε περισσότερα μέρη δεν είναι μία αξιόπιστη διαδικασία.

4.3.1.1 Ο αλγόριθμος των Kernighan-Lin

Ο αλγόριθμος των Kernighan-Lin ([KL70]) είναι μία από τις πρώτες μεθόδους που έχουν προταθεί, και χρησιμοποιείται ακόμη συχνά, πολλές φορές σε συνδυασμό με άλλες τεχνικές. Πρόκειται για μία διαδικασία βελτιστοποίησης μίας συνάρτησης κέρδους (benefit function) *Q*, η οποία παριστάνει τη διαφορά μεταξύ του πλήθους των ακμών εντός των ομάδων και του

πλήθους των ακμών μεταξύ τους. Η αφετηρία είναι μία αρχική διαμέριση του γράφου σε δύο ομάδες του προκαθορισμένου μεγέθους. Στη συνέχεια, ισομεγέθη υποσύνολα κορυφών ανταλλάσσονται μεταξύ των δύο ομάδων, έτσι ώστε να επιτυγχάνεται η μέγιστη αύξηση της Q. Για να μειωθεί ο κίνδυνος να παγιδευθεί σε τοπικά μέγιστα της Q, η διαδικασία περιλαμβάνει και κάποιες ανταλλαγές που μειώνουν την Q. Μετά μία σειρά ανταλλαγών, η διαμέριση με τη μεγαλύτερη τιμή της Q επιλέγεται και χρησιμοποιείται ως αφετηρία για μία νέα σειρά επαναλήψεων. Οι διαμερίσεις που βρίσκει η διαδικασία εξαρτώνται ισχυρά από την αρχική διάταξη (configuration). Είναι προτιμότερο να ξεκινάμε με μία καλή εκτίμηση της αναζητούμενης διαμέρισης, διαφορετικά τα αποτελέσματα είναι μάλλον φτωχά. Επομένως, η μέθοδος χρησιμοποιείται συνήθως για να βελτιώσει τις διαμερίσεις που βρίσκονται με άλλες τεχνικές, χρησιμοποιώντας τις ως αρχικές διατάξεις του αλγορίθμου. Ο αλγόριθμος των Kernighan-Lin έχει επεκταθεί ώστε να εξάγει διαμερίσεις με οποιοδήποτε αριθμό ομάδων ([SK88]).

4.3.2 Ιεραρχική ομαδοποίηση

Οι αλγόριθμοι ιεραρχικής ομαδοποίησης αποκαλύπτουν την πολυεπίπεδη δομή του γράφου, την οποία συχνά παρουσιάζουν τα κοινωνικά δίκτυα. Πολλές φορές δίνουν σαν έξοδο ένα δενδρόγραμμα. Μπορούν να διακριθούν σε συσσωρευτικούς (agglomerative), οι οποίοι ξεκινούν θεωρώντας ότι κάθε κόμβος ανήκει σε ξεχωριστή ομάδα και σε κάθε βήμα συγχωνεύουν ομάδες που είναι αρκετά παρόμοιες, και διαιρετικούς (divisive), οι οποίοι ξεκινούν θεωρώντας το σύνολο των κόμβων σαν μία ομάδα και σε κάθε βήμα διαχωρίζουν κάποια ομάδα απομακρύνοντας ακμές που συνδέουν ανόμοιους κόμβους. Είναι προφανές ότι οι αλγόριθμοι των δύο κατηγοριών ακολουθούν αντίθετες κατευθύνσεις. Η αφετηρία των αλγορίθμων ιεραρχικής ομαδοποίησης είναι ο ορισμός ενός μέτρου ομοιότητας μεταξύ των κορυφών. Στη συνέχεια υπολογίζεται η ομοιότητα κάθε ζεύγους κορυφών, είτε συνδέονται είτε όχι, και τελικά λαμβάνεται ένας πίνακας X, ο πίνακας ομοιότητας. Επίσης, στην περίπτωση των συσσωρευτικών αλγορίθμων είναι απαραίτητος ο ορισμός ενός μέτρου που να εκτιμά την ομοιότητα δύο ομάδων, με βάση τον πίνακα X. Υπάρχουν πολλές επιλογές. Στην ομαδοποίηση απλής διασύνδεσης (single linkage clustering) η ομοιότητα μεταξύ δύο ομάδων είναι το ελάχιστο στοιχείο x_{ij}, με το I στη μία ομάδα και το j στην άλλη. Αντίθετα, στην ομαδοποίηση πλήρους διασύνδεσης (complete linkage clustering) χρησιμοποιείται το μέγιστο στοιχείο x_{ij}. Στην ομαδοποίηση μέσης διασύνδεσης (average linkage clustering) υπολογίζεται ο μέσος των x_{ij}. Στους αλγορίθμους ιεραρχικής ομαδοποίησης μερικές φορές επιβάλλονται συνθήκες τερματισμού ώστε να επιλεγεί μία διαμέριση ή ένα σύνολο διαμερίσεων που ικανοποιούν κάποιο κριτήριο, όπως ένα δεδομένο αριθμό ομάδων ή τη βελτιστοποίηση μίας συνάρτησης ποιότητας (π.χ. της τμηματικότητας).

4.3.2.1 Συσσωρευτικοί αλγόριθμοι: ο άπληστος συσσωρευτικός αλγόριθμος βελτιστοποίησης της τμηματικότητας του Newman

Ο Newman ([New04]) πρότεινε έναν άπληστο αλγόριθμο συσσωρευτικής ομαδοποίησης για τη βελτιστοποίηση της τμηματικότητας. Η βασική του ιδέα είναι ότι σε κάθε στάδιο ομάδες κορυφών συγχωνεύονται διαδοχικά ώστε να σχηματίσουν μεγαλύτερες ομάδες, με τρόπο ώστε η τμηματικότητα του δικτύου να αυξάνεται ύστερα από κάθε συγχώνευση.

Αρχικά, κάθε κόμβος θεωρείται ότι ανήκει σε ξεχωριστή ομάδα και σε κάθε στάδιο επιλέγονται οι δύο ομάδες η συγχώνευση των οποίων οδηγεί στη μέγιστη αύξηση της τμηματικότητας. Δε χρειάζεται να ληφθούν υπόψη ζεύγη ομάδων που δε συνδέονται με ακμή, καθώς η συγχώνευση τέτοιων ομάδων δεν οδηγεί σε αύξηση της τμηματικότητας. Χρησιμοποιείται μία δομή δεδομένων η οποία διατηρεί το κλάσμα των ακμών που μοιράζεται κάθε ζεύγος ομάδων στην τρέχουσα διαμέριση. Οι Clauset κ.ά. ([CNM04]) βελτίωσαν την πολυπλοκότητα του αλγορίθμου χρησιμοποιώντας αποδοτικές δομές δεδομένων, όπως σωρούς μεγίστων.

4.3.2.2 Διαιρετικοί αλγόριθμοι: ο αλγόριθμος των Girvan και Newman

Οι διαιρετικές τεχνικές χρησιμοποιήθηκαν σπάνια στο παρελθόν. Τα τελευταία χρόνια, όμως, έχουν γίνει πιο δημοφιλείς.

Ο πιο δημοφιλής διαιρετικός αλγόριθμος είναι ο αλγόριθμος των Girvan και Newman ([GN02], [NG04]). Η μέθοδός τους είναι ιστορικά σημαντική, καθώς σήμανε την αρχή μίας νέας εποχής στην περιοχή της ανίχνευσης κοινοτήτων και ‘άνοιξε’ το πεδίο στους φυσικούς. Οι συγγραφείς χρησιμοποιούν την ιδέα του βαθμού διαμεσολάβησης ακμών (edge betweenness), ο οποίος εκφράζει τη συχνότητα συμμετοχής των ακμών σε μία διεργασία. Τα μέτρα βαθμού διαμεσολάβησης ορίζονται με τρόπο ώστε οι ακμές με υψηλότερο βαθμό διαμεσολάβησης να είναι πιθανότερο να είναι οι ακμές που συνδέουν διαφορετικές κοινότητες. Τα βήματα του αλγορίθμου είναι τα εξής:

1. Υπολογισμός του βαθμού διαμεσολάβησης όλων των ακμών.
2. Απομάκρυνση της ακμής με το μεγαλύτερο βαθμό διαμεσολάβησης (σε περίπτωση ‘ισοπαλίας’, επιλέγεται μία από τις ακμές με το μεγαλύτερο βαθμό διαμεσολάβησης τυχαία).
3. Επαναϋπολογισμός των βαθμών διαμεσολάβησης στον τρέχοντα γράφο.
4. Επανάληψη του κύκλου από το βήμα 2.

Οι συγγραφείς θεώρησαν τρεις εναλλακτικούς ορισμούς του βαθμού διαμεσολάβησης: το γεωδαιτικό βαθμό διαμεσολάβησης (geodesic betweenness), το βαθμό διαμεσολάβησης

τυχαίου περιπάτου (random-walk betweenness) και το βαθμό διαμεσολάβησης ροής ρεύματος (current-flow betweenness).

Ο γεωδαιτικός βαθμός διαμεσολάβησης μίας ακμής είναι ο αριθμός των συντομότερων μονοπατιών μεταξύ όλων των ζευγών κορυφών που περνούν από την ακμή. Για τον υπολογισμό του βαθμού διαμεσολάβησης τυχαίου περιπάτου μίας ακμής θεωρούμε μία κορυφή-αφετηρία, μία κορυφή-στόχο και έναν τυχαίο περιπατητή που ξεκινά από την αφετηρία και κινείται στο γράφο μέχρι να φθάσει στο στόχο, υπολογίζουμε την πιθανότητα να περάσει από την εν λόγω ακμή για κάθε ζεύγος αφετηρίας-στόχου και θεωρούμε το μέσο όρο των πιθανοτήτων. Για τον υπολογισμό του βαθμού διαμεσολάβησης ροής ρεύματος μίας ακμής θεωρούμε το δίκτυο που προκύπτει από το γράφο με αντικατάσταση κάθε ακμής με μία μοναδιαία αντίσταση, επιλέγουμε δύο κορυφές, μεταξύ των οποίων υποθέτουμε ότι εφαρμόζεται τάση, υπολογίζουμε το ρεύμα που διαρρέει την εν λόγω ακμή για κάθε τέτοιο ζεύγος κορυφών και θεωρούμε τη μέση απόλυτη τιμή των ρευμάτων.

Στη μέθοδο των Girvan και Newman η επιλογή του μέτρου βαθμού διαμεσολάβησης δεν είναι τόσο κρίσιμη. Διαφορετικές θεωρήσεις του βαθμού διαμεσολάβησης οδηγούν σε παρόμοιες κοινοτικές δομές.

4.3.3 Φασματική ομαδοποίηση

Οι φασματικοί αλγόριθμοι συγκαταλέγονται στις κλασικές μεθόδους ομαδοποίησης και ανίχνευσης κοινοτήτων. Πρόκειται για αλγορίθμους που εκχωρούν κόμβους σε ομάδες με βάση τα ιδιοδιανύσματα πινάκων όπως ο πίνακας γειτνίασης του δικτύου ή άλλοι σχετικοί πίνακες. Τα ανώτερα k ιδιοδιανύσματα καθορίζουν μία ενσωμάτωση (embedding) των κόμβων του δικτύου ως σημείων ενός k-διάστατου χώρου. Στη συνέχεια μπορεί κανείς να χρησιμοποιήσει κλασικές τεχνικές ομαδοποίησης δεδομένων, όπως ομαδοποίηση K-μέσων, για να εξαγάγει την τελική εκχώρηση των κόμβων σε ομάδες ([Von07]). Η κύρια ιδέα πίσω από τη φασματική ομαδοποίηση είναι ότι η αναπαράσταση χαμηλής διάστασης που επάγεται από τα ανώτερα ιδιοδιανύσματα αποκαλύπτει την κοινοτική δομή του αρχικού γράφου με μεγαλύτερη σαφήνεια.

Ο κύριος πίνακας που χρησιμοποιείται στη φασματική ομαδοποίηση είναι ο Λαπλασιανός πίνακας $\mathcal{L}$. Αν A είναι ο πίνακας γειτνίασης του δικτύου και D ο διαγώνιος πίνακας με τους βαθμούς των κορυφών στη διαγώνιο, ο μη κανονικοποιημένος Λαπλασιανός πίνακας L του δικτύου δίνεται από τη σχέση $L = D - A$. Ο (κανονικοποιημένος) Λαπλασιανός πίνακας $\mathcal{L}$ δίνεται από τη σχέση $\mathcal{L} = D^{-1/2}(D - A)D^{-1/2} = I - D^{-1/2}AD^{-1/2}$. Μπορεί να επαληθευθεί ότι οι πίνακες L και $\mathcal{L}$ είναι συμμετρικοί και θετικά ορισμένοι, επομένως έχουν πραγματικές

και θετικές ιδιοτιμές ([Chu97], [Von07]). Ο Λαπλασιανός πίνακας έχει το 0 ως ιδιοτιμή με πολλαπλότητα ίση με το πλήθος των συνεκτικών συνιστωσών του γράφου.

Το κύριο μειονέκτημα των φασματικών αλγορίθμων είναι η υπολογιστική τους πολυπλοκότητα. Στην πράξη, η φασματική ομαδοποίηση δύσκολα επεκτείνεται σε δίκτυα με εκατοντάδες χιλιάδες κόμβους χωρίς τη χρήση παράλληλων αλγορίθμων.

4.4 Πολυεπίπεδος διαμερισμός γράφου

Οι πολυεπίπεδες μέθοδοι επιτυγχάνουν ποιοτικούς διαμερισμούς χωρίς υψηλό υπολογιστικό κόστος και έχουν χρησιμοποιηθεί για την επίλυση ποικιλίας προβλημάτων ([Ten99]). Η κύρια ιδέα τους είναι η διαδοχική συρρίκνωση του αρχικού γράφου ώστε να ληφθεί ένας μικρός γράφος, ο διαμερισμός αυτού του μικρού γράφου και, τελικά, η διαδοχική προβολή αυτής της διαμέρισης στον αρχικό γράφο, με εκλέπτυνση σε κάθε βήμα. Μεταξύ των μεθόδων πολυεπίπεδου διαμερισμού γράφου αναφέρονται η πολυεπίπεδη φασματική ομαδοποίηση (multilevel spectral clustering) ([BS94]), η Metis ([KK99]), η Graclus ([DGK07]) και η MLR-MCL ([SP09]).

Τα κύρια συστατικά μίας στρατηγικής πολυεπίπεδου διαμερισμού γράφου είναι:

1. Συμπίεση (coarsening). Ο στόχος εδώ είναι η εξαγωγή ενός μικρότερου γράφου παρόμοιου με τον αρχικό. Αυτό το βήμα μπορεί να εφαρμοσθεί επανειλημμένα ώστε να ληφθεί ένας γράφος αρκετά μικρός ώστε να διαμερισθεί γρήγορα και ποιοτικά. Μία δημοφιλής στρατηγική συμπίεσης είναι η κατασκευή αρχικά ενός ταιριάσματος του γράφου (ταίριασμα ονομάζεται ένα σύνολο ακμών τέτοιο ώστε να μην υπάρχουν δύο ακμές του συνόλου που να προσπίπτουν στην ίδια κορυφή). Για κάθε ακμή του ταιριάσματος, οι κορυφές που συνδέει συμπτύσσονται και παριστάνονται ως ένας κόμβος στο συμπιεσμένο γράφο.
2. Αρχικός διαμερισμός. Σε αυτό το βήμα, εκτελείται ένας διαμερισμός του τελικού γράφου που εξήχθη στο βήμα 1. Καθώς ο γράφος αυτός είναι αρκετά μικρός, μπορούν να χρησιμοποιηθούν στρατηγικές όπως ο φασματικός διαμερισμός (spectral partitioning) οι οποίες είναι αργές αλλά δίνουν ποιοτικές διαμερίσεις.
3. Αποσυμπίεση (uncoarsening). Σε αυτό το βήμα, η διαμέριση του γράφου χρησιμοποιείται για να ληφθεί μία αρχική διαμέριση του μεγαλύτερου γράφου. Η λεπτότερη δομή του μεγαλύτερου γράφου χρησιμοποιείται για την εκλέπτυνση της διαμέρισης, συνήθως εκτελώντας τοπική αναζήτηση. Το βήμα αυτό επαναλαμβάνεται μέχρι να φθάσουμε στον αρχικό γράφο εισόδου.

4.5 Μέθοδοι ανίχνευσης επικαλυπτόμενων κοινοτήτων

Στα πραγματικά δίκτυα οι κορυφές συχνά ανήκουν σε διάφορες κοινότητες. Για παράδειγμα, ένα άτομο συνήθως ανήκει σε πολλές κοινωνικές ομάδες, όπως οικογένεια, φίλοι και συνάδελφοι. Επίσης, στα online κοινωνικά δίκτυα ένας χρήστης μπορεί να ανήκει σε οσεσδήποτε κοινότητες. Έτσι, οι αλγόριθμοι ανίχνευσης επικαλυπτόμενων κοινοτήτων, οι οποίοι καθορίζουν ένα σύνολο ομάδων όχι κατ' ανάγκη ξένων, κερδίζουν όλο και μεγαλύτερο ενδιαφέρον ([RG12], [CSLF10]). Για μία επισκόπηση και συγκριτική μελέτη σύγχρονων αλγορίθμων ανίχνευσης επικαλυπτόμενων κοινοτήτων παραπέμπουμε στο [XKS13].

4.5.1 Η μέθοδος διάχυσης κλίκας

Η δημοφιλέστερη τεχνική αυτής της κατηγορίας είναι η μέθοδος διάχυσης κλίκας (clique percolation method – CPM), που πρότειναν οι Palla κ.ά. ([PDFV05]). Βασίζεται στην ιδέα ότι οι εσωτερικές ακμές μίας κοινότητας είναι πιθανό να σχηματίζουν κλίκες λόγω της υψηλής τους πυκνότητας, ενώ είναι απίθανο διακοινοτικές ακμές να σχηματίζουν κλίκες.

Οι συγγραφείς χρησιμοποιούν τον όρο *k*-κλίκα για να δηλώσουν έναν πλήρη γράφο με *k* κορυφές (η έννοια της *k*-κλίκας δεν πρέπει να συγχέεται με την έννοια της *n*-κλίκας). Ονομάζουν γειτονικές (adjacent) δύο *k*-κλίκες οι οποίες έχουν $k-1$ κοινές κορυφές. Η ένωση γειτονικών *k*-κλικών ονομάζεται αλυσίδα *k*-κλικών (*k*-clique chain). Δύο *k*-κλίκες ονομάζονται συνδεδεμένες (connected) αν ανήκουν σε μία αλυσίδα *k*-κλικών. Τέλος, κοινότητα *k*-κλικών (*k*-clique community) ονομάζεται ο μεγαλύτερος συνεκτικός υπογράφος που προκύπτει από την ένωση μίας *k*-κλίκας και όλων των συνδεδεμένων με αυτή *k*-κλικών.

Οι κοινότητες *k*-κλικών μπορεί να επικαλύπτονται. Επίσης, μπορεί να υπάρχουν κορυφές που δεν ανήκουν σε καμμία κοινότητα *k*-κλικών.

Για την εύρεση των κοινοτήτων *k*-κλικών, αρχικά αναζητούνται μέγιστες κλίκες. Στη συνέχεια κατασκευάζεται ένας πίνακας επικάλυψης μεταξύ κλικών **O**, ο οποίος είναι τύπου $n_c \times n_c$, όπου n_c είναι ο αριθμός των κλικών. Το στοιχείο O_{ij} έχει τιμή τον αριθμό των κοινών κορυφών των κλικών *I* και *j*. Έπειτα, διατηρούνται τα στοιχεία του πίνακα **O** που είναι μεγαλύτερα ή ίσα με $k-1$, τα υπόλοιπα μηδενίζονται και βρίσκονται οι συνεκτικές συνιστώσες του πίνακα που προκύπτει.

Μικρές τιμές του *k* (τυπικά μεταξύ 3 και 6) έχει φανεί ότι δίνουν καλά αποτελέσματα.

4.6 Έλεγχος αλγορίθμων ανίχνευσης κοινοτήτων

Όταν σχεδιάζεται ένας αλγόριθμος ομαδοποίησης, είναι αναγκαίο να ελεγχθεί η επίδοσή του και να συγκριθεί με αυτήν άλλων αλγορίθμων. Το ζήτημα του ελέγχου αλγορίθμων δεν έχει μελετηθεί ικανοποιητικά, με αποτέλεσμα να είναι αδύνατο να δηλώσει κανείς ποια μέθοδος είναι η πιο αξιόπιστη στις εφαρμογές και οι τεχνικές ομαδοποίησης γράφων να έχουν πολλαπλασιαστεί τα τελευταία χρόνια.

Εδώ παρουσιάζουμε γράφους αναφοράς με εγγενή κοινοτική δομή, την οποία οι μέθοδοι καλούνται να αναγνωρίσουν, και μέτρα σύγκρισης διαμερίσεων γράφων.

4.6.1 Γράφοι αναφοράς

Ο έλεγχος ενός αλγορίθμου συνίσταται στην εφαρμογή του σε ένα συγκεκριμένο πρόβλημα του οποίου η λύση είναι γνωστή και τη σύγκριση αυτής της λύσης με αυτήν που προτείνει ο αλγόριθμος. Για τον έλεγχο αλγορίθμων ομαδοποίησης γράφων χρησιμοποιούνται γράφοι αναφοράς με σαφή κοινοτική δομή. Αυτοί είναι είτε γράφοι παραγόμενοι από υπολογιστή (computer-generated graphs), όπου η κοινοτική δομή σχεδιάζεται όπως είναι επιθυμητό, είτε πραγματικά δίκτυα, όπου οι κοινότητες είναι καλώς ορισμένες με βάση πληροφορία για το σύστημα.

4.6.1.1 Γράφοι αναφοράς παραγόμενοι από υπολογιστή

Τα τελευταία χρόνια έχει γίνει δημοφιλής μία ειδική κατηγορία γράφων. Οι γράφοι αυτοί παράγονται με το μοντέλο εμφυτευμένης l-διαμέρισης (planted l-partition model) ([CK01]). Το μοντέλο διαμερίζει ένα γράφο με $n = g \cdot l$ κορυφές σε l ομάδες με g κορυφές η καθεμία. Κορυφές της ίδιας ομάδας συνδέονται με πιθανότητα p_{in}, ενώ κορυφές διαφορετικών ομάδων με πιθανότητα p_{out}. Ο μέσος βαθμός μίας κορυφής είναι $\langle k \rangle = p_{in}(g-1) + p_{out}\, g(l-1)$. Αν $p_{in} > p_{out}$ η ενδοομαδική πυκνότητα ακμών υπερβαίνει την διομαδική και ογράφος παρουσιάζει κοινοτική δομή. Οι Girvan και Newman θεώρησαν μία ειδική περίπτωση του μοντέλου ([GN02]). Έθεσαν l = 4, g = 32 (οπότε n = 128) και σταθεροποίησαν το μέσο βαθμό $\langle k \rangle$ στην τιμή 16. Επομένως, $p_{in} + 3p_{out} \approx 1/2$, οπότε οι πιθανότητες p_{in} και p_{out} δεν είναι ανεξάρτητες. Συχνά χρησιμοποιούνται ως παράμετροι τα μεγέθη $z_{in} = p_{in}(g-1) = 31 p_{in}$ και $z_{out} = p_{out}\, g(l-1) = 96 p_{out}$, τα οποία δηλώνουν τον αναμενόμενο εσωτερικό και εξωτερικό βαθμό μίας κορυφής αντίστοιχα. Αυτοί οι γράφοι ήδη αποτελούν καθιερωμένους γράφους αναφοράς.

Ο έλεγχος μίας μεθόδου με το γράφο αναφοράς των Girvan-Newman συνίσταται στον υπολογισμό της ομοιότητας μεταξύ των διαμερίσεων που δίνει η μέθοδος και της φυσικής διαμέρισης του γράφου στις τέσσερις ισομεγέθεις ομάδες. Συνήθως κανείς κτίζει πολλές πραγματώσεις του γράφου για μία συγκεκριμένη τιμή του z_{out} και υπολογίζει τη μέση ομοιότητα μεταξύ των λύσεων της μεθόδου και της φυσικής διαμέρισης. Έπειτα, η διαδικασία επαναλαμβάνεται για διαφορετικές τιμές του z_{out}. Τα αποτελέσματα συνήθως απεικονίζονται σε μία γραφική παράσταση όπου η μέση ομοιότητα σχεδιάζεται ως συνάρτηση του z_{out}. Οι περισσότεροι αλγόριθμοι δίνουν καλά αποτελέσματα για μικρό z_{out} και αρχίζουν να αποτυγχάνουν όταν το z_{out} πλησιάζει την τιμή 8.

Έχουν προταθεί αρκετές τροποποιήσεις του μοντέλου.

4.6.1.2 Πραγματικοί γράφοι αναφοράς

Οι έλεγχοι πάνω σε πραγματικά δίκτυα συνήθως εστιάζουν σε έναν πολύ περιορισμένο αριθμό παραδειγμάτων, για τα οποία υπάρχει ακριβής πληροφορία για τις κορυφές και τις ιδιότητές τους.

Ένα δημοφιλές πραγματικό δίκτυο με γνωστή κοινοτική δομή είναι το κοινωνικό δίκτυο της σχολής καράτε του Zachary ([Zac77]). Αποτελείται από 34 κορυφές, τα μέλη μίας σχολής καράτε στις ΗΠΑ, τα οποία παρατηρήθηκαν για μία περίοδο τριών ετών. Οι ακμές συνδέουν τα άτομα που παρατηρήθηκε να αλληλεπιδρούν εκτός των δραστηριοτήτων της σχολής. Κάποια στιγμή, μία διαμάχη μεταξύ του προέδρου και του εκπαιδευτή της σχολής οδήγησε στη διάσπασή της σε δύο ξεχωριστές ομάδες από τις οποίες η μία υποστήριζε τον πρόεδρο και η άλλη τον εκπαιδευτή. Το ερώτημα είναι αν είναι δυνατό από την τοπολογία του γράφου να προβλεφθεί ο πραγματικός διαχωρισμός σε δύο ομάδες. Το δίκτυο της σχολής καράτε του Zachary είναι μακράν το πιο μελετημένο σύστημα.

Ένα άλλο πραγματικό δίκτυο αναφοράς είναι το δίκτυο των δελφινιών που ζουν στο Doubtful Sound στη Νέα Ζηλανδία που αναλύθηκε από το Lusseau ([Lus03]). Αποτελείται από 62 δελφίνια και ακμές συνδέουν ζώα που εθεάθησαν μαζί συχνότερα από ό,τι αναμενόταν. Μετά την αναχώρηση ενός δελφινιού για κάποιο χρόνο, τα υπόλοιπα χωρίστηκαν σε δύο ομάδες, αρκετά συνεκτικές, με πολλές εσωτερικές κλίκες και εύκολα αναγνωρίσιμες.

Ένα άλλο γνωστό παράδειγμα είναι το δίκτυο των αμερικανικών κολεγιακών ποδοσφαιρικών ομάδων, που προέρχεται από τους Girvan και Newman ([GN02]). Αποτελείται από 115 κορυφές, που παριστάνουν τις ομάδες, και δύο κορυφές συνδέονται αν οι αντίστοιχες ομάδες παίζουν μεταξύ τους. Οι ομάδες χωρίζονται σε 12 ομίλους. Τα παιχνίδια μεταξύ ομάδων του ίδιου ομίλου είναι πιο συχνά από τα παιχνίδια μεταξύ ομάδων διαφορετικών ομίλων, οπότε έχουμε μία φυσική διαμέριση όπου οι κοινότητες αντιστοιχούν στους ομίλους.

4.6.2 Μέτρα σύγκρισης διαμερίσεων

Ας θεωρήσουμε δύο γενικές (generic) διαμερίσεις $x = (X_1, X_2, \ldots, X_{n_X})$ και $y = (Y_1, Y_2, \ldots, Y_{n_Y})$ ενός γράφου *G*, με n_X και n_Y ομάδες αντίστοιχα. Παριστάνουμε με *n* τον αριθμό των κορυφών του γράφου, με n_i^X και n_j^Y τον αριθμό των κορυφών στις ομάδες X_i και Y_j αντίστοιχα και με n_{ij} τον αριθμό των κοινών κορυφών των ομάδων X_i και Y_j.

Οι Girvan και Newman πρότειναν ένα μέτρο που ονομάζεται κλάσμα των ορθώς ταξινομημένων κορυφών (fraction of correctly classified vertices). Μία κορυφή είναι ορθώς ταξινομημένη αν βρίσκεται στην ίδια ομάδα με τουλάχιστον τις μισές κορυφές της κοινότητάς της. Ο αριθμός των ορθώς ταξινομημένων κορυφών διαιρείται με το συνολικό αριθμό κορυφών του γράφου, ώστε προκύπτει ένας αριθμός μεταξύ 0 και 1.

Πέρα από το κλάσμα των ορθώς ταξινομημένων κορυφών, τα περισσότερα μέτρα ομοιότητας μπορούν να χωρισθούν σε τρεις κατηγορίες: μέτρα βασισμένα σε μέτρηση ζευγών, αντιστοίχιση (matching) ομάδων και θεωρία πληροφορίας.

4.6.2.1 Μέτρα βασισμένα στη μέτρηση ζευγών

Τα μέτρα αυτά εξαρτώνται από τον αριθμό των ζευγών κορυφών που ταξινομούνται στις ίδιες (διαφορετικές) ομάδες στις δύο διαμερίσεις. Με a_{11} παριστάνουμε τον αριθμό των ζευγών κορυφών που βρίσκονται στην ίδια κοινότητα και στις δύο διαμερίσεις, με a_{01} (a_{10}) τον αριθμό των ζευγών κορυφών που βρίσκονται στην ίδια κοινότητα στη διαμέριση *x* (*y*) και σε διαφορετικές κοινότητες στη διαμέριση *y* (*x*) και με a_{00} τον αριθμό των ζευγών κορυφών που βρίσκονται σε διαφορετικές κοινότητες και στις δύο διαμερίσεις.

Ο Wallace ([Wal83]) πρότεινε τους δείκτες

$$W_I = \frac{a_{11}}{\sum_k n_k^X (n_k^X - 1)/2} \quad (4.13)$$

και

$$W_{II} = \frac{a_{11}}{\sum_k n_k^Y (n_k^Y - 1)/2} \quad (4.14),$$

οι οποίοι παριστάνουν την πιθανότητα ζεύγη κορυφών στην ίδια ομάδα της διαμέρισης *x* να βρίσκονται επίσης στην ίδια ομάδα της διαμέρισης *Y*, και αντίστροφα. Αυτοί οι δείκτες είναι μη συμμετρικοί. Οι Fowlkes και Mallows ([FM83]) πρότειναν τη χρήση του γεωμετρικού μέσου των παραπάνω δεικτών, ο οποίος είναι συμμετρικός.

Ο δείκτης Rand ([Ran71]) είναι ο λόγος του αριθμού των ορθώς ταξινομημένων ζευγών κορυφών και στις δύο διαμερίσεις (δηλαδή στην ίδια ή σε διαφορετικές ομάδες) προς το συνολικό αριθμό ζευγών

$$R(X,Y) = \frac{a_{11}+a_{00}}{a_{11}+a_{01}+a_{10}+a_{00}} \quad (4.15).$$

Ένα μέτρο ισοδύναμο με το δείκτη Rand είναι η μετρική Mirkin ([Mir96])

$$M(X,Y) = 2(a_{01}+a_{10}) = n(n-1)[1-R(X,Y)] \quad (4.16).$$

Ο δείκτης Jaccard είναι ο λόγος του αριθμού των ζευγών κορυφών που ταξινομούνται στην ίδια ομάδα και στις δύο διαμερίσεις προς τον αριθμό των ζευγών κορυφών που ταξινομούνται στην ίδια ομάδα σε μία τουλάχιστον διαμέριση

$$J(X,Y) = \frac{a_{11}}{a_{11}+a_{01}+a_{10}} \quad (4.17).$$

4.6.2.2 Μέτρα βασισμένα στην αντιστοίχιση ομάδων

Τα μέτρα ομοιότητας που βασίζονται στην αντιστοίχιση (matching) ομάδων στοχεύουν στην εύρεση των μεγαλύτερων επικαλύψεων μεταξύ ζευγών ομάδων διαφορετικών διαμερίσεων.

Το σφάλμα ταξινόμησης ορίζεται ως ([MH01])

$$H(X,Y) = 1 - \frac{1}{n}\max_{\pi}\sum_{k=1}^{n_X} n_{k\pi(k)} \quad (4.18),$$

όπου π είναι μία '1-1' απεικόνιση από τους δείκτες των ομάδων της διαμέρισης Y στους δείκτες των ομάδων της διαμέρισης X. Το μέγιστο λαμβάνεται επί όλων των δυνατών '1-1' απεικονίσεων. Με αυτόν τον τρόπο βρίσκεται η μέγιστη επικάλυψη μεταξύ των ομάδων των δύο διαμερίσεων.

Ένα εναλλακτικό μέτρο είναι η κανονικοποιημένη μετρική Van Dongen ([Van00]), που ορίζεται ως

$$D(X,Y) = 1 - \frac{1}{2n}\left[\sum_{k=1}^{n_X}\max_{k'} n_{kk'} + \sum_{k'=1}^{n_Y}\max_{k} n_{kk'}\right] \quad (4.19).$$

4.6.2.3 Μέτρα βασισμένα στη θεωρία πληροφορίας

Τα μέτρα αυτής της κατηγορίας βασίζονται στην επαναδιατύπωση του προβλήματος της σύγκρισης διαμερίσεων ως ενός προβλήματος αποκωδικοποίησης μηνύματος στο πλαίσιο της

θεωρίας πληροφορίας ([Mac03]). Η ιδέα είναι ότι, αν δύο διαμερίσεις είναι παρόμοιες, κανείς χρειάζεται λίγη πληροφορία για να συμπεράνει τη μία διαμέριση δεδομένης της άλλης. Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί ως ένα μέτρο ανομοιότητας. Για την αποτίμηση του πληροφοριακού περιεχομένου κατά Shannon ([Mac03]) μίας διαμέρισης, αρχικά θεωρούμε τις ταξινομήσεις των κορυφών σε κοινότητες $\{x_i\}$ και $\{y_i\}$, όπου x_i και y_i είναι οι δείκτες των ομάδων της κορυφής I στις διαμερίσεις x και $\mathcal{Y}$ αντίστοιχα. Γίνεται η υπόθεση ότι οι ετικέτες x και y είναι τιμές δύο τυχαίων μεταβλητών X και Y, με από κοινού συνάρτηση κατανομής πιθανότητας $P(x,y) = P(X = x, Y = y) = n_{xy}/n$, το οποίο υποδηλώνει ότι $P(x) = P(X = x) = n_x^X/n$ και $P(y) = P(Y = y) = n_y^Y/n$. Η αμοιβαία πληροφορία (mutual information) $I(X,Y)$ δύο τυχαίων μεταβλητών X και Y ορίζεται ως

$$I(X,Y) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \quad (4.20)$$

και μπορεί να ορισθεί και για τις διαμερίσεις x και $\mathcal{Y}$, καθώς περιγράφονται από τυχαίες μεταβλητές. Ισχύει $I(X,Y) = H(X) - H(X \mid Y)$, όπου $H(X) = -\sum_x P(x) \log P(x)$ είναι η εντροπία Shannon της X και $H(X \mid Y) = -\sum_{x,y} P(x,y) \log P(x \mid y)$ είναι η υπό συνθήκη εντροπία (conditional entropy) της X δεδομένης της Y. Η αμοιβαία πληροφορία δεν είναι ένα ιδανικό μέτρο ομοιότητας, καθώς, δεδομένης μίας διαμέρισης x, όλες οι διαμερίσεις που προκύπτουν από τη x με περαιτέρω διαμερισμό κάποιων από τις ομάδες της έχουν την ίδια αμοιβαία πληροφορία με τη x, ακόμη και αν διαφέρουν πολύ μεταξύ τους. Για την αποφυγή τέτοιων προβλημάτων, οι Danon κ.ά. υιοθέτησαν την κανονικοποιημένη αμοιβαία πληροφορία (normalized mutual information) ([DDDA05])

$$I_{norm}(x,\mathcal{Y}) = \frac{2I(X,Y)}{H(X) + H(Y)} \quad (4.21),$$

η οποία χρησιμοποιείται σήμερα πολύ συχνά σε ελέγχους αλγορίθμων ομαδοποίησης γράφων. Η κανονικοποιημένη αμοιβαία πληροφορία ισούται με 1 αν οι διαμερίσεις ταυτίζονται, ενώ έχει αναμενόμενη τιμή ίση με 0 αν οι διαμερίσεις είναι ανεξάρτητες. Το μέτρο αυτό επεκτάθηκε από τους Lancichinetti κ.ά. στην περίπτωση επικαλυπτόμενων ομάδων.

Η Meilă ([Mei07]) εισήγαγε τη μεταβολή πληροφορίας (variation of information)

$$V(x,\mathcal{Y}) = H(X|Y) + H(Y|X) \quad (4.22),$$

η οποία έχει κάποιες επιθυμητές ιδιότητες σε σχέση με την κανονικοποιημένη αμοιβαία πληροφορία και άλλα μέτρα. Συγκεκριμένα, ορίζει μία μετρική στο χώρο των διαμερίσεων και έχει τις ιδιότητες της απόστασης. Επίσης, είναι ένα τοπικό μέτρο, δηλαδή η ομοιότητα

διαμερίσεων που διαφέρουν μόνο σε ένα μικρό μέρος του γράφου εξαρτάται από τις διαφορές των ομάδων σε αυτήν την περιοχή και όχι από τη διαμέριση του υπόλοιπου του γράφου.

4.7 Κατευθύνσεις έρευνας στην ανίχνευση κοινοτήτων

4.7.1 Δυναμικά δίκτυα

Οι αλγόριθμοι που εξετάσαμε θεωρούν τα κοινωνικά δίκτυα αμετάβλητα. Ωστόσο, τα περισσότερα κοινωνικά δίκτυα, όπως και οι κοινότητες και τα μέλη τους, εξελίσσονται στο χρόνο. Η παρακολούθηση της εξέλιξης της κοινοτικής δομής στο χρόνο είναι πολύ σημαντική για την αποκάλυψη του τρόπου με τον οποίο οι κοινότητες δημιουργούνται και αλληλεπιδρούν. Συνήθως κανείς αναλύει ξεχωριστά στιγμιότυπα σε διαφορετικές χρονικές στιγμές και ελέγχει τι συνέβη τη στιγμή t+1 στις κοινότητες τη στιγμή t. Ίσως θα ήταν καλύτερο να χρησιμοποιηθεί συγχρόνως ολόκληρο το δυναμικό σύνολο δεδομένων.

4.7.2 Ετερογενή δίκτυα

Οι περισσότεροι αλγόριθμοι ανίχνευσης κοινοτήτων υποθέτουν ότι το υπό μελέτη δίκτυο είναι ομοιογενές, δηλαδή αποτελείται από κόμβους και ακμές του ίδιου τύπου. Όμως, πολλά πραγματικά δίκτυα είναι ετερογενή, δηλαδή οι κόμβοι τους ή/και οι ακμές τους είναι διαφόρων τύπων. Αυτή η ποικιλία αφενός προσφέρει μία ευκαιρία, καθώς η αναγνώριση της ετερογένειας του δικτύου ίσως μπορεί να δώσει πολύτιμες πληροφορίες, αφετέρου αποτελεί μία πρόκληση, καθώς δεν είναι ακόμη προφανές πώς πρέπει να χειριστούμε κόμβους και ακμές διαφορετικών τύπων. Η ανίχνευση κοινοτήτων σε ετερογενή δίκτυα εξετάζεται στο [TWL09].

4.7.3 Κατευθυνόμενα δίκτυα – δίκτυα με βάρη

Οι περισσότεροι αλγόριθμοι ασχολούνται με μη κατευθυνόμενους γράφους χωρίς βάρη. Τα πραγματικά δίκτυα, ωστόσο, μπορεί να είναι κατευθυνόμενα ή να έχουν βάρη. Πρόσφατα, κυρίως, αναπτύχθηκαν μέθοδοι που ασχολούνται με τέτοια συστήματα, αφήνοντας όμως χώρο για βελτιώσεις. Για μία συγκριτική επισκόπηση των αλγορίθμων ανίχνευσης κοινοτήτων σε κατευθυνόμενους γράφους παραπέμπουμε στο [MV13].

4.7.4 Μη δομική πληροφορία

Αν και η πληροφορία που προέρχεται από τις κοινωνικές σχέσεις στα κοινωνικά δίκτυα έχει ερευνηθεί εκτεταμένα, ο συνδυασμός της με πληροφορία περιεχομένου που σχετίζεται με τους κόμβους ή τις ακμές των δικτύων για την ανίχνευση κοινοτήτων δεν έχει μελετηθεί

πλήρως. Αυτή η πληροφορία μπορεί να έχει τη μορφή προφίλ χρηστών ή υλικού δημιουργημένου από χρήστες (κείμενο, εικόνες κ.ά.) ή να σχετίζεται και με τις ακμές του δικτύου. Με τη διαθεσιμότητα της πληροφορίας περιεχομένου, αναμένεται οι κοινότητες που θα εξαχθούν να είναι όχι μόνο τοπολογικά πυκνά συνδεδεμένες αλλά και σημασιολογικά συνεπείς. Η αξιοποίηση μη δομικής πληροφορίας για ανίχνευση κοινοτήτων εξετάζεται στα [Yos10] και [ZFW+11].

5

Κοινωνική σύσταση

Τα συστήματα συστάσεων (recommender systems) βοηθούν σημαντικά τους online χρήστες προτείνοντάς τους πληροφορίες που ενδεχομένως τους ενδιαφέρουν. Λόγω της δυνητικής αξίας των κοινωνικών σχέσεων στα συστήματα συστάσεων, η κοινωνική σύσταση (social recommendation) ([THL13]) προσελκύει όλο και περισσότερη προσοχή τα τελευταία χρόνια. Σε αυτό το κεφάλαιο παρουσιάζουμε μία επισκόπηση των υφιστάμενων συστημάτων συστάσεων και κάποιες κατευθύνσεις της σχετικής έρευνας. Αρχικά, δίνουμε τυπικούς ορισμούς της κοινωνικής σύστασης και συζητούμε μία ιδιαιτερότητα της κοινωνικής σύστασης και τις συνέπειές της σε σύγκριση με τα παραδοσιακά συστήματα συστάσεων. Στη συνέχεια, ταξινομούμε τα υφιστάμενα κοινωνικά συστήματα συστάσεων (social recommender systems) σε βασισμένα στη μνήμη (memory-based) και βασισμένα σε μοντέλα (model-based), σύμφωνα με τα βασικά μοντέλα που υιοθετούνται για το κτίσιμό τους, και εξετάζουμε αντιπροσωπευτικά συστήματα από κάθε κατηγορία. Επίσης, παρουσιάζουμε κατευθύνσεις της έρευνας για τη βελτίωση των δυνατοτήτων της κοινωνικής σύστασης.

5.1 Εισαγωγή

Με την ανάπτυξη του Παγκόσμιου Ιστού, η πληροφορία αυξάνεται με άνευ προηγουμένου ρυθμούς και το πρόβλημα της υπερφόρτωσης πληροφορίας (information overload) γίνεται όλο και πιο σοβαρό για τους online χρήστες. Τα συστήματα συστάσεων, τα οποία επιχειρούν

να αντιμετωπίσουν το πρόβλημα της υπερφόρτωσης πληροφορίας προτείνοντας πληροφορίες που ενδεχομένως ενδιαφέρουν τους online χρήστες, έχουν γίνει δημοφιλή ([SKKR01], [Gol06], [MA07], [Kor09], [MZL+11]). Καλές συστάσεις επιτρέπουν σε χρήστες που αναζητούν πληροφορίες να βρουν γρήγορα πληροφορίες σχετικές με τα ενδιαφέροντά τους μέσα από μία μεγάλη ποσότητα άσχετων πληροφοριών. Επίσης, τα συστήματα συστάσεων βοηθούν τους παρόχους πληροφοριών όχι μόνο να αποφασίζουν ποιες πληροφορίες να προσφέρουν στους πελάτες αλλά και να βελτιώνουν την καταναλωτική εμπιστοσύνη (consumer loyalty), καθώς οι πελάτες τείνουν να επιστρέφουν στους ιστότοπους οι οποίοι εξυπηρετούν καλύτερα τις ανάγκες τους ([SKR01]). Τέτοια συστήματα έχουν υλοποιηθεί ευρέως σε διάφορους τομείς (domains) συμπεριλαμβανομένων της σύστασης προϊόντων (product recommendation) στο Amazon και της σύστασης ταινιών (movie recommendation) στο Netflix.

Τα συστήματα συστάσεων αποτέλεσαν ανεξάρτητη ερευνητική περιοχή στα μέσα της δεκαετίας του 1990 ([AT05]) και έχουν προσελκύσει την προσοχή πολλών επιστημών, όπως των μαθηματικών, της φυσικής, της ψυχολογίας και της πληροφορικής ([Ell08]). Για παράδειγμα, μεταξύ των νικητών του διαγωνισμού για το βραβείο Netflix, ενός από τους πιο γνωστούς διαωνισμούς για τη σύσταση, περιλαμβάνονται ψυχολόγοι, επιστήμονες πληροφορικής και φυσικοί. Για το κτίσιμο συστημάτων συστάσεων χρησιμοποιούνται πολλές τεχνικές, οι οποίες μπορούν να ταξινομηθούν γενικά σε βασισμένες στο περιεχόμενο (content-based), βασισμένες στη συνεργατική διήθηση (ΣΔ) (collaborative filtering (CF) based) και υβριδικές ([AT05]). Οι μέθοδοι που βασίζονται στο περιεχόμενο, έχοντας τις ρίζες τους στην ανάκτηση πληροφορίας (information retrieval) ([BR99]) και τη διήθηση πληροφορίας (information filtering) ([BC92]), συνιστούν αντικείμενα παρόμοια με αυτά τα οποία προτίμησε ο χρήστης στο παρελθόν. Οι μέθοδοι που βασίζονται στη ΣΔ προβλέπουν τα ενδιαφέροντα των χρηστών αποκαλύπτοντας πολύπλοκα και μη αναμενόμενα πρότυπα (patterns) από την συμπεριφορά των χρηστών στο παρελθόν και τους συνιστούν αντικείμενα με βάση άλλους χρήστες με παρόμοια ενδιαφέροντα και προτιμήσεις στο παρελθόν ([Kor08], [SK09]). Οι υβριδικές μέθοδοι συνδυάζουν τις παραπάνω κατηγορίες.

Η αυξανόμενη δημοφιλία των κοινωνικών μέσων εμπλουτίζει τις κοινωνικές δραστηριότητες των ανθρώπων με τις οικογένειές τους, τους φίλους τους και τους συναδέλφους τους, με αποτέλεσμα πλούσιες κοινωνικές σχέσεις, όπως οι φιλίες στο Facebook, σχέσεις ακόλουθου/ακολουθούμενου στο Twitter και σχέσεις εμπιστοσύνης στο Epinions. Οι online κοινωνικές σχέσεις παρέχουν ένα διαφορετικό τρόπο ψηφιακής επικοινωνίας των ατόμων και επιτρέπουν στους online χρήστες να μοιράζονται ιδέες και απόψεις με τους χρήστες που συνδέονται με αυτούς. Οι προτιμήσεις ενός χρήστη είναι παρόμοιες ή επηρεάζονται από αυτές των φίλων του. Η λογική πίσω από αυτήν την υπόθεση μπορεί να εξηγηθεί από θεωρίες

κοινωνικής συσχέτισης όπως η ομοιοφιλία (homophily) ([MSC01]) και η κοινωνική επιρροή (social influence) ([MF93]). Η ομοιοφιλία υποδεικνύει ότι χρήστες με παρόμοιες προτιμήσεις είναι πιο πιθανό να συνδεθούν, ενώ η κοινωνική επιρροή ότι χρήστες οι οποίοι είναι συνδεδεμένοι είναι πιο πιθανό να έχουν παρόμοιες προτιμήσεις. Κατ' αναλογία προς το γεγονός ότι οι χρήστες στο φυσικό κόσμο είναι πιθανό να αναζητήσουν προτάσεις από τους φίλους τους πριν πάρουν μια απόφαση για αγορά και οι φίλοι των χρηστών δίνουν με συνέπεια καλές συστάσεις, οι κοινωνικές σχέσεις μπορούν να αξιοποιηθούν για τη βελτίωση της επίδοσης των online συστημάτων συστάσεων ([Gol06], [MYLK08], [JE09], [MZL+11]).

5.2 Παραδοσιακά συστήματα συστάσεων

Ένα τυπικό σύστημα συστάσεων χαρακτηρίζεται από ένα σύνολο χρηστών και ένα σύνολο αντικειμένων. Έστω $\mathbf{u} = \{u_1, u_2, ..., u_n\}$ και $\mathbf{v} = \{v_1, v_2, ..., v_m\}$ τα σύνολα των χρηστών και των αντικειμένων αντίστοιχα, όπου n είναι ο αριθμός των χρηστών και m ο αριθμός των αντικειμένων. Ένας χρήστης u_i βαθμολογεί ένα υποσύνολο του συνόλου των αντικειμένων με κάποιες βαθμολογίες. Έστω $\mathbf{R} \in \mathbb{R}^{n \times m}$ ο πίνακας βαθμολογιών, όπου $\mathbf{R}_{ij}$ είναι η βαθμολογία αν ο χρήστης u_i δίνει μία βαθμολογία στο αντικείμενο v_j, διαφορετικά συμβολίζουμε την τιμή του με το σύμβολο "?" (άγνωστη βαθμολογία). Συνήθως, ο πίνακας βαθμολογιών είναι πολύ αραιός με την έννοια ότι υπάρχουν πολλές άγνωστες βαθμολογίες στον $\mathbf{R}$. Για παράδειγμα, η πυκνότητα του πίνακα βαθμολογιών στα εμπορικά συστήματα συστάσεων είναι συχνά μικρότερη από 1% ([SKKR01]). Αν το αντικείμενο v_j έχει γνωρίσματα, το παριστάνουμε ως $\mathbf{x}_j \in \mathbb{R}^{\ell}$, όπου ℓ είναι ο αριθμός των γνωρισμάτων. Η εργασία των συστημάτων συστάσεων είναι η πρόβλεψη της βαθμολογίας του χρήστη u_i για ένα μη βαθμολογημένο αντικείμενο v_j ή η σύσταση κάποιων αντικειμένων για δεδομένους χρήστες, δηλαδή να προβλέπουν απούσες τιμές στον πίνακα $\mathbf{R}$ με βάση γνωστές βαθμολογίες.

Από τότε που τα συστήματα συστάσεων αποτέλεσαν ανεξάρτητη ερευνητική περιοχή, στα μέσα της δεκαετίας του 1990, έχουν προταθεί πολλά συστήματα συστάσεων, τα οποία μπορούν γενικά να ταξινομηθούν σε βασισμένα στο περιεχόμενο, βασισμένα στη συνεργατική διήθηση και υβριδικά ([RRS11], [JZFF10]). Παρακάτω, εξετάζουμε κάθε κατηγορία και αντιπροσωπευτικά συστήματά της.

5.2.1 Συστήματα συστάσεων βασισμένα στο περιεχόμενο

Τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο έχουν τις ρίζες τους στην ανάκτηση πληροφορίας ([BR99]) και τη διήθηση πληροφορίας ([BC92]). Συνιστούν αντικείμενα παρόμοια με αυτά που προτίμησε ο χρήστης στο παρελθόν. Τα περισσότερα υφιστάμενα συστήματα συστάσεων που βασίζονται στο περιεχόμενο εστιάζουν στη σύσταση αντικειμένων με πληροφορία κειμένου (textual information), όπως ειδήσεις, βιβλία και έγγραφα. Το περιεχόμενο (content) σε αυτά τα συστήματα συνήθως περιγράφεται με λέξεις-κλειδιά ([BS97], [PB97]) και η πληροφορητικότητα (informativeness) μίας λέξης-κλειδιού σε ένα έγγραφο συχνά μετράται με το βάρος TFIDF (TFIDF weight) ([WLWK08]). Το βάρος TF (TF weight) μίας λέξης-κλειδιού σε ένα έγγραφο δηλώνει τη συχνότητά της στο έγγραφο, ενώ το βάρος IDF (IDF weight) μίας λέξης-κλειδιού ορίζεται ως η αντίστροφη συχνότητα εγγράφων (inverse document frequency) της λέξης-κλειδιού.

Έστω x_{jk} το βάρος TFIDF της k-οστής λέξης-κλειδιού στο v_j και ότι το περιεχόμενο του v_j μπορεί να παρασταθεί ως $\mathbf{x}_j = (x_{j1}, x_{j2}, ..., x_{j\ell})$. Με αυτήν την αναπαράσταση, τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο συνιστούν αντικείμενα σε ένα χρήστη τα οποία είναι παρόμοια με αντικείμενα που του άρεσαν στο παρελθόν ([PB97]). Συγκεκριμένα, διάφορα υποψήφια αντικείμενα συγκρίνονται με αντικείμενα που προηγουμένως βαθμολόγησε ο χρήστης. Για να βαθμολογηθούν τα υποψήφια αντικείμενα υιοθετείται ένα μέτρο ομοιότητας όπως η ομοιότητα συνημιτόνου (cosine similarity). Εκτός από τις παραδοσιακές μεθόδους ανάκτησης πληροφορίας που βασίζονται σε ευρετικές τεχνικές (heuristics), χρησιμοποιούνται και άλλες τεχνικές, όπως διάφοροι αλγόριθμοι ταξινόμησης και ομαδοποίησης ([PB97], [MBR98]).

Όπως παρατηρήθηκε στα [BS97] και [AT05], τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο υπόκεινται σε κάποιους περιορισμούς: (1) *περιορισμένη ανάλυση περιεχομένου* (*limited content analysis*) – είναι δύσκολο να εφαρμοσθούν σε πεδία (domains) που έχουν εγγενές πρόβλημα με την αυτόματη εξαγωγή χαρακτηριστικών (features), όπως τα πολυμεσικά δεδομένα (multimedia data), (2) *υπερ-εξειδίκευση* (*over-specialization*) – τα αντικείμενα που συνιστώνται σε ένα χρήστη περιορίζονται σε αυτά που είναι παρόμοια με αυτά που ήδη βαθμολόγησε και (3) *το πρόβλημα του νέου χρήστη* – για να κατανοήσει ένα σύστημα συστάσεων που βασίζεται στο περιεχόμενο τις προτιμήσεις ενός χρήστη, πρέπει αυτός να βαθμολογήσει έναν ικανό αριθμό αντικειμένων, οπότε τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο αποτυγχάνουν να συστήσουν αντικείμενα σε χρήστες με λίγες ή καμμία βαθμολογία.

5.2.2 Συστήματα συστάσεων βασισμένα στη συνεργατική διήθηση

Η συνεργατική διήθηση είναι μία από τις πιο δημοφιλείς τεχνικές για το κτίσιμο συστημάτων συστάσεων ([SKKR01], [Kor08], [SK09]). Μπορεί να προβλέψει τα ενδιαφέροντα του χρήστη απευθείας αποκαλύπτοντας πολύπλοκα ή μη αναμενόμενα πρότυπα από την συμπεριφορά του χρήστη στο παρελθόν χωρίς καμμία γνώση του πεδίου ([Kor08], [SK09]). Η υπόθεση πίσω από τα συστήματα συστάσεων που βασίζονται στη συνεργατική διήθηση είναι ότι αν οι χρήστες συμφώνησαν μεταξύ τους στο παρελθόν είναι πιθανότερο να συμφωνήσουν μεταξύ τους στο μέλλον από ό,τι με τυχαία επιλεγμένους χρήστες. Οι υφιστάμενες μέθοδοι συνεργατικής διήθησης μπορούν να κατηγοριοποιηθούν σε βασισμένες στη μνήμη (memory-based) και βασισμένες σε μοντέλα (model-based) ([GNOT92], [BHK98], [SK09]).

5.2.2.1 Συνεργατική διήθηση βασισμένη στη μνήμη

Οι μέθοδοι που βασίζονται στη μνήμη χρησιμοποιούν είτε ολόκληρο τον πίνακα χρηστών-αντικειμένων ή ένα δείγμα για να παραγάγουν μία πρόβλεψη ([SK09]), και μπορούν να διακριθούν περαιτέρω σε προσανατολισμένες στους χρήστες (user-oriented) ([HKBR99], [BHK98]) και προσανατολισμένες στα αντικείμενα (item-oriented) ([SKKR01], [Kar01]). Οι προσανατολισμένες στους χρήστες μέθοδοι προβλέπουν μία άγνωστη βαθμολογία ενός χρήστη για ένα αντικείμενο ως τη σταθμισμένη μέση τιμή όλων των βαθμολογιών των παρόμοιων χρηστών για το ίδιο αντικείμενο, ενώ οι προσανατολισμένες στα αντικείμενα μέθοδοι με βάση τη μέση βαθμολογία του ίδιου χρήστη για παρόμοια αντικείμενα. Τα κύρια προβλήματα που καλείται να λύσει μία μέθοδος ΣΔ βασισμένη στη μνήμη είναι ο υπολογισμός του βαθμού της ομοιότητας και η συνάθροιση (aggregation) των βαθμολογιών. Οι προσανατολισμένες στους χρήστες και οι προσανατολισμένες στα αντικείμενα μέθοδοι μπορούν να αξιοποιήσουν παρόμοιες τεχνικές για να αντιμετωπίσουν αυτά τα δύο προβλήματα. Έτσι, χρησιμοποιούμε τις προσανατολισμένες στους χρήστες μεθόδους ως παραδείγματα για να εξηγήσουμε αντιπροσωπευτικές μεθόδους υπολογισμού της ομοιότητας και συνάθροισης των βαθμολογιών.

Υπολογισμός ομοιότητας για προσανατολισμένες στους χρήστες μεθόδους: Ο υπολογισμός της ομοιότητας μεταξύ χρηστών είναι ένα κρίσιμο βήμα των προσανατολισμένων στους χρήστες μεθόδων. Έχουν προταθεί πολλές τεχνικές για την επίλυση αυτού του προβλήματος, όπως ο συντελεστής συσχέτισης του Pearson ([RIS+94]), η ομοιότητα συνημιτόνου ([Cho10]) και η ομοιότητα που βασίζεται στην πιθανότητα (probability-based similarity) ([Kar01], [DK04]),

μεταξύ των οποίων οι ευρύτερα χρησιμοποιούμενες είναι ο συντελεστής συσχέτισης του Pearson και η ομοιότητα συνημιτόνου.

- *Συντελεστής συσχέτισης του Pearson:* Κάθε χρήστης παριστάνεται ως ένα διάνυσμα βαθμολογιών. Για παράδειγμα, ο i-οστός χρήστης θα συμβολίζεται με $\mathbf{R}_i$. Ο συντελεστής συσχέτισης του Pearson μετρά το βαθμό στον οποίο δύο μεταβλητές σχετίζονται γραμμικά μεταξύ τους ([RIS+94]). Ο συντελεστής συσχέτισης του Pearson μεταξύ των χρηστών u_i και u_j μπορεί να υπολογισθεί ως:

$$\mathbf{S}_{ij} = \frac{\sum_{k \in I} (R_{ik} - \bar{R}_i) \cdot (R_{jk} - \bar{R}_j)}{\sqrt{\sum_{k \in I} (R_{ik} - \bar{R}_i)^2} \sqrt{\sum_{k \in I} (R_{jk} - \bar{R}_j)^2}} \text{ (5.1),}$$

  όπου με I δηλώνεται το σύνολο των αντικειμένων που βαθμολογήθηκαν από αμφότερους τους u_i και u_j, $\mathbf{S} \in \mathbb{R}^{n \times n}$ είναι ο πίνακας ομοιότητας μεταξύ των χρηστών και $\bar{R}_i$ η μέση βαθμολογία του u_i.

- *Ομοιότητα συνημιτόνου:* Η ομοιότητα συνημιτόνου υπολογίζει το συνημίτονο της γωνίας των διανυσμάτων βαθμολογιών ([Cho10]). Για παράδειγμα, η ομοιότητα συνημιτόνου μεταξύ των u_i και u_j μπορεί να υπολογισθεί ως:

$$\mathbf{S}_{ij} = \frac{\sum_{k \in I} R_{ik} \cdot R_{jk}}{\sqrt{\sum_{k \in I} R_{ik}^2} \sqrt{\sum_{k \in I} R_{jk}^2}} \text{ (5.2).}$$

Συνάθροιση των βαθμολογιών για προσανατολισμένες στους χρήστες μεθόδους: Αφού λάβουν έναν πίνακα ομοιότητας μεταξύ των χρηστών, οι προσανατολισμένες στους χρήστες μέθοδοι προβλέπουν μία απούσα βαθμολογία για ένα δεδομένο χρήστη συναθροίζοντας τις βαθμολογίες των χρηστών που είναι παρόμοιοι με αυτόν. Έχουν προταθεί πολλές στρατηγικές συνάθροισης ([RIS+94], [SKKR01], [Kar01]), από τις οποίες η ευρύτερα χρησιμοποιούμενη είναι ο σταθμισμένος μέσος:

$$\hat{\mathbf{R}}_{ij} = \bar{\mathbf{R}}_i + \frac{\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik} (\mathbf{R}_{kj} - \bar{\mathbf{R}}_k)}{\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik}} \text{ (5.3),}$$

όπου $\mathcal{N}_i$ είναι το σύνολο των χρηστών που έχουν βαθμολογήσει το αντικείμενο v_j.

5.2.2.2 Συνεργατική διήθηση βασισμένη σε μοντέλα

Οι μέθοδοι που βασίζονται σε μοντέλα θεωρούν ένα μοντέλο που παράγει τις βαθμολογίες και εφαρμόζουν τεχνικές εξόρυξης από δεδομένα και μηχανικής μάθησης για να βρουν πρότυπα από δεδομένα εκπαίδευσης ([YK08], [SK09]), τα οποία μπορούν να

χρησιμοποιηθούν για να προβλεφθούν άγνωστες βαθμολογίες. Σε σύγκριση με τη ΣΔ που βασίζεται στη μνήμη, η ΣΔ που βασίζεται σε μοντέλα έχει έναν πιο 'ολιστικό' στόχο, να αποκαλύψει λανθάνοντες παράγοντες (latent factors) που επεξηγούν τις παρατηρηθείσες βαθμολογίες ([YK08]). Κάποιες γνωστές μέθοδοι που βασίζονται σε μοντέλα είναι τα μοντέλα ΣΔ δικτύων πεποίθησης Bayes (Bayesian belief nets) ([BHK98], [MP00]), τα μοντέλα ΣΔ ομαδοποίησης (clustering) ([UF98], [CHW01]), οι μέθοδοι που βασίζονται σε τυχαίους περιπάτους ([YK08], [HZC04]) και τα μοντέλα ΣΔ που βασίζονται σε παραγοντοποίηση ([GRGP01], [Hof04], [Pat07], [Kor08], [SM08]). Οι μέθοδοι ΣΔ που βασίζονται σε παραγοντοποίηση ([Kor08], [SM08]) είναι πολύ καλές αν όχι οι καλύτερες και υιοθετούνται ευρέως για το κτίσιμο συστημάτων συστάσεων ([CGTY11]).

Τα μοντέλα ΣΔ που βασίζονται σε παραγοντοποίηση υποθέτουν ότι λίγα λανθάνοντα πρότυπα (latent patterns) επηρεάζουν τη συμπεριφορά βαθμολόγησης των χρηστών και εκτελούν μία παραγοντοποίηση χαμηλής τάξης (low-rank factorization) στον πίνακα βαθμολογιών. Έστω $\mathbf{u}_i \in \mathbb{R}^K$ και $\mathbf{v}_j \in \mathbb{R}^K$ το διάνυσμα προτιμήσεων του χρήστη u_i και το χαρακτηριστικό διάνυσμα του αντικειμένου v_j αντίστοιχα, όπου K είναι ο αριθμός των λανθανόντων παραγόντων. Τα μοντέλα ΣΔ που βασίζονται σε παραγοντοποίηση επιλύουν το ακόλουθο πρόβλημα:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{W}_{ij}(\mathbf{R}_{ij} - \mathbf{U}_i\mathbf{V}_j^T)^2 + \alpha(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2) \quad (5.4),$$

όπου $\mathbf{U} = [\mathbf{U}_1^T, \mathbf{U}_2^T, ..., \mathbf{U}_n^T]^T \in \mathbb{R}^{n\times K}$ και $\mathbf{V} = [\mathbf{V}_1^T, \mathbf{V}_2^T, ..., \mathbf{V}_m^T]^T \in \mathbb{R}^{m\times K}$. K είναι ο αριθμός των λανθανόντων παραγόντων (προτύπων), ο οποίος συνήθως καθορίζεται μέσω διασταυρωμένης επικύρωσης. Ο όρος $\alpha(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2)$ εισάγεται για την αποφυγή υπερ-προσαρμογής (over-fitting), που ελέγχεται με την παράμετρο α. $\mathbf{W} \in \mathbb{R}^{n\times m}$ είναι ένας πίνακας βαρών όπου $\mathbf{W}_{ij}$ είναι το βάρος της βαθμολογίας του χρήστη u_i για το αντικείμενο v_j. Πολύ συχνά θέτουμε $\mathbf{W}_{ij} = 1$ αν $\mathbf{R}_{ij} \neq 0$. Ο πίνακας βαρών $\mathbf{W}$ μπορεί επίσης να χρησιμοποιηθεί για το χειρισμό της έμμεσης ανατροφοδότησης (implicit feedback) και την κωδικοποίηση παράπλευρης πληροφορίας όπως η συμπεριφορά επιλογών του χρήστη ([FS11]), η ομοιότητα μεταξύ χρηστών και αντικειμένων ([PZC+08], [LHZC10]), η ποιότητα των επισκοπήσεων (reviews) ([RGG12]) και η φήμη (reputation) του χρήστη ([THGL13]).

Τα συστήματα συστάσεων που βασίζονται στη συνεργατική διήθηση μπορούν να ξεπεράσουν κάποιες αδυναμίες των συστημάτων συστάσεων που βασίζονται στο περιεχόμενο. Για παράδειγμα, τα συστήματα που βασίζονται στη ΣΔ χρησιμοποιούν βαθμολογική πληροφορία (rating information), άρα είναι ανεξάρτητα του πεδίου και μπορούν να συστήσουν

οποιοδήποτε αντικείμενο. Ωστόσο, τα συστήματα που βασίζονται στη ΣΔ έχουν τους δικούς τους περιορισμούς, όπως το πρόβλημα της ψυχρής έναρξης (cold-start problem) (νέα αντικείμενα ή νέοι χρήστες) και το πρόβλημα της αραιότητας των δεδομένων (data sparsity) ([AT05], [SK09]).

5.2.3 Υβριδικά συστήματα συστάσεων

Για την υπέρβαση κάποιων περιορισμών των συστημάτων που βασίζονται στο περιεχόμενο ή τη ΣΔ, αναπτύχθηκαν υβριδικές προσεγγίσεις, οι οποίες συνδυάζουν μεθόδους βασισμένες στο περιεχόμενο και τη ΣΔ. Για το συνδυασμό μεθόδων βασισμένων στο περιεχόμενο και στη ΣΔ προτάθηκαν διάφορες στρατηγικές, οι οποίες μπορούν χονδρικά να ταξινομηθούν σε τρεις κατηγορίες ([AT05]). Παρακάτω εξετάζουμε σε συντομία κάθε κατηγορία.

Συνδυασμός διαφορετικών συστημάτων συστάσεων: σε αυτήν τη στρατηγική, μέθοδοι βασισμένες στο περιεχόμενο και στη ΣΔ υλοποιούνται ξεχωριστά και οι προβλέψεις τους συνδυάζονται για να ληφθεί η τελική σύσταση. Έχουν προταθεί διάφοροι τρόποι για το συνδυασμό των προβλέψεων, όπως ένα σχήμα ψηφοφορίας (voting scheme) ([Paz99]) και ένας γραμμικός συνδυασμός των βαθμολογιών ([CGM+99]).

Προσθήκη χαρακτηριστικών βασισμένων στο περιεχόμενο σε μοντέλα ΣΔ: τα συστήματα που προκύπτουν με αυτήν τη στρατηγική χρησιμοποιούν προφίλ βασισμένα σε περιεχόμενο και μη από κοινού βαθμολογημένα αντικείμενα για να υπολογίσουν τις ομοιότητες μεταξύ των χρηστών. Μπορούν να ξεπεράσουν κάποιους περιορισμούς των μεθόδων ΣΔ σχετικά με την αραιότητα ([Paz99], [GSK+99]).

Προσθήκη χαρακτηριστικών ΣΔ σε μοντέλα βασισμένα στο περιεχόμενο: η πιο δημοφιλής προσέγγιση που ακολουθεί αυτήν τη στρατηγική είναι η χρήση τεχνικής μείωσης διαστατικότητας (dimensionality reduction) στον πίνακα προφίλ. Για παράδειγμα, στο [SN99] χρησιμοποιείται λανθάνουσα σημασιολογική δεικτοδότηση (latent semantic indexing) για τη δημιουργία μίας συνεργατικής όψης (collaborative view) ενός συνόλου προφίλ χρηστών, με αποτέλεσμα τη βελτίωση της επίδοσης της σύστασης σε σύγκριση με αμιγείς προσεγγίσεις βασισμένες στο περιεχόμενο.

5.3 Κοινωνική σύσταση

Ένα από τα πρώτα κοινωνικά συστήματα συστάσεων (social recommendation systems) εμφανίσθηκε το 1997 ([KSS97]). Πολλές υπηρεσίες κοινωνικών μέσων, όπως το Facebook και το Twitter, εμφανίσθηκαν τα τελευταία χρόνια, επιτρέποντας στους ανθρώπους να

επικοινωνούν και να εκφράζονται με ευκολία. Η γενικευμένη χρήση των κοινωνικών μέσων παράγει κοινωνική πληροφορία (social information) με έναν άνευ προηγουμένου ρυθμό. Η ταχεία ανάπτυξη των κοινωνικών μέσων έχει επιταχύνει πολύ την ανάπτυξη των κοινωνικών συστημάτων συστάσεων ([KLM10], [GC11]). Σε αυτήν τη ενότητα θα δώσουμε τους ορισμούς της κοινωνικής σύστασης, θα συζητήσουμε τις ευκαιρίες για κοινωνικά συστήματα συστάσεων συγκριτικά με παραδοσιακά συστήματα συστάσεων, θα ταξινομήσουμε και θα εξετάσουμε τα υφιστάμενα κοινωνικά συστήματα συστάσεων.

5.3.1 Ορισμοί της κοινωνικής σύστασης

Η κοινωνική σύσταση μελετάται από το 1997 ([KSS97]) και προσελκύει όλο και περισσότερη προσοχή με την αυξανόμενη δημοφιλία των κοινωνικών μέσων ([KLM10], [GC11]), ωστόσο δεν υπάρχει κοινώς αποδεκτός ορισμός για την κοινωνική σύσταση. Σε αυτήν τη υποενότητα, δίνουμε ένα στενό και έναν ευρύ ορισμό της κοινωνικής σύστασης με βάση υπάρχοντες ορισμούς στη βιβλιογραφία.

Ένας στενός ορισμός της κοινωνικής σύστασης αναφέρει ότι είναι οποιαδήποτε σύσταση με online κοινωνικές σχέσεις ως επιπρόσθετη είσοδο. Οι κοινωνικές σχέσεις μπορεί να είναι σχέσεις εμπιστοσύνης, φιλίες, σχέσεις συμμετοχής (memberships) ή σχέσεις ακόλουθου/ακολουθούμενου (following relations). Σε αυτόν τον ορισμό τα κοινωνικά συστήματα συστάσεων υποθέτουν ότι οι χρήστες συσχετίζονται όταν συνάπτουν κοινωνικές σχέσεις ([Mas07], [MA07], [MYLK08]). Για παράδειγμα, οι προτιμήσεις των χρηστών είναι πιθανό να είναι παρόμοιες ή να επηρεάζονται από αυτές των φίλων τους. Κάνοντας αυτήν την υπόθεση, η κοινωνική σύσταση αξιοποιεί τις συσχετίσεις μεταξύ των χρηστών οι οποίες υποδηλώνονται από τις κοινωνικές σχέσεις για να βελτιώσει την επίδοση της σύστασης. Αντιπροσωπευτικά συστήματα που ανταποκρίνονται σε αυτόν τον ορισμό είναι τα TidalTrust ([Gol06]), MoleTrust ([MA07], [VDC11]), SoRec ([MLN+09]), SocialMF ([JE09]), SoReg ([MZL+11]) και LOCALBAL ([THGL13]).

Ένας άλλος, ευρύς ορισμός της κοινωνικής σύστασης αναφέρει ότι κοινωνικό σύστημα συστάσεων είναι κάθε σύστημα συστάσεων που στοχεύει σε τομείς (domains) κοινωνικών μέσων ([GC11]). Ο ορισμός αυτός περιλαμβάνει συστήματα συστάσεων που συνιστούν οποιαδήποτε οντότητα (object) σε τομείς κοινωνικών μέσων, όπως αντικείμενα (items), ετικέτες (tags) ([SV08]), ανθρώπους ([CGD+09], [AB12]) και κοινότητες (communities) ([CCL+09]). Οι πηγές που χρησιμοποιούν δεν περιορίζονται σε online κοινωνικές σχέσεις, αλλά περιλαμβάνουν κάθε είδος δεδομένων από κοινωνικά μέσα, όπως κοινωνική επισήμανση (social tagging) ([MZLK11]), αλληλεπιδράσεις μεταξύ των χρηστών ([JCL+12]) και συμπεριφορά επιλογών (click behaviors) των χρηστών ([MYH+07]).

Σε αυτό το κεφάλαιο, εστιάζουμε σε κοινωνικά συστήματα συστάσεων που ανταποκρίνονται στο στενό ορισμό. Παρόμοιες τεχνικές μπορούν να εφαρμοσθούν για την υλοποίηση κοινωνικών συστημάτων συστάσεων που ανταποκρίνονται σε οποιονδήποτε από τους δύο ορισμούς. Ο στενός ορισμός είναι απλός και βοηθά στην καλύτερη κατανόηση των κοινωνικών συστημάτων συστάσεων αιχμής.

5.3.2 Μία ιδιαιτερότητα της κοινωνικής σύστασης και οι συνέπειές της

Η αυξανόμενη δημοφιλία των κοινωνικών μέσων επιτρέπει στους online χρήστες να συμμετέχουν σε online δραστηριότητες οι οποίες παράγουν πλούσιες κοινωνικές σχέσεις. Στα κοινωνικά συστήματα συστάσεων, επιπρόσθετα προς τη βαθμολογική πληροφορία $\mathbf{R}$, οι χρήστες μπορούν να συνδέονται μεταξύ τους. Έστω $\mathbf{T} \in \mathbb{R}^{n \times n}$ ένας πίνακας που δηλώνει τις κοινωνικές σχέσεις μεταξύ των χρηστών για τον οποίο $\mathbf{T}_{ij} = 1$ αν ο u_j συνδέεται με τον u_i και $\mathbf{T}_{ij} = 0$ διαφορετικά. Η διαθεσιμότητα των κοινωνικών σχέσεων $\mathbf{T}$ παρέχει μία ανεξάρτητη πηγή για σύσταση και αυτή η ιδιαιτερότητα της κοινωνικής σύστασης προσφέρει νέες ευκαιρίες.

Πρώτον, τα παραδοσιακά συστήματα συστάσεων υποθέτουν ότι οι χρήστες προέρχονται από ανεξάρτητες και ισόνομες κατανομές (υπόθεση γνωστή ως υπόθεση i.i.d.). Ωστόσο, οι online χρήστες είναι εγγενώς συνδεδεμένοι μέσω σχέσεων διαφόρων τύπων, όπως φιλίες και σχέσεις εμπιστοσύνης. Επιπρόσθετα προς τον πίνακα βαθμολογιών στα παραδοσιακά συστήματα συστάσεων, οι χρήστες στα κοινωνικά συστήματα συστάσεων είναι συνδεδεμένοι, παρέχοντας κοινωνική πληροφορία. Επειδή είναι συνδεδεμένοι, οι χρήστες είναι συσχετισμένοι και όχι i.i.d.. Για παράδειγμα, οι συγγραφείς στο [WLJH10] βρίσκουν ότι χρήστες με σχέσεις ακόλουθου/ακολουθούμενου είναι πιθανότερο να έχουν παρόμοια ενδιαφέροντα από ό,τι δύο τυχαία επιλεγμένοι χρήστες, και οι συγγραφείς στο [TGHL13b] δείχνουν ότι χρήστες με σχέσεις εμπιστοσύνης είναι πιθανότερο να έχουν παρόμοιες προτιμήσεις σε βαθμολογίες αντικειμένων. Αυτό το φαινόμενο παρατηρείται στα περισσότερα online κοινωνικά δίκτυα και μπορεί να επεξηγηθεί με θεωρίες κοινωνικής συσχέτισης (social correlation) όπως η ομοιοφιλία ([MSC01]) και η κοινωνική επιρροή ([MF93]). Ας σημειωθεί ότι το γεγονός ότι ένας χρήστης είναι πιθανότερο να έχει παρόμοια ενδιαφέροντα με τους χρήστες που συνδέονται με αυτόν από ό,τι με τυχαία επιλεγμένους χρήστες δε σημαίνει ότι οι χρήστες που συνδέονται με έναν χρήστη ταυτίζονται με τους πιο παρόμοιους με αυτόν χρήστες ως προς τη βαθμολογική πληροφορία. Υποθέτουμε ότι ο χρήστης u_i συνδέεται με d χρήστες. Εστω C_i και S_i τα σύνολα (μεγέθους d) των χρηστών που συνδέονται με τον u_i και των d πιο παρόμοιων με αυτόν χρηστών, αντίστοιχα.

Οι συγγραφείς στο [TGHL13a] έδειξαν ότι η επικάλυψη (overlap) μεταξύ των C_i και S_i είναι μικρότερη από 10%, το οποίο είναι συνεπές με την παρατήρηση στο [CCH+08]. Αν αλλάξουμε τη θεώρησή μας και θεωρήσουμε τις συνδέσεις ως μετρήσεις ομοιότητας, η κοινωνική πληροφορία παρέχει μαρτυρία ομοιότητας (similarity evidence) και το σύνολο C_i είναι η λίστα των παρόμοιων χρηστών ως προς την κοινωνική πληροφορία ([GJS+08]). Τότε, η παραπάνω παρατήρηση μπορεί να επεξηγηθεί με την καίρια διαπίστωση στο [GJP+10] – λίστες παρόμοιων χρηστών που λαμβάνονται από διαφορετικές πηγές βρέθηκε ότι διαφέρουν πολύ μεταξύ τους. Επίσης υποδεικνύει ότι μία συνάθροιση της βαθμολογικής και κοινωνικής πληροφορίας θα μπορούσε να είχε μεγάλη αξία. Εκτός από μαρτυρία ομοιότητας, η κοινωνική πληροφορία παρέχει μία άλλη μοναδική μαρτυρία, δηλαδή μαρτυρία οικειότητας (familiarity evidence), η οποία είναι πολύ σημαντική στη σύσταση, καθώς στο φυσικό κόσμο συνήθως ζητούμε προτάσεις από τους φίλους μας που είναι εξοικειωμένοι (familiar) με τις προτιμήσεις μας ([GJP+10], [GJS+08]). Και η μαρτυρία ομοιότητας και η μαρτυρία οικειότητας από κοινωνική πληροφορία υποδηλώνουν ότι τα κοινωνικά δίκτυα περιέχουν πληροφορία συμπληρωματική προς τη βαθμολογική πληροφορία και παρέχουν μία ανεξάρτητη πηγή πληροφορίας για τους online χρήστες. Η αξιοποίηση των κοινωνικών σχέσεων μπορεί ενδεχομένως να βελτιώσει την επίδοση της σύστασης.

Δεύτερον, εκτός από τα συστήματα συστάσεων, η κοινωνική σύσταση εμπλέκει το ανεξάρτητο ερευνητικό πεδίο της ανάλυσης κοινωνικών δικτύων ([WF94], [Sco11], [Sco12], [DLC13]). Τα κοινωνικά συστήματα συστάσεων μπορούν να επωφεληθούν από τα ερευνητικά αποτελέσματα της ανάλυσης κοινωνικών δικτύων, όπως οι θεωρίες κοινωνικής συσχέτισης ([MSC01], [MF93]), η ανάλυση κατάστασης (status analysis) ([PBMW99], [Kle99]), η ανίχνευση κοινοτήτων ([LHK10], [TL10]), η online εμπιστοσύνη (online trust) ([Mas07], [TGL12]) και τα ετερογενή (heterogeneous) δίκτυα ([SH12]), στη σύσταση. Τα ερευνητικά επιτεύγματα της ανάλυσης κοινωνικών δικτύων ανοίγουν δρόμους στην αξιοποίηση των κοινωνικών σχέσεων και μπορούν να εφαρμοσθούν στο κτίσιμο κοινωνικών συστημάτων συστάσεων.

5.3.3 Υφιστάμενα κοινωνικά συστήματα συστάσεων

Όπως αναφέρθηκε παραπάνω, η συνεργατική διήθηση υιοθετείται ευρέως για το κτίσιμο συστημάτων συστάσεων και τα περισσότερα υφιστάμενα κοινωνικά συστήματα συστάσεων βασίζονται σε τεχνικές ΣΔ. Έτσι, σε αυτό το κεφάλαιο εστιάζουμε σε κοινωνικά συστήματα συστάσεων που βασίζονται στη ΣΔ. Η κοινωνική σύσταση δέχεται δύο εισόδους, τη βαθμολογική πληροφορία και την κοινωνική πληροφορία. Τα περισσότερα υφιστάμενα

κοινωνικά συστήματα συστάσεων επιλέγουν μοντέλα ΣΔ ως τα βασικά τους μοντέλα για το κτίσιμό τους και προτείνουν προσεγγίσεις για τη σύλληψη της κοινωνικής πληροφορίας βασισμένες σε αποτελέσματα από την ανάλυση κοινωνικών δικτύων. Επομένως, ένα γενικό πλαίσιο κοινωνικής σύστασης βασισμένης στη ΣΔ αποτελείται από δύο μέρη: (1) ένα βασικό μοντέλο ΣΔ και (2) ένα μοντέλο κοινωνικής πληροφορίας. Αυτό μπορεί να δηλωθεί ως:

μοντέλο κοινωνικής σύστασης βασισμένης στη ΣΔ = βασικό μοντέλο ΣΔ + μοντέλο κοινωνικής πληροφορίας (5.5).

Το βασικό μοντέλο ΣΔ σε ένα παρέχει ένα κριτήριο ταξινόμησης των κοινωνικών συστημάτων συστάσεων. Ταξινομούμε τα κοινωνικά συστήματα συστάσεων σε δύο μεγάλες κατηγορίες σύμφωνα με τα βασικά τους μοντέλα ΣΔ: κοινωνικά συστήματα συστάσεων βασισμένα στη μνήμη και κοινωνικά συστήματα συστάσεων βασισμένα σε μοντέλα.

5.3.3.1 Κοινωνικά συστήματα συστάσεων βασισμένα στη μνήμη

Τα κοινωνικά συστήματα συστάσεων που βασίζονται στη μνήμη χρησιμοποιούν μοντέλα ΣΔ βασισμένα στη μνήμη, ιδιαίτερα μεθόδους προσανατολισμένες στους χρήστες, ως τα βασικά τους μοντέλα. Μία απούσα βαθμολογία για έναν δεδομένο χρήστη υπολογίζεται με συνάθροιση των βαθμολογιών των χρηστών που συσχετίζονται με αυτόν, οι οποίοι συμβολίζονται με N^+. Για ένα δεδομένο χρήστη, οι παραδοσιακές προσανατολισμένες στους χρήστες μέθοδοι χρησιμοποιούν παρόμοιους χρήστες N, ενώ τα κοινωνικά συστήματα συστάσεων που βασίζονται στη μνήμη χρησιμοποιούν συσχετισμένους χρήστες N^+, οι οποίοι λαμβάνονται τόσο από τη βαθμολογική πληροφορία όσο και από την κοινωνική πληροφορία. Τα κοινωνικά συστήματα συστάσεων αυτής της κατηγορίας συνήθως ακολουθούν δύο βήματα. Στο πρώτο βήμα, λαμβάνουν λαμβάνουν τους συσχετισμένους χρήστες $N^+(i)$ για ένα δεδομένο χρήστη u_i και το δεύτερο βήμα είναι το κλασικό τελευταίο βήμα των μεθόδων ΣΔ που βασίζονται στη μνήμη – συνάθροιση των βαθμολογιών από τους συσχετισμένους χρήστες που ελήφθησαν από το πρώτο βήμα για τις απούσες βαθμολογίες. Διαφορετικά κοινωνικά συστήματα συστάσεων αυτής της κατηγορίας χρησιμοποιούν διαφορετικές προσεγγίσεις για να λάβουν τους συσχετισμένους χρήστες N^+ στο πρώτο βήμα. Παρακάτω, παρουσιάζουμε λεπτομέρειες σχετικά με κάποιες αντιπροσωπευτικές προσεγγίσεις.

Social based Weight Mean ([VCDT09], [VDC11]): Για ένα δεδομένο χρήστη u_i, η στρατηγική αυτή απλώς θεωρεί τους απευθείας συνδεδεμένους χρήστες με τον u_i, $\mathcal{F}(i)$, ως το σύνολο των συσχετισμένων με αυτόν χρηστών $N^+(i)$:

$$N^+(i) = \{u_j \mid \mathbf{T}(i, j) = 1\} \text{ (5.6).}$$

TidalTrust ([Gol06]): Σε αυτό το σύστημα οι χρήστες συνδέονται με σχέσεις εμπιστοσύνης. Οι συγγραφείς σχεδίασαν το μέτρο TidalTrust για να εκτιμήσουν τις τιμές εμπιστοσύνης μεταξύ των χρηστών βασισμένοι στις ακόλουθες δύο παρατηρήσεις: (1) συντομότερα μονοπάτια διάδοσης (propagation) παράγουν ακριβέστερες εκτιμήσεις εμπιστοσύνης και (2) μονοπάτια με υψηλότερες τιμές εμπιστοσύνης δίνουν καλύτερα αποτελέσματα. Για να εκτιμήσει τις τιμές εμπιστοσύνης μεταξύ των χρηστών, το TidalTrust ακολουθεί τα παρακάτω βήματα:

- αναζητά ένα συντομότερο μονοπάτι από το χρήστη-πηγή προς τους χρήστες που έδωσαν βαθμολογίες και θέτει το μήκος του συντομότερου μονοπατιού ως το βάθος μονοπατιού του αλγορίθμου.
- υπολογίζει την τιμή εμπιστοσύνης από το χρήστη-πηγή προς έναν χρήστη που έδωσε βαθμολογία στο δεδομένο βάθος. Για ένα ζεύγος χρηστών u_i και u_j οι οποίοι δε συνδέονται άμεσα, η τιμή εμπιστοσύνης συναθροίζεται από τις τιμές εμπιστοσύνης από τους άμεσους γείτονες του u_i προς τον u_j σταθμισμένες με τις τιμές εμπιστοσύνης από τον u_i προς τους άμεσους γείτονές του:

$$\mathbf{S}_{ij} = \frac{\sum_{u_k \in \mathcal{F}(i)} \mathbf{S}_{ik}\mathbf{S}_{kj}}{\sum_{u_k \in \mathcal{F}(i)} \mathbf{S}_{ik}} \text{ (5.7).}$$

- αφού υπολογισθούν οι τιμές εμπιστοσύνης $\mathbf{S}$, καθορίζεται το σύνολο $N^+(i)$ ως το σύνολο των χρηστών προς τους οποίους οι τιμές εμπιστοσύνης από τον u_i υπερβαίνουν ένα δεδομένο κατώφλι τ:

$$N^+(i) = \{u_j \mid \mathbf{S}_{ij} \geq \tau\} \text{ (5.8).}$$

MoleTrust ([MA04], [MA05], [MA07]): Ο υπολογισμός του μέτρου MoleTrust αποτελείται από δύο κύρια βήματα. Αρχικά, απαλείφονται οι κύκλοι στο δίκτυο. Για να ληφθούν οι τιμές εμπιστοσύνης, πρέπει να εκτελεσθεί ένας μεγάλος αριθμός διαδόσεων εμπιστοσύνης. Επομένως, η απαλοιφή των κύκλων μπορεί να επιταχύνει σημαντικά τον προτεινόμενο αλγόριθμο, καθώς κάθε χρήστης χρειάζεται να δεχθεί επίσκεψη μόνο μία φορά για τον υπολογισμό των τιμών εμπιστοσύνης. Με αυτήν την πράξη το αρχικό δίκτυο εμπιστοσύνης μετατρέπεται σε έναν κατευθυνόμενο ακυκλικό γράφο (directed acyclic graph). Στο δεύτερο βήμα, οι τιμές εμπιστοσύνης υπολογίζονται με βάση τον κατευθυνόμενο ακυκλικό γράφο με έναν απλό τυχαίο περίπατο σε αυτόν: πρώτα υπολογίζεται η εμπιστοσύνη προς τους χρήστες που βρίσκονται 1 βήμα μακριά, μετά η εμπιστοσύνη προς τους χρήστες που βρίσκονται 2 βήματα μακριά κ.ο.κ. Αφού υπολογισθούν οι τιμές εμπιστοσύνης, το MoleTrust καθορίζει

τους χρήστες εντός του μέγιστου βάθους που έχουν βαθμολογήσει το αντικείμενο-στόχο ως τους συσχετισμένους χρήστες N^+, όπου το μέγιστο βάθος είναι μία προκαθορισμένη παράμετρος.

TrustWalker ([JE09]): Η διαισθητική ιδέα αυτού του συστήματος προέρχεται από δύο καίριες παρατηρήσεις. Πρώτον, το κοινωνικό δίκτυο ενός χρήστη (δηλαδή, οι γείτονές του) έχει μικρή επικάλυψη με τους χρήστες που είναι παρόμοιοι με αυτόν ([CCH+08]), πράγμα που σημαίνει ότι η κοινωνική πληροφορία αποτελεί ανεξάρτητη πηγή πληροφορίας. Δεύτερον, οι βαθμολογίες για παρόμοια αντικείμενα χρηστών στους οποίους ο χρήστης έχει μεγάλη εμπιστοσύνη είναι πιο αξιόπιστες από βαθμολογίες για το ίδιο αντικείμενο-στόχο χρηστών στους οποίους έχει λίγη εμπιστοσύνη. Η πρώτη παρατήρηση υποδεικνύει τη σημασία των προσεγγίσεων που βασίζονται στην εμπιστοσύνη, ενώ η δεύτερη τη δύναμη των προσανατολισμένων στα αντικείμενα προσεγγίσεων. Για να επωφεληθεί και από τα δύο είδη προσεγγίσεων, το TrustWalker προτείνει ένα μοντέλο βασισμένο σε τυχαίους περιπάτους ώστε να συνδυάσει προσεγγίσεις βασισμένες στην εμπιστοσύνη και προσανατολισμένες στους χρήστες σε ένα συνεκτικό πλαίσιο. Αναζητά τις βαθμολογίες των άμεσων και έμμεσων φίλων ενός χρήστη για το αντικείμενο-στόχο καθώς επίσης και για παρόμοια αντικείμενα εκτελώντας έναν τυχαίο περίπατο σε online κοινωνικά δίκτυα. Για παράδειγμα, ας θεωρήσουμε ότι θέλουμε να λάβουμε μία βαθμολογία του χρήστη u_i για το αντικείμενο v_j. Υποθέτουμε ότι βρισκόμαστε στον κόμβο u_k. Το TrustWalker λειτουργεί όπως παρακάτω σε κάθε βήμα του τυχαίου περιπάτου: αν ο u_k βαθμολόγησε το v_j τότε σταματά τον τυχαίο περίπατο και επιστρέφει την τιμή $\mathbf{R}_{kj}$ σαν αποτέλεσμα, διαφορετικά έχει δύο επιλογές – (1) σταματά τον τυχαίο περίπατο, επιλέγει τυχαία ένα αντικείμενο v_l παρόμοιο με το v_j, το οποίο βαθμολόγησε ο u_i και επιστρέφει την τιμή $\mathbf{R}_{il}$, ή (2) συνεχίζει τον τυχαίο περίπατο και προχωρά στο χρήστη u_k στα δίκτυα εμπιστοσύνης του u_i.

Το TrustWalker χρησιμοποιεί το συντελεστή συσχέτισης του Pearson των βαθμολογιών για να υπολογίσει την ομοιότητα μεταξύ αντικειμένων. Καθώς οι τιμές του συντελεστή συσχέτισης του Pearson βρίσκονται στο διάστημα $[-1,1]$, μόνο αντικείμενα με θετική συσχέτιση με το αντικείμενο-στόχο λαμβάνονται υπόψη. Η ομοιότητα μεταξύ των αντικειμένων v_i και v_j υπολογίζεται ως:

$$sim(i,j) = \frac{1}{1+e^{-\frac{N_{ij}}{2}}} \times PCC(i,j) \quad (5.9),$$

όπου N_{ij} είναι ο αριθμός των χρηστών οι οποίοι βαθμολόγησαν αμφότερα τα v_i και v_j και $PCC(i, j)$ είναι ο συντελεστής συσχέτισης του Pearson των v_i και v_j.

5.3.3.2 Κοινωνικά συστήματα συστάσεων βασισμένα σε μοντέλα

Τα κοινωνικά συστήματα συστάσεων που βασίζονται σε μοντέλα επιλέγουν μεθόδους ΣΔ βασισμένες σε μοντέλα ως τα βασικά τους μοντέλα. Τεχνικές παραγοντοποίησης πινάκων χρησιμοποιούνται ευρέως στις μεθόδους ΣΔ που βασίζονται σε μοντέλα. Αυτές οι τεχνικές παρουσιάζουν διάφορες 'ευχάριστες' ιδιότητες ([DKA11], [ME11]): (1) πολλές μέθοδοι βελτιστοποίησης, όπως μέθοδοι βασισμένες στην κλίση (gradient based), μπορούν να εφαρμοσθούν για να βρεθεί μία βέλτιστη λύση, επεκτάσιμη σε χιλιάδες χρήστες με εκατομμύρια σχέσεις εμπιστοσύνης, (2) η παραγοντοποίηση πινάκων έχει μία 'ευχάριστη' πιθανοτική ερμηνεία και (3) είναι πολύ ευέλικτη και επιτρέπει να περιληφθεί πρότερη (prior) γνώση. Τα περισσότερα υπάρχοντα κοινωνικά συστήματα συστάσεων αυτής της κατηγορίας βασίζονται σε παραγοντοποίηση πινάκων ([MYLK08], [MKL09], [YZC+09], [JE10], [VNLD10], [MZL+11], [AI11], [YLS+11], [STM11], [TGL12], [TGLD12], [NST+12], [HDD13], [TGHL13b]). Η λογική πίσω από αυτές τις μεθόδους είναι ότι οι προτιμήσεις των χρηστών είναι παρόμοιες ή επηρεάζονται από αυτές των χρηστών με τους οποίους αυτοί συνδέονται κοινωνικά. Ωστόσο, το χαμηλό κόστος του σχηματισμού κοινωνικών σχέσεων μπορεί να οδηγήσει σε κοινωνικές σχέσεις με ετερογενείς δυνάμεις (heterogeneous strengths) (π.χ. ασθενείς δεσμοί (weak ties) και ισχυροί δεσμοί (strong ties) ανάμικτοι) ([XNR10]). Καθώς οι χρήστες με ισχυρούς δεσμούς είναι πιθανότερο να έχουν κοινές προτιμήσεις από ό,τι αυτοί με ασθενείς δεσμούς, η ίση αντιμετώπιση όλων των κοινωνικών σχέσεων μπορεί να οδηγήσει σε υποβάθμιση (degradation) στην επίδοση της σύστασης. Γι' αυτό, με κάθε κοινωνική σχέση αυτές οι μέθοδοι συσχετίζουν μία δύναμη (strength), η οποία συνήθως υπολογίζεται με την ομοιότητα των βαθμολογιών. Για παράδειγμα, όταν επιλέγουμε την ομοιότητα συνημιτόνου, αν οι u_i και u_k συνδέονται, η τιμή $\mathbf{S}_{ik}$ υπολογίζεται ως η ομοιότητα συνημιτόνου μεταξύ των διανυσμάτων βαθμολογιών των u_i και u_k, διαφορετικά θέτουμε $\mathbf{S}_{ik} = 0$. Επομένως, η δύναμη μεταξύ των u_i και u_k, $\mathbf{S}_{ik}$, μπορεί να ορισθεί τυπικά ως:

$$\mathbf{S}_{ik} = \begin{cases} \frac{\sum_j \mathbf{R}_{ij} \cdot \mathbf{R}_{kj}}{\sqrt{\sum_j \mathbf{R}_{ij}^2} \sqrt{\sum_j \mathbf{R}_{kj}^2}} & \text{αν οι } u_i \text{ και } u_j \text{ συνδέονται} \\ 0 & \text{διαφορετικά} \end{cases} \quad (5.10).$$

Σε αντίθεση με τα παραδοσιακά συστήματα συστάσεων που βασίζονται στην παραγοντοποίηση πινάκων, τα κοινωνικά συστήματα συστάσεων αυτής της κατηγορίας

αξιοποιούν την κοινωνική πληροφορία και ένα ενοποιημένο πλαίσιο μπορεί να δηλωθεί ως εξής:

$$\min_{\mathbf{U},\mathbf{V},\Omega} \left\| \mathbf{W} \odot (\mathbf{R} - \mathbf{U}^T \mathbf{V}) \right\|_F^2 + \alpha \; Social(\mathbf{T}, \mathbf{S}, \Omega) + \lambda (\left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{V} \right\|_F^2 + \left\| \Omega \right\|_F^2) \; (5.11),$$

όπου ο όρος $Social(\mathbf{T}, \mathbf{S}, \Omega)$ εισάγεται για να συλλάβει κοινωνική πληροφορία, Ω είναι το σύνολο των παραμέτρων που μαθαίνονται από την κοινωνική πληροφορία, η παράμετρος α ελέγχει τη συνεισφορά της κοινωνικής πληροφορίας και ο πίνακας $\mathbf{W}$ ελέγχει τα βάρη των γνωστών βαθμολογιών στη διαδικασία της μάθησης. Ανάλογα με το πώς ορίζεται ο όρος $Social(\mathbf{T}, \mathbf{S}, \Omega)$, διακρίνουμε τα κοινωνικά συστήματα συστάσεων αυτής της κατηγορίας σε τρεις ομάδες: μέθοδοι συν-παραγοντοποίησης, μέθοδοι συνόλου και μέθοδοι ομαλοποίησης (regularization). Παρακάτω, εξετάζουμε λεπτομερώς αντιπροσωπευτικά συστήματα από κάθε ομάδα.

Μέθοδοι συν-παραγοντοποίησης (co-factorization methods) ([MYLK08], [TGHL13b]): Η υπόθεση πίσω από τα συστήματα αυτής της ομάδας είναι ότι ο i-οστός χρήστης u_i θα έπρεπε να έχει το ίδιο διάνυσμα προτιμήσεων $\mathbf{u}_i$ στο χώρο των βαθμολογιών (βαθμολογική πληροφορία) και στον κοινωνικό χώρο (κοινωνική πληροφορία). Τα κοινωνικά συστήματα συστάσεων αυτής της ομάδας εκτελούν συν-παραγοντοποίηση στον πίνακα χρηστών-αντικειμένων και τον πίνακα κοινωνικών σχέσεων μεταξύ των χρηστών χρησιμοποιώντας τον ίδιο λανθάνοντα παράγοντα προτιμήσεων χρηστών. Τα SoRec ([MYLK08]) και LOCABAL ([TGHL13b]) είναι δύο αντιπροσωπευτικά συστήματα αυτής της ομάδας.

SoRec ([MYLK08]): Το SoRec ορίζει το $Social(\mathbf{T}, \mathbf{S}, \Omega)$ ως:

$$\min \sum_{i=1}^{n} \sum_{u_k \in \mathcal{N}_i} (\mathbf{S}_{ik} - \mathbf{u}_i^T \mathbf{z}_k)^2 \; (5.12),$$

όπου $\mathcal{N}_i$ είναι το σύνολο των χρηστών που συνδέονται απευθείας με τον u_i και $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_n] \in \mathbb{R}^{K \times n}$ ο πίνακας λανθανόντων χαρακτηριστικών. Ο πίνακας προτιμήσεων των χρηστών $\mathbf{U}$ μαθαίνεται από αμφότερες την κοινωνική πληροφορία και τη βαθμολογική πληροφορία με επίλυση του παρακάτω προβλήματος βελτιστοποίησης:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{Z}} \left\| \mathbf{W} \odot (\mathbf{R} - \mathbf{U}^T \mathbf{V}) \right\|_F^2 + \alpha \sum_{i=1}^{n} \sum_{u_k \in \mathcal{N}_i} (\mathbf{S}_{ik} - \mathbf{u}_i^T \mathbf{z}_k)^2 + \lambda (\left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{V} \right\|_F^2 + \left\| \mathbf{Z} \right\|_F^2) \; (5.13).$$

LOCABAL ([TGHL13b]): Το LOCABAL ορίζει το $Social(\mathbf{T}, \mathbf{S}, \Omega)$ ως:

$$\min \sum_{i=1}^{n} \sum_{u_k \in \mathcal{N}_i} (\mathbf{S}_{ik} - \mathbf{u}_i^T \mathbf{H} \mathbf{u}_k)^2 \; (5.14).$$

Το LOCABAL βασίζεται σε θεωρίες κοινωνικής συσχέτισης όπου οι προτιμήσεις δύο κοινωνικά συνδεδεμένων χρηστών συσχετίζονται μέσω του πίνακα συσχέτισης $\mathbf{H}$. Στην παραπάνω εξίσωση, τα διανύσματα προτιμήσεων $\mathbf{u}_i$ και $\mathbf{u}_k$ δύο κοινωνικά συνδεδεμένων χρηστών u_i και u_k συσχετίζονται μέσω του $\mathbf{H} \in \mathbb{R}^{K \times K}$, που ελέγχεται από τη δύναμή τους $\mathbf{S}_{ik}$. Μεγάλη τιμή του $\mathbf{S}_{ik}$, δηλαδή ισχυρή σύνδεση μεταξύ των u_i και u_k, δείχνει ότι οι προτιμήσεις τους πρέπει να συσχετισθούν σφικτά μέσω του $\mathbf{H}$, ενώ μικρή τιμή δείχνει ότι πρέπει να συσχετισθούν χαλαρά. Το $Social(\mathbf{T}, \mathbf{S}, \Omega)$ μπορεί να εφαρμοσθεί τόσο σε κατευθυνόμενα όσο και σε μη κατευθυνόμενα κοινωνικά δίκτυα μέσω του πίνακα συσχέτισης $\mathbf{H}$. Για ένα κατευθυνόμενο δίκτυο ο $\mathbf{H}$ είναι μη συμμετρικός, ενώ για ένα μη κατευθυνόμενο δίκτυο είναι συμμετρικός.

Με αυτόν τον ορισμό για το $Social(\mathbf{T}, \mathbf{S}, \Omega)$, το LOCABAL επιλύει το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{H}} \left\| \mathbf{W} \odot (\mathbf{R} - \mathbf{U}^T \mathbf{V}) \right\|_F^2 + \alpha \sum_{i=1}^{n} \sum_{u_k \in \mathcal{N}_i} (\mathbf{S}_{ik} - \mathbf{u}_i^T \mathbf{H} \mathbf{u}_k)^2 + \lambda (\left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{V} \right\|_F^2 + \left\| \mathbf{H} \right\|_F^2) \quad (5.15).$$

Ένα πλεονέκτημα των προσεγγίσεων αυτής της ομάδας είναι ότι μπορούν να χρησιμοποιηθούν τόσο για σύσταση όσο και για πρόβλεψη κοινωνικών σχέσεων.

Μέθοδοι συνόλου ([MKL09], [TGL12]): Η βασική ιδέα των μεθόδων συνόλου είναι ότι οι χρήστες και τα κοινωνικά τους δίκτυα θα έπρεπε να έχουν παρόμοιες βαθμολογίες για τα αντικείμενα, και μία απούσα βαθμολογία για ένα δεδομένο χρήστη μπορεί να προβλεφθεί σαν γραμμικός συνδυασμός των βαθμολογιών του χρήστη και του κοινωνικού του δικτύου. Παρουσιάζουμε δύο αντιπροσωπευτικά συστήματα παρακάτω.

STE ([MKL09]): Η βαθμολογία του i-οστού χρήστη u_i για το j-οστό αντικείμενο v_j εκτιμάται από το STE ως:

$$\hat{\mathbf{R}}_{ij} = \mathbf{u}_i^T \mathbf{v}_j + \beta \sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik} \mathbf{u}_k^T \mathbf{v}_j \quad (5.16),$$

όπου $\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik} \mathbf{u}_k^T \mathbf{v}_j$ είναι ένα σταθμισμένο άθροισμα των προβλεπόμενων βαθμολογιών (των γειτόνων του u_i) για το v_j και η παράμετρος β ελέγχει τη συνεισφορά της κοινωνικής πληροφορίας. Εύκολα επαληθεύεται ότι η προηγούμενη εξίσωση είναι ισοδύναμη με την παρακάτω:

$$\hat{\mathbf{R}} = (I + \beta \mathbf{S}) \mathbf{U}^T \mathbf{V} \quad (5.17).$$

Το STE ελαχιστοποιεί τον όρο:

$$\left\|\mathbf{W}\odot((\mathbf{R}-\mathbf{U}^T\mathbf{V})-\beta\mathbf{S}\mathbf{U}^T\mathbf{V})\right\|_F^2 = \left\|\mathbf{W}\odot(\mathbf{R}-\mathbf{U}^T\mathbf{V})\right\|_F^2+\alpha\ Social(\mathbf{T},\mathbf{S},\Omega) \quad (5.18),$$

όπου

$$Social(\mathbf{T},\mathbf{S},\Omega)=\left\|\mathbf{W}\odot(\mathbf{R}-\beta\mathbf{S}\mathbf{U}^T\mathbf{V})\right\|_F^2-2Tr(\mathbf{W}\odot((\mathbf{R}-\mathbf{U}^T\mathbf{V})(\beta\mathbf{S}\mathbf{U}^T\mathbf{V}))) \quad (5.19).$$

mTrust ([TGL12]): Μία παραλλαγή του mTrust προβλέπει τη βαθμολογία του χρήστη u_i για το αντικείμενο v_j ως:

$$\hat{\mathbf{R}}_{ij}=\mathbf{u}_i^T\mathbf{v}_j+\beta\frac{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}\mathbf{R}_{kj}}{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}} \quad (5.20),$$

όπου $\frac{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}\mathbf{R}_{kj}}{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}}$ είναι ένας σταθμισμένος μέσος των βαθμολογιών για το v_j του κοινωνικού δικτύου του u_i. Το mTrust επιλύει το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{S}}\sum_i\sum_j(\mathbf{R}_{ij}-\mathbf{u}_i^T\mathbf{v}_j-\beta\frac{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}\mathbf{R}_{kj}}{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}})^2 \quad (5.21).$$

Ομοίως προς το STE, μπορούμε να βρούμε μία standard μορφή της εξίσωσης (5.11) από την εξίσωση (5.21) για το mTrust, όπου

$$Social(\mathbf{T},\mathbf{S},\Omega)=\sum_i\sum_j\mathbf{W}_{ij}((\mathbf{R}_{ij}-\beta\frac{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}\mathbf{R}_{kj}}{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}})^2-2(\mathbf{R}_{ij}-\mathbf{u}_i^T\mathbf{v}_j)(\beta\frac{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}\mathbf{R}_{kj}}{\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}})) \quad (5.22).$$

Μέθοδοι ομαλοποίησης (regularization methods) ([JE10], [MZL+11]): Οι μέθοδοι ομαλοποίησης εστιάζουν στην προτίμηση ενός χρήστη και υποθέτουν ότι είναι παρόμοια με αυτή του κοινωνικού του δικτύου. Οι μέθοδοι ομαλοποίησης επιβάλλουν η προτίμηση $\mathbf{u}_i$ ενός χρήστη u_i να είναι κοντύτερα σε αυτήν των χρηστών του κοινωνικού δικτύου $\mathcal{N}_i$ του u_i. Δύο αντιπροσωπευτικά συστήματα αυτής της ομάδας είναι τα SocialMF ([JE10]) και Social Regularization ([MZL+11]).

SocialMF ([JE10]): Το SocialMF επιβάλλει η προτίμηση ενός χρήστη να είναι κοντύτερα στη μέση προτίμηση του κοινωνικού του δικτύου και ορίζει την ποσότητα $Social(\mathbf{T},\mathbf{S},\Omega)$ ως:

$$\min\sum_{i=1}^n(\mathbf{u}_i-\sum_{u_k\in\mathcal{N}_i}\mathbf{S}_{ik}\mathbf{u}_k)^2 \quad (5.23),$$

όπου $\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik}\mathbf{u}_k$ είναι η σταθμισμένη μέση προτίμηση των χρηστών του κοινωνικού δικτύου $\mathcal{N}_i$ του u_i και κάθε γραμμή του $\mathbf{S}$ είναι κανονικοποιημένη στη μονάδα. Οι συγγραφείς έδειξαν ότι το SocialMF αντιμετωπίζει την μεταβατικότητα της εμπιστοσύνης σε δίκτυα εμπιστοσύνης καθώς το διάνυσμα λανθανόντων χαρακτηριστικών ενός χρήστη εξαρτάται από αυτά των άμεσων γειτόνων του (το οποίο μπορεί να διαδοθεί μέσα στο δίκτυο), με αποτέλεσμα το διάνυσμα λανθανόντων χαρακτηριστικών ενός χρήστη πιθανώς να εξαρτάται από όλους τους χρήστες στο δίκτυο.

Το SocialMF επιλύει το ακόλουθο πρόβλημα βελτιστοποίησης:

$$\min_{\mathbf{U},\mathbf{V}} \left\| \mathbf{W} \odot (\mathbf{R} - \mathbf{U}^T\mathbf{V}) \right\|_F^2 + \alpha \sum_{i=1}^{n} (\mathbf{u}_i - \sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik}\mathbf{u}_k)^2 + \lambda(\left\|\mathbf{U}\right\|_F^2 + \left\|\mathbf{V}\right\|_F^2) \quad (5.24).$$

Social Regularization ([MZL+11]): Οι χρήστες στο κοινωνικό δίκτυο ενός χρήστη μπορεί να έχουν διαφορετικές προτιμήσεις. Με αυτήν τη διαισθητική ιδέα, το Social Regularization προτείνει μία ομαλοποίηση ως εξής:

$$\min \sum_{i=1}^{n} \sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik} (\mathbf{u}_i - \mathbf{u}_k)^2 \quad (5.25),$$

όπου η ομοιότητα της προτίμησης δύο συνδεδεμένων χρηστών ελέγχεται από την ομοιότητά τους με βάση τις προηγούμενες βαθμολογίες τους. Η ομοιότητα μπορεί να υπολογισθεί με το συντελεστή συσχέτισης του Pearson ή την ομοιότητα συνημιτόνου των αντικειμένων που βαθμολόγησαν δύο συνδεδεμένοι χρήστες. Μικρή τιμή του $\mathbf{S}_{ik}$ δείχνει ότι η απόσταση μεταξύ των διανυσμάτων λανθανόντων χαρακτηριστικών $\mathbf{u}_i$ και $\mathbf{u}_k$ θα πρέπει να είναι μεγαλύτερη, ενώ μεγάλη τιμή δείχνει ότι αυτή η απόσταση θα πρέπει να είναι μικρότερη. Η εξίσωση (5.25) μπορεί να γραφεί ως:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik}(\mathbf{u}_i - \mathbf{u}_k)^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{n} \mathbf{S}_{ik}(\mathbf{u}_i - \mathbf{u}_k)^2 \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{K} \mathbf{S}_{ik}(\mathbf{U}_{ij} - \mathbf{U}_{kj})^2 \qquad (5.26), \\
&= \sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{K} \mathbf{S}_{ik}\mathbf{U}_{ij}^2 - \sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{K} \mathbf{S}_{ik}\mathbf{U}_{ij}\mathbf{U}_{kj} \\
&= Tr(\mathbf{U}^T \mathcal{L} \mathbf{U})
\end{aligned}$$

όπου $\mathcal{L} = \mathbf{D} - \mathbf{S}$ ο Λαπλασιανός πίνακας και $\mathbf{D}$ ο διαγώνιος πίνακας με $\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{S}(j,i)$.

Τα συστήματα συστάσεων με κοινωνική ομαλοποίηση (social regularization) επιλύουν το παρακάτω πρόβλημα:

$$\min_{\mathbf{U},\mathbf{V}} \left\| \mathbf{W} \odot (\mathbf{R} - \mathbf{U}^T \mathbf{V}) \right\|_F^2 + \alpha \sum_{i=1}^{n} \sum_{u_k \in \mathcal{N}_i} \mathbf{S}_{ik} (\mathbf{u}_i - \mathbf{u}_k)^2 + \lambda (\left\| \mathbf{U} \right\|_F^2 + \left\| \mathbf{V} \right\|_F^2) \quad (5.27).$$

Ένα πλεονέκτημα των προσεγγίσεων αυτής της κατηγορίας είναι ότι μοντελοποιούν έμμεσα τη διάδοση (propagation) των προτιμήσεων στα κοινωνικά δίκτυα, πράγμα που μπορεί να χρησιμοποιηθεί για τη μείωση των χρηστών ψυχρής έναρξης και την αύξηση της κάλυψης των αντικειμένων προς σύσταση.

5.3.4 Δείκτες αξιολόγησης

Αν και οι είσοδοι της παραδοσιακής σύστασης και της κοινωνικής σύστασης είναι διαφορετικές, οι έξοδοί τους είναι ίδιες, δηλαδή οι προβλεπόμενες τιμές για άγνωστες βαθμολογίες. Επομένως, οι δείκτες που χρησιμοποιούνται για την αξιολόγηση παραδοσιακών συστημάτων συστάσεων μπορούν επίσης να χρησιμοποιηθούν για την αξιολόγηση κοινωνικών συστημάτων συστάσεων.

Για την αξιολόγηση των συστημάτων συστάσεων, τα δεδομένα συνήθως χωρίζονται σε δύο μέρη – το σύνολο εκπαίδευσης $\mathcal{K}$ (γνωστές βαθμολογίες) και το σύνολο δοκιμής $\mathcal{U}$ (άγνωστες βαθμολογίες). Τα συστήματα συστάσεων εκπαιδεύονται με βάση το $\mathcal{K}$ και η ποιότητα της σύστασης αξιολογείται με βάση το $\mathcal{U}$. Έχουν προταθεί διάφοροι δείκτες για την αξιολόγηση της ποιότητας της σύστασης από διάφορες προοπτικές, όπως η ορθότητα της πρόβλεψης (prediction accuracy), η ορθότητα της κατάταξης (ranking accuracy), η ποικιλία και η νεωτερικότητα (diversity and novelty) και η κάλυψη (coverage). Η ορθότητα της πρόβλεψης και η ορθότητα της κατάταξης είναι δύο ευρέως υιοθετούμενοι δείκτες.

5.3.4.1 Ορθότητα της πρόβλεψης

Ορθότητα της πρόβλεψης (prediction accuracy): Η ορθότητα της πρόβλεψης μετρά την ομοιότητα των προβλεπόμενων βαθμολογιών προς τις πραγματικές βαθμολογίες. Δύο ευρέως χρησιμοποιούμενοι δείκτες αυτής της κατηγορίας είναι το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error – MAE) και η Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error – RMSE).

Ο δείκτης RMSE ορίζεται ως:

$$RMSE = \sqrt{\frac{\sum_{(u_i, v_j) \in \mathcal{U}} (\mathbf{R}_{ij} - \hat{\mathbf{R}}_{ij})^2}{|\mathcal{U}|}} \quad (5.28),$$

όπου $|\mathcal{U}|$ είναι το μέγεθος του συνόλου $\mathcal{U}$ και $\hat{\mathbf{R}}_{ij}$ η προβλεπόμενη βαθμολογία του χρήστη u_i για το αντικείμενο v_j.

Ο δείκτης MAE ορίζεται ως:

$$MAE = \frac{1}{|\mathcal{U}|} \sum_{(u_i, v_j) \in \mathcal{U}} |\mathbf{R}_{ij} - \hat{\mathbf{R}}_{ij}| \quad (5.29).$$

Μικρή τιμή του RMSE ή του MAE σημαίνει καλύτερη επίδοση. Ας σημειωθεί ότι έχει αποδειχθεί ότι μικρή βελτίωση σε όρους RMSE ή MAE μπορεί να έχει σημαντική επίδραση στην ποιότητα της σύστασης των ανώτερων λίγων αντικειμένων (top-few recommendation) ([Kor08]).

5.3.4.2 Ορθότητα της κατάταξης

Ορθότητα της Κατάταξης (ranking accuracy): Η ορθότητα της κατάταξης αξιολογεί πόσα συνιστώμενα αντικείμενα αγοράστηκαν από το χρήστη. Η ακρίβεια (precision) και η ανάκληση (recall) είναι δύο δημοφιλείς δείκτες αυτής της κατηγορίας. Η ανάκληση συλλαμβάνει (captures) πόσα από τα αποκτημένα αντικείμενα συστήθηκαν, ενώ η ακρίβεια πόσα συνιστώμενα αντικείμενα αποκτήθηκαν. Για παράδειγμα, το Prec@n χρησιμοποιείται για να δείξει πόσα από τα N πιο συνιστώμενα (top-N recommended) αντικείμενα αποκτήθηκαν. Μακριές λίστες συστάσεων βελτιώνουν την ανάκληση ενώ μειώνουν την ακρίβεια. Έτσι, χρησιμοποιείται ο δείκτης F-score, ο οποίος τις συνδυάζει και εξαρτάται λιγότερο από το μήκος της λίστας συστάσεων .

Ένας άλλος δημοφιλής δείκτης είναι το αθροιστικό κέρδος (Discount Cumulative Gain – DCG), το οποίο ορίζεται ως:

$$DCG = \frac{1}{|\mathbf{u}|} \sum_{u_i \in \mathbf{u}} \sum_{j=1}^{|L|} \frac{\hat{\mathbf{R}}_{ij}}{\max(1, \log_b j)} \quad (5.30)$$

όπου L είναι η λίστα των συνιστώμενων αντικειμένων.

5.4 Κατευθύνσεις έρευνας στην κοινωνική σύσταση

Καθώς η επίδοση διαφέρει από πεδίο σε πεδίο, η κοινωνική σύσταση, βρίσκεται ακόμη στα πρώτα στάδια ανάπτυξης και αποτελεί μία ενεργή περιοχή έρευνας. Σε αυτήν τη ενότητα παρουσιάζουμε διάφορες κατευθύνσεις έρευνας που μπορούν να βελτιώσουν τις ικανότητες των κοινωνικών συστημάτων συστάσεων και να καταστήσουν την κοινωνική σύσταση εφαρμόσιμη σε μεγαλύτερο εύρος εφαρμογών.

5.4.1 Η ετερογένεια των κοινωνικών δικτύων

Τα περισσότερα υπάρχοντα κοινωνικά συστήματα συστάσεων μεταχειρίζονται τις συνδέσεις ενός χρήστη ομοιογενώς (homogeneously). Ωστόσο, οι συνδέσεις σε ένα online κοινωνικό δίκτυο είναι εγγενώς ετερογενείς (heterogeneous) και αποτελούνται από διάφορους τύπους σχέσεων ([TL09], [SH12], [TGL12]). Στο [TGL12] οι συγγραφείς βρήκαν ότι οι άνθρωποι δείχνουν εμπιστοσύνη με διαφορετικό τρόπο σε χρήστες σε διαφορετικούς τομείς. Για παράδειγμα, ένας χρήστης μπορεί να εμπιστεύεται έναν άλλο σε κάποιο θέμα αλλά όχι σε κάποιο άλλο. Για διαφορετικά σύνολα αντικειμένων η αξιοποίηση διαφορετικών τύπων κοινωνικών σχέσεων μπορεί να ωφελήσει τα υπάρχοντα κοινωνικά συστήματα συστάσεων ([TGL12]).

5.4.2 Συνδέσεις ασθενούς εξάρτησης

Τα περισσότερα κοινωνικά συστήματα συστάσεων που βασίζονται σε μοντέλα κάνουν χρήση μόνο των συνδέσεων ισχυρής εξάρτησης (strong dependence connections), δηλαδή των άμεσων συνδέσεων (direct connections), με αποτέλεσμα να υποτιμούν την ποικιλία των απόψεων και των προτιμήσεων των χρηστών ([Quo12]). Στο φυσικό κόσμο οι χρήστες δεν έχουν μόνο συνδέσεις ισχυρής εξάρτησης. Μπορούν να έχουν συνδέσεις ασθενούς εξάρτησης (weak dependence connections) με άλλους χρήστες με τους οποίους δεν είναι άμεσα συνδεδεμένοι. Οι συνδέσεις ασθενούς εξάρτησης μπορούν να δώσουν σημαντική πληροφορία πλαισίου (context information) για τα ενδιαφέροντα των χρηστών και έχουν αποδειχθεί χρήσιμες στην αναζήτηση εργασίας (job hunting) ([Gra73]), τη διάχυση των ιδεών ([Gra83]), τη μεταφορά γνώσης ([LC04]) και τη σχεσιακή μάθηση (relational learning) ([TL09]), ενώ σπάνια χρησιμοποιούνται στη σύσταση.

Ο προσδιορισμός των συνδέσεων ασθενούς εξάρτησης για σύσταση είναι μία ενδιαφέρουσα κατεύθυνση προς εξερεύνηση. Ένας δυνατός τρόπος προσδιορισμού των συνδέσεων ασθενούς εξάρτησης είναι η αξιοποίηση των γεωγραφικών τοποθεσιών των χρηστών. Για

παράδειγμα, το [SMML10] παρατηρεί ότι χρήστες που βρίσκονται γεωγραφικά κοντά είναι πιθανό να έχουν παρόμοια ενδιαφέροντα, ενώ το [GTL12] ότι είναι πιθανό να επισκεφθούν παρόμοιες τοποθεσίες. Οι χρήστες των online κοινωνικών δικτύων σχηματίζουν ομάδες, όπου υπάρχουν περισσότερες συνδέσεις μεταξύ χρηστών της ίδιας ομάδας από ό,τι μεταξύ χρηστών διαφορετικών ομάδων ([New05], [For10]). Σύμφωνα με τις θεωρίες κοινωνικής συσχέτισης, παρόμοιοι χρήστες αλληλεπιδρούν σε μεγαλύτερο βαθμό από ό,τι ανόμοιοι. Έτσι, χρήστες στην ίδια ομάδα είναι πιθανό να έχουν παρόμοιες προτιμήσεις, συνάπτοντας συνδέσεις ασθενούς εξάρτησης αν δε συνδέονται άμεσα ([TL09]).

5.4.3 Κατάτμηση χρηστών

Στα παραδοσιακά συστήματα συστάσεων, οι βαθμολογίες των χρηστών που είναι πιο παρόμοιοι με ένα δεδομένο χρήστη συναθροίζονται για την πρόβλεψη μίας απούσας βαθμολογίας. Όταν χρησιμοποιείται κοινωνική πληροφορία, εκτός των παρόμοιων χρηστών λαμβάνονται υπόψη και οι κοινωνικά συνδεδεμένοι. Οι πιο παρόμοιοι χρήστες με ένα δεδομένο χρήστη έχουν μικρή επικάλυψη με τους χρήστες που συνδέονται με αυτόν ([CCH+08]). Επομένως, το σύνολο των χρηστών μπορεί να κατατμηθεί σε τέσσερις ομάδες – I: συνδεδεμένοι και ανόμοιοι χρήστες, II: συνδεδεμένοι και παρόμοιοι χρήστες, III: μη συνδεδεμένοι και παρόμοιοι χρήστες και IV: μη συνδεδεμένοι και ανόμοιοι χρήστες. Ανάλογα με τον αριθμό των βαθμολογιών που έχουν λάβει, τα αντικείμενα μπορούν να κατατμηθούν σε ψυχρής έναρξης και κανονικά. Διαφορετικοί τύποι χρηστών μπορεί να συνεισφέρουν διαφορετικά για διαφορετικούς τύπους αντικειμένων. Για παράδειγμα, οι συνδεδεμένοι χρήστες μπορούν να βελτιώσουν την ορθότητα της σύστασης των τοποθεσιών ψυχρής έναρξης ([GTL12]), ενώ οι παρόμοιοι χρήστες είναι σημαντικοί για τη σύσταση κανονικών αντικειμένων ([MA04]).

5.4.4 Χρονική πληροφορία

Οι προτιμήσεις των χρηστών μεταβάλλονται με το χρόνο. Για παράδειγμα, οι προτιμήσεις των ανθρώπων που ενδιαφέρονται για τα "ηλεκτρονικά" τη στιγμή t ίσως μετακινηθούν στα "αθλητικά" τη στιγμή $t+1$. Η χρονική πληροφορία είναι σημαντικός παράγοντας στα συστήματα συστάσεων και υπάρχουν παραδοσιακά συστήματα συστάσεων που τη λαμβάνουν υπόψη ([DL05], [Kor09]). Η δυναμική στα δεδομένα μπορεί να έχει σημαντικότερη επίδραση στην ορθότητα από τη σχεδίαση πιο πολύπλοκων αλγορίθμων μάθησης ([Kor09]). Η αξιοποίηση της χρονικής πληροφορίας στα συστήματα συστάσεων

αποτελεί ακόμη μία πρόκληση λόγω της πολυπλοκότητας των χρονικών προτύπων (temporal patterns) των χρηστών ([Kor09]).

Οι κοινωνικές σχέσεις επίσης μεταβάλλονται με το χρόνο. Για παράδειγμα, νέες κοινωνικές σχέσεις προστίθενται ενώ υπάρχουσες καθίστανται ανενεργές ή διαγράφονται. Οι αλλαγές των βαθμολογιών και των κοινωνικών σχέσεων αυξάνουν περαιτέρω τη δυσκολία της αξιοποίησης της χρονικής πληροφορίας στην κοινωνική σύσταση. Μία προκαταρκτική μελέτη της επίδρασης των αλλαγών των βαθμολογιών και των σχέσεων εμπιστοσύνης στα συστήματα συστάσεων αποδεικνύει ότι η χρονική πληροφορία μπορεί να ωφελήσει την κοινωνική σύσταση ([TGLD12]).

5.4.5 Αρνητικές σχέσεις

Σήμερα, τα περισσότερα υπάρχοντα κοινωνικά συστήματα συστάσεων χρησιμοποιούν θετικές σχέσεις, όπως φιλίες και σχέσεις εμπιστοσύνης. Ωστόσο, στα κοινωνικά μέσα οι χρήστες επίσης ορίζουν αρνητικές σχέσεις, όπως δυσπιστία (distrust) και δυσαρέσκεια (dislike). Οι συγγραφείς στο [AAH13] βρήκαν ότι οι αρνητικές σχέσεις είναι ακόμη πιο σημαντικές από τις θετικές σχέσεις, αποκαλύπτοντας τη σημασία των αρνητικών σχέσεων για την κοινωνική σύσταση. Υπάρχουν διάφορες εργασίες που αξιοποιούν τη δυσπιστία ([MLK09], [VCDT09]) στα κοινωνικά συστήματα συστάσεων. Μεταχειρίζονται την εμπιστοσύνη και τη δυσπιστία ξεχωριστά και απλώς χρησιμοποιούν τη δυσπιστία με αντίθετο τρόπο προς την εμπιστοσύνη, όπως φιλτράροντας τους χρήστες στους οποίους οι άλλοι χρήστες δείχνουν δυσπιστία ή θεωρώντας τις σχέσεις δυσπιστίας ως αρνητικά βάρη. Ωστόσο, η εμπιστοσύνη και η δυσπιστία διαμορφώνονται από διαφορετικές διαστάσεις αξιοπιστίας και επηρεάζουν διαφορετικά τις προθέσεις της συμπεριφοράς η καθεμία ([Cho06]). Επιπλέον, οι σχέσεις δυσπιστίας είναι ανεξάρτητες από τις σχέσεις εμπιστοσύνης ([VDC11]). Βαθύτερη κατανόηση των αρνητικών σχέσεων και των συσχετίσεών τους με τις θετικές σχέσεις μπορεί να βοηθήσει στην ανάπτυξη αποδοτικών κοινωνικών συστημάτων συστάσεων με αξιοποίηση τόσο των θετικών όσο και των αρνητικών σχέσεων.

5.4.6 Δεδομένα από πολλά μέσα

Γενικά, ένας χρήστης έχει πολλούς λογαριασμούς στα κοινωνικά μέσα. Για παράδειγμα, ένας χρήστης που έχει λογαριασμό στο Epinions μπορεί να έχει λογαριασμό και στο eBay. Ένας νέος χρήστης σε έναν ιστότοπο μπορεί να βρισκόταν σε κάποιον άλλο ιστότοπο για πολλή ώρα. Για παράδειγμα, θεωρούμε ένα χρήστη που έχει ήδη ορίσει τα ενδιαφέροντά του στο Epinions και έχει γράψει πολλές επισκοπήσεις για αντικείμενα. Όταν ο χρήστης εγγραφεί στο

eBay για πρώτη φορά σαν χρήστης ψυχρής έναρξης, τα δεδομένα σχετικά με το χρήστη στο Epinions μπορούν να βοηθήσουν το eBay να λύσει το πρόβλημα της ψυχρής έναρξης και να συστήσει με ορθότητα αντικείμενα στο χρήστη. Η ενοποίηση δικτύων από πολλούς ιστότοπους μπορεί να έχει μεγάλη επίδραση στα κοινωνικά συστήματα συστάσεων και να προσφέρει έναν αποδοτικό και αποτελεσματικό τρόπο επίλυσης του προβλήματος της ψυχρής έναρξης. Η πρώτη δυσκολία της ενοποίησης δεδομένων είναι η σύνδεση αντίστοιχων χρηστών από πολλούς ιστοτόπους και υπάρχει πρόσφατη εργασία για την αντιμετώπιση αυτού του προβλήματος αντιστοίχισης ([NS09], [ZL09], [LZS+13]). Η μελέτη του προβλήματος αυτού καθιστά δυνατή την ενοποίηση δεδομένων από πολλά μέσα και προσφέρει νέες ευκαιρίες για τα κοινωνικά συστήματα συστάσεων.

6

Επίλογος

Στο κεφάλαιο αυτό συνοψίζονται τα συμπεράσματα που προέκυψαν από αυτήν τη διπλωματική εργασία και προτείνονται μελλοντικές επεκτάσεις της.

6.1 Σύνοψη και συμπεράσματα

Σε αυτήν την εργασία είδαμε ότι έχει προταθεί μία πληθώρα μοντέλων για τα κοινωνικά δίκτυα. Κάθε ένα από αυτά επιτυγχάνει να παραστήσει κάποια από τα χαρακτηριστικά τους, ενώ αδυνατεί να παραστήσει κάποια άλλα.

Στην πρόβλεψη συνδέσμου έχει προταθεί μία ποικιλία τεχνικών που διαφέρουν ως προς την επίδοση της πρόβλεψης, την επεκτασιμότητα (scalability) και τη γενικευσιμότητα. Οι μέθοδοι αυτές γενικά αγνοούν δομικά χαρακτηριστικά των δικτύων, όπως η ιεραρχική οργάνωση και η κοινοτική δομή, από τα οποία μπορεί να προκύψει χρήσιμη πληροφορία για την πρόβλεψη συνδέσμου.

Στην ανίχνευση κοινοτήτων έχει καταβληθεί μεγάλη προσπάθεια και έχουν προταθεί ποικίλες μέθοδοι, ωστόσο δεν έχει δοθεί ικανοποιητική λύση. Αυτό που κυρίως λείπει είναι ένα θεωρητικό πλαίσιο που να ορίζει τι πρέπει να κάνει ένας αλγόριθμος ανίχνευσης κοινοτήτων.

Η κοινωνική σύσταση έχει κατορθώσει να λύσει προβλήματα που δυσκολεύονταν να λύσουν τα παραδοσιακά συστήματα συστάσεων, όπως η αραιότητα των δεδομένων και το πρόβλημα της ψυχρής έναρξης. Ωστόσο, πρέπει να σημειωθεί ότι λόγω της ευκολίας και του μηδενικού

κόστους σύναψης online κοινωνικών σχέσεων, οι online κοινωνικές σχέσεις τις οποίες αξιοποιούν τα κοινωνικά συστήματα συστάσεων περιέχουν πολύ ‘θόρυβο’, με αποτέλεσμα η χρήση όλων των διαθέσιμων σχέσεων να οδηγεί σε χαμηλότερη επίδοση σε σχέση με τα παραδοσιακά συστήματα συστάσεων.

6.2 Μελλοντικές επεκτάσεις

Τα παρακάτω αφορούν κάθε εφαρμογή ανάλυσης κοινωνικών δικτύων, συμπεριλαμβανομένης της κοινωνικής σύστασης.

6.2.1 Δυναμικά δίκτυα

Στην εργασία αυτή θεωρήσαμε ένα στιγμιότυπο (snapshot) του κοινωνικού δικτύου. Στην πραγματικότητα, όμως, τα κοινωνικά δίκτυα εξελίσσονται με το χρόνο. Η εισαγωγή της διάστασης του χρόνου στις εφαρμογές ανάλυσης κοινωνικών δικτύων αναμένεται να βελτιώσει την επίδοση των αλγορίθμων που επιλύουν προβλήματα ανάλυσης κοινωνικών δικτύων.

6.2.2 Κατευθυνόμενα δίκτυα – δίκτυα με βάρη

Στα προηγούμενα περιοριστήκαμε στην περίπτωση που ένα κοινωνικό δίκτυο παριστάνεται ως ένας μη κατευθυνόμενος γράφος χωρίς βάρη. Ωστόσο, πολλές φορές είναι δυνατό να παρασταθεί ως ένας γράφος κατευθυνόμενος ή με βάρη. Οι κατευθύνσεις και τα βάρη των ακμών προσφέρουν πολύτιμη πληροφορία, που μπορεί να βελτιώσει σημαντικά τις λύσεις που έχουν προταθεί. Σε πολλές περιπτώσεις η επέκταση των υφιστάμενων μεθόδων στην περίπτωση των κατευθυνόμενων δικτύων ή των δικτύων με βάρη δεν είναι τετριμμένη.

6.2.3 Ετερογενή δίκτυα

Σε όλη την έκταση της εργασίας θεωρήσαμε ότι οι κόμβοι και οι ακμές των δικτύων είναι ομοειδείς. Σε πολλές, όμως, περιπτώσεις μπορούμε να θεωρήσουμε ότι οι κόμβοι ή οι ακμές ανήκουν σε διάφορες κατηγορίες (π. χ. διαφορετικοί τρόποι αλληλεπίδρασης μεταξύ των χρηστών ενός ιστότοπου κοινωνικής δικτύωσης). Η αξιοποίηση αυτής της πρόσθετης πληροφορίας στις εφαρμογές ανάλυσης κοινωνικών δικτύων αποτελεί πρόκληση.

6.2.4 Μη δομική πληροφορία

Πολλές φορές διατίθεται πληροφορία πέραν της τοπολογίας του δικτύου, όπως γνωρίσματα των κόμβων ή των ακμών. Χαρακτηριστικά παραδείγματα τέτοιας πληροφορίας είναι το περιεχόμενο που αναρτάται από χρήστες ιστότοπων κοινωνικής δικτύωσης και οι

πληροφορίες που δίνουν για τον εαυτό τους στο προφίλ τους. Η συνδυασμένη αξιοποίηση δομικής και μη δομικής πληροφορίας αναμένεται να οδηγήσει σε αποτελεσματικότερη επίλυση των προβλημάτων της ανάλυσης κοινωνικών δικτύων.

6.2.5 Επεκτασιμότητα

Η παραδοσιακή ανάλυση κοινωνικών δικτύων συχνά ασχολείται με δίκτυα μερικών εκατοντάδων κόμβων. Από την άλλη πλευρά, τα μεγαλύτερα online κοινωνικά δίκτυα αποτελούνται από πολλά εκαρομμύρια κόμβους. Επομένως, είναι σημαντική η ανάπτυξη νέων αλγορίθμων, επεκτάσιμων (scalable) στα μεγέθη που χαρακτηρίζουν τα σύγχρονα κοινωνικά δίκτυα.

7

Βιβλιογραφία

[AA03] L. A. Adamic, E. Adar. Friends and neighbors on the web. Social Networks, 25(3) pp. 211-230, 2003.

[AAH13] Abbassi, Z., Aperjis, C., Huberman, B.A.: Friends versus the crowd: tradeoffs and dynamics. HP Report (2013)

[AB12] Agarwal, V., Bharadwaj, K.: A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity. Social Network Analysis and Mining pp. 1–21 (2012)

[Agg11] Charu C. Aggarwal (Ed.): Social Network Data Analytics. Springer 2011

[AI11] Au Yeung, C., Iwata, T.: Strength of social influence in trust networks in product review sites. In: Proceedings of the fourth ACM international conference on Web search and data mining, pp. 495–504. ACM (2011)

[Alb73] Alba, R. D., A graph-theoretic definition of a sociometric clique, J. Math. Sociol. 3, 113, 1973.

[AT05] Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734–749 (2005)

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. Science, 286:509-512, 1999.

[BC92] Belkin, N.J., Croft, W.B.: Information filtering and information retrieval: two sides of the same coin? Communications of the ACM 35(12), 29–38 (1992)

[BES90] Borgatti, S., M. Everett, and P. Shirey, LS sets, lambda sets, and other cohesive subsets, Soc. Netw. 12, 337, 1990.

[BHK98] Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pp. 43–52. (1998)

[BJNR02] Barabasi, Albert-Laszlo, and Jeong, H., and Neda, Z. and Ravasz, E. Evolution of the social network of scientific collaboration. Physics A, 311(3-4):590-614. (2002)

[BP98] S. Brin, L. Page. The anatomy of a largescale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7):pp. 107-117, 1998.

[BPDA04] Boguna, M., Pastor-Satorras, R., Díaz-Guilera, A., Arenas, A. Emergence of clustering, correlations, and communities in a social network model. Physical Review E 70, 056122, 2004

[BR99] Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol.

463. ACM press New York. (1999)

[BS94] S. T. Barnard and H. D. Simon. Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. Concurrency Practice and Experience, 6(2):101-118, 1994.

[BS97] Balabanovi´c, M., Shoham, Y.: Fab: content-based, collaborative recommendation. Communications of the ACM 40(3), 66–72 (1997)

[Bur76] Burt, R. S., Positions in Networks, Soc. Forces 55, 93, 1976.

[CCH+08] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 160–168. ACM (2008)

[CCL+09] Chen, W.Y., Chu, J.C., Luan, J., Bai, H., Wang, Y., Chang, E.Y.: Collaborative filtering for orkut communities: discovery of user latent behavior. In: Proceedings of the 18th international conference on World wide web, pp. 681–690. ACM (2009)

[CGD+09] Chen, J., Geyer, W., Dugan, C., Muller, M., Guy, I.: Make new friends, but keep the old: recommending people on social networking sites. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 201–210. ACM (2009)

[CGM+99] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. In: Proceedings of ACM SIGIR workshop on recommender systems, vol. 60. Citeseer (1999)

[CGTY11] Chen, B., Guo, J., Tseng, B., Yang, J.: User reputation in a comment rating environment. In: Proceedings of the 17th ACM SIGKDD international

conference on Knowledge discovery and data mining, pp. 159–167. ACM (2011)

[Cho06] Cho, J.: The mechanism of trust and distrust formation and their relational outcomes. Journal of Retailing 82(1), 25–35 (2006)

[Cho10] Chowdhury, G.: Introduction to modern information retrieval. Facet publishing (2010)

[Chu97] F. Chung. Spectral graph theory. CBMS Regional Conference Series in Mathematics, 1997.

[CHW01] Chee, S.H.S., Han, J., Wang, K.: Rectree: An efficient collaborative filtering method. In: Data Warehousing and Knowledge Discovery, pp. 141–151. Springer (2001)

[CK01] Condon, A., and R.M. Karp, Algorithms for graph partitioning on the planted partition model, Random Struct. Algor. 18, 116, 2001

[CMN08] A. Clause, C. Moore, M. E. J. Newman. Hierarchical structure and the prediction of missing links in network. Nature, 453:pp. 98-101, 2008.

[CNM04] B.Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Physical ReviewE, 70(6):66111, 2004.

[CRRS89] Chandra, A. K., P. Raghavan, W. L. Ruzzo, and R. Smolensky, The electrical resistance of a graph captures its commute and cover times, in STOC ’89: Proceedings of the twenty-first annual ACM symposium on Theory of computing (ACM, New York, NY, USA), pp. 574-586, 1989.

[CSLF10] D. Chen, M. Shang, Z. Lv and Y. Fu. Detecting overlapping communities of weighted networks via a local algorithm. Physica A: Statistical Mechanics and its Applications, Volume 389, Issue 19, p. 4177-4187. 2010.

[CZ10] Chung, Fan, and Zhao,Wenbo,. PageRank and random walks on graphs. Proceedings of the ”Fete of Combinatorics” conference in honor of Lovasz. (2010)

[DDDA05] Danon, L., A. Diaz-Guilera, J. Duch, and A. Arenas, Comparing community structure identification, J. Stat. Mech. P09008, 2005.

[DEB02] Davidsen, J., Ebel, H., Bornholdt, S.. Emergence of a small world from local interaction: Modeling acquaintance networks. Physical Review Letters 88, 128701, 2002.

[DGK07] B. S. Dhillon, Y. Guan and B. Kulis. Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. IEEE Trans. Pattern Anal. Mach. Intell., 29(11):1944-1957, 2007.

[DK04] Deshpande, M., Karypis, G.: Item-based top-n recommendation algorithms. ACM Transactions on Information Systems 22(1), 143–177 (2004)

[DKA11] Dunlavy, D., Kolda, T., Acar, E.: Temporal link prediction using matrix and tensor factorizations. ACM Transactions on Knowledge Discovery from Data (TKDD) 5(2), 10 (2011)

[DL05] Ding, Y., Li, X.: Time weight collaborative filtering. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 485–492. ACM (2005)

[DLC13] Davis, D., Lichtenwalter, R., Chawla, N.V.: Supervised methods for multi-relational link prediction. Social Network Analysis and Mining pp. 1–15 (2013)

[DYTG09] Doppa, Janardhan R., and Yu, Jun, and Tadepalli, Prasad, and Getoor, Lise. Chance-Constrained Programs for Link Prediction. In Proceedings of

Workshop on Analyzing Networks and Learning with Graphs at NIPS Conference. (2009).

[EH08] Estrada, E., and N. Hatano, Communicability in Complex Networks, Phys. Rev. E 77(3), 036111, 2008.

[EH09] Estrada, E., and N. Hatano, Communicability Graph and Community Structures in Complex Networks, Appl. Math. Comput. 214, 500, 2009.

[EK10] David Easley and Jon Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010

[Ell08] Ellenberg, J.: This psychologist might outsmart the math brains competing for the netflix prize. Wired Magazine, March pp. 114–122 (2008)

[ER59] Erdős, P., and A. Rényi, Publ. On random graphs. Math. Debrecen 6, 290, 1959.

[FKV98] Frieze, A, and Kannan, R., and Vempala, S. Fast montecarlo algorithms for finding low-rank approximations. In Journal of the ACM (JACM), 51(6):10251041. (1998)

[FM83] Fowlkes, E. B., and C. L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78, 553, 1983.

[For10] Fortunato S., Community detection in graphs, Phys. Rep. 486(3-5):75-174 2010

[FS11] Fang, Y., Si, L.: Matrix co-factorization for recommendation with rich side information and implicit feedback. In: Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, pp. 65–69. ACM (2011)

[GAM89] Gupta, S., Anderson, R. M., and May, R. M., Networks of sexual contacts: Implications for the pattern of spread of HIV, *AIDS* 3, 807–817 (1989).

[GC11] Guy, I., Carmel, D.: Social recommender systems. In: Proceedings of the 20th international conference companion on World wide web, pp. 283–284. ACM (2011)

[Gei93] S. Geisser, Predictive inference: An introduction, Chapman and Hall, New York, 1993.

[GFKT02] Getoor, Lise, and Friedman, Nir, and Koller, Dephne, and Taskar, Benjamin. Learning Probabilistic Models of Link structure. Journal of Machine Learning Research, 3:679-707. (2002)

[GJP+10] Guy, I., Jacovi,M., Perer, A., Ronen, I., Uziel, E.: Same places, same things, same people?: mining user similarity on social media. In: Proceedings of the 2010 ACM conference on Computer supported cooperative work, pp. 41–50. ACM (2010)

[GJS+08] Guy, I., Jacovi, M., Shahar, E., Meshulam, N., Soroka, V., Farrell, S.: Harvesting with sonar: the value of aggregating social network information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1017–1026. ACM (2008)

[GN02] Girvan, M., and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99(12), 7821. 2002

[GNOT92] Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Communications of the ACM 35(12), 61–70 (1992)

[Gol06] Golbeck, J.: Generating predictive movie recommendations from trust in

social networks. Trust Management pp. 93–104 (2006)

[Gra73] Granovetter, M.: The strength of weak ties. American Journal of Socialogy 78(6), 1360–1380 (1973)

[Gra83] Granovetter, M.: The strength of weak ties: A network theory revisited. Sociological theory 1(1), 201–233 (1983)

[GRGP01] Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste: A constant time collaborative filtering algorithm. Information Retrieval 4(2), 133–151 (2001)

[GSK+99] Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J.: Combining collaborative filtering with personal agents for better recommendations. In: Proceedings of the National Conference on Artificial Intelligence, pp. 439–446. (1999)

[GTL12] Gao, H., Tang, J., Liu, H.: gscorr: Modeling geo-social correlations for new check-ins on location-based social networks. In: Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 1582–1586. ACM (2012)

[HCSZ06] Hasan, Mohammad A., and Chaoji, Vineet, and Salem, Saeed and Zaki, Mohammed. Link Prediction using Supervised Learning. In Proceedings of SDMWorkshop of Link Analysis, Counterterrorism and Security. (2006)

[HCZ+08] Hu, Y., H. Chen, P. Zhang, M. Li, Z. Di, and Y. Fan, . A New Comparative Definition of Community and Corresponding Identifying Algorithm Phys. Rev. E 78(2), 026121, 2008

[HDD13] Hong, L., Doumith, A.S., Davison, B.D.: Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In: Proceedings of the sixth ACM international conference on Web search and

data mining, pp. 557–566. ACM (2013)

[HK01] Harel, D., and Y. Koren, On Clustering Using Random Walks, in FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science (Springer-Verlag, London, UK), pp. 18-41, 2001

[HKBR99] Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 230–237. ACM (1999)

[HKTR04] J. L. Herlocker, J. A. Konstann, K. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst. 22 5. (2004)

[HL09] Z. Huang, D. K. J. Lin, The time-series link prediction problem with applications in communication surveillance, INFORMS J. Comput. 21 286. (2009)

[HM82] J. A. Hanely, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 29. (1982)

[Hof04] Hofmann, T.: Latent semantic models for collaborative filtering. ACM Transactions on Information Systems (TOIS) 22(1), 89–115 (2004)

[HRH02] P. D. Hoff, A. E. Raftery, M. S. Handcock, Latent Space Approaches to Social Network Analysis, Journal of the American Statistical Association 97 (460), 1090-1098, 2002

[HZC04] Huang, Z., Zeng, D., Chen, H.: A link analysis approach to recommendation under sparse data. In: Proc. 2004 Americas Conf. Information Systems (2004)

[JCL+12] Jiang, M., Cui, P., Liu, R., Yang, Q., Wang, F., Zhu, W., Yang, S.: Social contextual recommendation. In: Proceedings of the 22th ACM international conference on Information and knowledge management. ACM (2012)

[JE09] Jamali, M., Ester, M.: 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)

[JE10] Jamali, M., Ester, M.: A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems, pp. 135–142. ACM (2010)

[JZFF10] Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender systems: an introduction. Cambridge University Press (2010)

[KA06] Kashima, Hisashi, and Abe, Naoke. A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction. ICDM '06: Proceedings of the Sixth IEEE International Conference on Data Mining. 340-349. (2006)

[Kar01] Karypis, G.: Evaluation of item-based top-n recommendation algorithms. In: Proceedings of the tenth international conference on Information and knowledge management, pp. 247–254. ACM (2001)

[Kat53] Katz, Leo. A new status index derived from sociometric analysis. Psychometrika, 18(1):39-43. (1953)

[KK99] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing, 20, 1999.

[KL09] Kunegis, Jerome, and Lommatzsch, Andreas. Learning Spectral Graph Transformations for Link Prediction. In Proceedings of the International Conference on Machine Learning, pp 561-568. (2009)

[KL70] B. Kernighan and S. Lin. An Efficient Heuristic Procedure for partitioning graphs. The Bell System Technical J., 49, 1970.

[Kle99] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46(5), 604–632 (1999)

[KLM10] King, I., Lyu, M.R., Ma, H.: Introduction to social recommendation. In: Proceedings of the 19th international conference on World wide web, pp. 1355–1356. ACM (2010)

[Kor08] Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 426–434. ACM (2008)

[Kor09] Koren, Y.: Collaborative filtering with temporal dynamics. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 447–456. ACM (2009)

[KOS+07] Kumpula, J., Onnela, J.-P., Saramäki, J., Kaski, K., Kertész, J.. Emergence of communities in weighted networks. Physical Review Letters 99, 228701, 2007

[KR93] Klein, D. J., and M. Randic, Resistance Distance, J. Math. Chem. 12, 81. 1993

[KSS97] Kautz, H., Selman, B., Shah, M.: Referral web: combining social networks and collaborative filtering. Communications of the ACM 40(3), 63–65

(1997)

[LC04] Levin, D.Z., Cross, R.: The strength of weak ties you can trust: The mediating role of trust in effective knowledge transfer. Management science 50(11), 1477–1490 (2004)

[LHK10] J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting Positive and Negative Links in Online Social Networks, In Proceedings of WWW'2010, ACM Press, New York, 2010.

[LHZC10] Li, Y., Hu, J., Zhai, C., Chen, Y.: Improving one-class collaborative filtering by incorporating rich user information. In: Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 959–968. ACM (2010)

[Lin98] D. Lin, An information-theoretic definition of similarity, in Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufman Publishers, San Francisco, 1998.

[LK07] Liben-Nowell, David, and Kleinberg, Jon. The Link Prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology, 58(7):1019-1031. (2007).

[LKF05] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proc. Of ACM SIGKDD, pages 177-187, Chicago, Illinois, USA, 2005

[LP49] Luce, R. D., and A. D. Perry, A method of matrix analysis of group structure, Psychometrika 14(2), 95. 1949

[LS69] Luccio, F., and M. Sami, On the decomposition of networks into minimally interconnected networks, IEEE Trans. Circuit Th. CT 16, 184. 1969

[Luc50] Luce, R. D., Connectivity and generalized cliques in sociometric group

structure, Psychometrika 15(2), 169. 1950

[Lus03] Lusseau, D., The emergent properties of a dolphin social network, Proc. Royal Soc. London B 270, S186. 2003

[LZ10] L L¨u, T. Zhou, Link prediction in weighted networks: The role of weak ties, EPL 89 18001. (2010)

[LZ11] L. Lu, T. Zhou, Link Prediction in ComplexNetworks: A Survey, Physica A 390, 1150 (2011)

[LZS+13] Liu, J., Zhang, F., Song, X., Song, Y.I., Lin, C.Y., Hon, H.W.: What's in a name?: an unsupervised approach to link users across communities. In: Proceedings of the sixth ACM international conference on Web search and data mining, pp. 495–504. ACM (2013)

[MA04] Massa, P., Avesani, P.: Trust-aware collaborative filtering for recommender systems. In: On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, pp. 492–508. Springer (2004)

[MA05] Massa, P., Avesani, P.: Controversial users demand local trust metrics: An experimental study on epinions. Com community. In: Proceedings of the National Conference on Artificial Intelligence, vol. 20, p. 121. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2005)

[MA07] Massa, P., Avesani, P.: Trust-aware recommender systems. In: Proceedings of the 2007 ACM conference on Recommender systems, pp. 17–24. ACM (2007)

[Mac03] Mackay, D. J. C., , Information Theory, Inference, and Learning Algorithms (Cambridge University Press, Cambridge, UK). 2003

[Mas07] Massa, P.: A survey of trust use and modeling in real online systems. Trust

in E-services: Technologies, Practices and Challenges (2007)

[MBR98] Mooney, R.J., Bennett, P.N., Roy, L.: Book recommending using text categorization with extracted information. In: Proc. Recommender Systems Papers from 1998 Workshop, Technical Report WS-98-08 (1998)

[ME11] Menon, A., Elkan, C.: Link prediction via matrix factorization. Machine Learning and Knowledge Discovery in Databases pp. 437–452 (2011)

[Mei07] Meila, M., Comparing clusterings – an information based distance, J. Multivar. Anal. 98(5), 873. 2007

[MF93] Marsden, P., Friedkin, N.: Network studies of social influence. Sociological Methods and Research 22(1), 127–151 (1993)

[MH01] Meila, M., and D. Heckerman, An Experimental Comparison of Several Clustering and Initialization Methods, Mach. Learn. 42(1), 9. 2001

[Mil67] Stanley Milgram. The small-world problem. Psychology Today, 2:60-67, 1967.

[Mir96] Mirkin, B., Mathematical classification and clustering (Kluwer Academic Press, Norwell, USA). 1996

[MKL09] Ma, H., King, I., Lyu, M.R.: Learning to recommend with social trust ensemble. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 203–210. ACM (2009)

[MLK09] Ma, H., Lyu, M.R., King, I.: Learning to recommend with trust and distrust relationships. In: Proceedings of the third ACM conference on Recommender systems, pp. 189–196. ACM (2009)

[MLN+09] Ma, N., Lim, E., Nguyen, V., Sun, A., Liu, H.: Trust relationship prediction using online product review data. In: Proceeding of the 1st ACM international workshop on Complex networks meet information & knowledge management, pp. 47–54. ACM (2009)

[MM07] T. Murata, S. Moriyasu, Link prediction of social networks based on weighted proximity measure, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, ACM Press, New York, 2007.

[MMG+07] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, Measurement and analysis of online social networks, in 7th ACM conference on Internet measurement, pp. 29-42, 2007.

[Mok79] Mokken, R. J., Cliques, clubs and clans, Qual. Quant. 13(2), 161. 1979

[MP00] Miyahara, K., Pazzani, M.J.: Collaborative filtering with the simple Bayesian classifier. In: PRICAI 2000 Topics in Artificial Intelligence, pp. 679–689. Springer (2000)

[MSC01] McPherson, M., Smith-Lovin, L., Cook, J.: Birds of a feather: Homophily in social networks. Annual review of sociology pp. 415–444 (2001)

[MV13] F. D. Malliaros and M. Vazirgiannis. Clustering and Community Detection in Directed Networks: A Survey. Physics Reports, 533(4): 95-142, Elsevier, 2013.

[MVS04] Marsili, M., Vega-Redondo, F., Slanina, F.. The rise and fall of a networked society: A formal model. Proceedings of the National Academy of Sciences (PNAS) (USA) 101, 1439-1442, 2004

[MYH+07] Mei, T., Yang, B., Hua, X.S., Yang, L., Yang, S.Q., Li, S.: Videoreach: an online video recommendation system. In: Proceedings of the 30th annual

international ACM SIGIR conference on Research and development in information retrieval, pp. 767–768. ACM (2007)

[MYLK08] Ma, H., Yang, H., Lyu, M., King, I.: Sorec: social recommendation using probabilistic matrix factorization. In: Proceeding of the 17th ACM conference on Information and knowledge management, pp. 931–940. ACM (2008)

[MZL+11] Ma, H., Zhou, D., Liu, C., Lyu, M., King, I.: Recommender systems with social regularization. In: Proceedings of the fourth ACM international conference on Web search and data mining, pp. 287–296. ACM (2011)

[MZLK11] Ma, H., Zhou, T.C., Lyu, M.R., King, I.: Improving recommender systems by incorporating social contextual information. ACM Transactions on Information Systems 29(2), 9 (2011)

[New03a] Mark E. J. Newman. The structure and function of complex networks. SIAM Review, 45:167-256, 2003.

[New03b] Newman, M. E. J., Mixing patterns in networks, Phys.Rev. E 67, 026126 (2003).

[New04] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69(6):066133, 2004.

[New05] Newman, M.E.: Power laws, pareto distributions and zipf's law. Contemporary physics 46(5), 323–351 (2005)

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Phys. Rev. E, 69(2):026113, Feb 2004.

[NS09] Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: Security and Privacy, 2009 30th IEEE Symposium on, pp. 173–187. IEEE

(2009)

[NST+12] Noel, J., Sanner, S., Tran, K.N., Christen, P., Xie, L., Bonilla, E.V., Abbasnejad, E., Della Penna, N.: New objective functions for social collaborative filtering. In: Proceedings of the 21st international conference on World Wide Web, pp. 859–868. ACM (2012)

[Pat07] Paterek, A.: Improving regularized singular value decomposition for collaborative filtering. In: Proceedings of KDD cup and workshop, vol. 2007, pp. 5–8 (2007)

[Paz99] Pazzani, M.J.: A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review 13(5-6), 393–408 (1999)

[PB97] Pazzani, M., Billsus, D.: Learning and revising user profiles: The identification of interesting web sites. Machine learning 27(3), 313–331 (1997)

[PBMW99] Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Stanford InfoLab(1999)

[PDFV05] G. Palla, I. Derenyi, I. Farkas and T. Vicsec. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435, pp. 814-818, 2005.

[PF03] Palmer, C. R., and C. Faloutsos, Electricity Based External Similarity of Categorical Attributes, in Proceedings of PAKDD 2003, pp. 486-500. 2003

[PZC+08] Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q.: One-class collaborative filtering. In: Eighth IEEE International Conference on Data Mining, pp. 502–511. IEEE (2008)

[Quo12] Quora: Why does the startup idea of social recommendations consistently fail? In: http://www.quora.com/Why-does-the-startup-idea-of-social-recommendationsconsistently-fail (2012)

[Ran71] Rand, W. M., Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66(336), 846. 1971

[RCC+04] Radicchi, F., C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101, 2658. 2004

[RG12] B. S. Rees and K. B. Gallagher. Overlapping community detection using a community optimized graph swarm. Social Netw. Analys. Mining 405-417 2012

[RGG12] Raghavan, S., Gunasekar, S., Ghosh, J.: Review quality aware collaborative filtering. In: Proceedings of the sixth ACM conference on Recommender systems, pp. 123–130. ACM (2012)

[RIS+94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: 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)

[RRS11] Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. Recommender Systems Handbook pp. 1–35 (2011)

[Sch12] Rachel Schutt, Brief Introduction to Social Network Modeling, http://columbiadatascience.com/2012/11/02/brief-introduction-to-social-network-modeling/

[Sco11] Scott, J.: Social network analysis: developments, advances, and prospects. Social network analysis and mining 1(1), 21–26 (2011)

[Sco12] Scott, J.: Social network analysis. SAGE Publications Limited (2012)

[Sei83] Seidman, S. B., Network structure and minimum degree, Soc. Netw. 5, 269. 1983

[SF78] Seidman, S. B., and B. L. Foster, A graph-theoretic generalization of the clique concept, J. Math. Sociol. 6, 139. 1978

[SH12] Sun, Y., Han, J.: Mining heterogeneous information networks: Principles and methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery 3(2), 1–159 (2012)

[SK09] Su, X., Khoshgoftaar, T.: A survey of collaborative filtering techniques. Advances in Artificial Intelligence 2009, 4 (2009)

[SK88] Suaris, P. R., and G. Kedem, An algorithm for quadrisection and its application to standard cell placement, IEEE Trans. Circuits Syst. 35, 294. 1988

[SKKR01] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, pp. 285–295. ACM (2001)

[SKR01] Schafer, J.B., Konstan, J.A., Riedl, J.: E-commerce recommendation applications. Data mining and knowledge discovery 5(1), 115–153 (2001)

[SM08] Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. Advances in neural information processing systems 20, 1257–1264 (2008)

[SMML10] Scellato, S., Mascolo, C., Musolesi, M., Latora, V.: Distance matters: geo-social metrics for online social networks. Proceedings of WOSN 10 (2010)

[SN99] Soboroff, I., Nicholas, C.: Combining content and collaboration in text filtering. In: Proc. Intl Joint Conf. Artificial Intelligence Workshop: Machine Learning for Information Filtering (1999)

[SP09] V. Satuluri and S. Parthasarathy. Scalable graph clustering using stochastic flows: applications to community discovery. In KDD '09, pp.737-746, New York, NY, USA, 2009. ACM.

[STM11] Symeonidis, P., Tiakas, E., Manolopoulos, Y.: Product recommendation and rating prediction based on multi-modal social networks. In: Proceedings of the fifth ACMconference on Recommender systems, pp. 61–68. ACM (2011)

[SV08] Sigurbj¨ornsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th international conference on World Wide Web, pp. 327–336. ACM (2008)

[TAB09] T. Tylenda, R. Angelova, S. Bedathur, Towards time-aware link prediction in evolving social networks, In Proceedings of the 3rd Workshop on Social Network Mining and Analysis, ACM Press, New York, 2009.

[Ten99] S. H. Teng. Coarsening, sampling, and smoothing: Elements of the multilevel method. Algorithms for Parallel Processing, 105:247-276, 1999.

[TFP08] Tong, H., C. Faloutsos, and J.-Y. Pan, Random walk with restart: fast solutions and applications, Knowl. Inf. Syst. 14(3), 327. 2008

[TGHL13a] Tang, J., Gao, H., Hu, X., Liu, H.: Context-aware review helpfulness rating prediction. In: RecSys (2013)

[TGHL13b] Tang, J., Gao, H., Hu, X., Liu, H.: Exploiting homophily effect for trust prediction. In:Proceedings of the sixth ACM international conference on

Web search and data mining, pp. 53–62. ACM (2013)

[TGL12] Tang, J., Gao, H., Liu, H.: mTrust: Discerning multi-faceted trust in a connected world. In: Proceedings of the fifth ACM international conference onWeb search and data mining, pp. 93–102. ACM (2012)

[TGLD12] Tang, J., Gao, H., Liu, H., Das Sarma, A.: eTrust: Understanding trust evolution in an online world. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 253–261. ACM (2012)

[THGL13] Tang, J., Hu, X., Gao, H., Liu, H.: Exploiting local and global social context for recommendation. In: IJCAI (2013)

[THL13] Jiliang Tang, Xia Hu, Huan Liu: Social recommendation: a review. Social Netw. Analys. Mining 3(4): 1113-1133 (2013)

[TKK+09] Riitta Toivonen, Lauri Kovanen, Mikko Kivelä, Jukka-Pekka Onnela, Jari Saramäki, Kimmo Kaski: A comparative study of social network models: Network evolution models and nodal attribute models. Social Networks 31(4): 240-254 (2009)

[TL09] Tang, L., Liu, H.: Relational learning via latent social dimensions. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, p. 817-26 (2009)

[TL10] Tang, L., Liu, H.: Community detection and mining in social media. Synthesis Lectures on Data Mining and Knowledge Discovery 2(1), 1–137 (2010)

[TM69] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. Sociometry, 32(4):425-443, 1969.

[TOS+06] Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J., Kaski, K.. A model for social networks. Physica A 371(2), 851-860, 2006

[TWAK03] Tasker, Benjamin, and Wong, Ming F., and Abbeel, Pieter, and Koller, Daphne. Link Prediction in Relational Data. NIPS '03: In Proceedings of Neural Information Processing Systems. (2003).

[TWL09] L. Tang, X. Wang and H. Liu. Uncovering Groups via Heterogeneous Interaction Analysis. In IEEE International Conference on Data Mining (ICDM '09) 2009

[UF98] Ungar, L.H., Foster, D.P.: Clustering methods for collaborative filtering. In: AAAI Workshop on Recommendation Systems, 1 (1998)

[Van00] Dongen, S., Performance criteria for graph clustering and Markov cluster experiments, Technical Report, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, The Netherlands. 2000

[Váz03] Vázquez, A.. Growing networks with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Physical Review E 67, 056104, 2003

[VCDT09] Victor, P., Cornelis, C., De Cock, M., Teredesai, A.M.: A comparative analysis of trustenhanced recommenders for controversial items. In: Proc. of the International AAI Conference on Weblogs and Social Media, pp. 342–345 (2009)

[VDC11] Victor, P., De Cock, M., Cornelis, C.: Trust and recommendations. In: Recommender Systems Handbook, pp. 645–675. Springer (2011)

[VNLD10] Vasuki, V., Natarajan, N., Lu, Z., Dhillon, I.S.: Affiliation recommendation using auxiliary networks. In: Proceedings of the fourth ACM conference on

Recommender systems, pp. 103–110. ACM (2010)

[Von07] U. Von Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17(4):395-416, 2007.

[Wal83] Wallace, D. L., Method for Comparing Two Hierarchical Clusterings: Comment, J. Am. Stat. Assoc. 78, 569. 1983

[WF94] Wasserman, S., and K. Faust, Social network analysis (Cambridge University Press, Cambridge, UK). 1994

[WLJH10] Weng, J., Lim, E., Jiang, J., He, Q.: 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)

[WLWK08] Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) 26(3), 13 (2008)

[WPR06] Wong, L. H., Pattison, P., Robins, G.. A spatial model for social networks. Physica A 360, 99-120, 2006

[WS03] White, S., and P. Smyth, Algorithms for estimating relative importance in networks, in KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, New York, NY, USA), pp. 266-275. 2003

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of `small-world' networks. Nature, 393:440-442, 1998.

[WSP07] Wang, Chao, and Satuluri, Venu, and Parthasarathy, Srinivasan. Local Probabilistic Models for Link Prediction. ICDM ’07: In Proceedings of

International Conference on Data Mining. (2007)

[XKS13] Jierui Xie, Stephen Kelley and Boleslaw K. Szymanski, Overlapping Community Detection in Networks: the State of the Art and Comparative Study, ACM Computing Surveys, vol. 45, no. 4, 2013

[XNR10] Xiang, R., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: Proceedings of the 19th international conference on World wide web (2010)

[YK08] Yildirim, H., Krishnamoorthy, M.S.: A random walk method for alleviating the sparsity problem in collaborative filtering. In: Proceedings of the 2008 ACM conference on Recommender systems, pp. 131–138. ACM (2008)

[YLS+11] Yang, S.H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., Zha, H.: Like like alike: joint friendship and interest propagation in social networks. In: Proceedings of the 20th international conference on World wide web, pp. 537–546. ACM (2011)

[Yos10] T. Yoshida. Toward Finding Hidden Communities based on User Profile. IEEE International Conference on Data Mining Workshops, 2010.

[YZC+09] Yuan, Q., Zhao, S., Chen, L., Liu, Y., Ding, S., Zhang, X., Zheng, W.: Augmenting collaborative recommender by fusing explicit social relationships. In: Workshop on Recommender Systems and the Social Web, Recsys 2009 (2009)

[Zac77] W. W. Zachary. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33(4):452-473, 1977.

[ZFW+11] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams and J. Fan. Topic oriented community detection through social objects and link analysis in social networks. Knowledge-Based Systems 26: 164-173 (2012) 2011

[Zho03] Zhou, H., Distance, dissimilarity index, and network community structure. Phys. Rev. E 67(6), 061901. 2003

[ZL09] Zafarani, R., Liu, H.: Connecting corresponding identities across communities. In: Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM09) (2009)

[ZZZ10] Z.-K. Zhang, T. Zhou, Y.-C. Zhang, Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs, Physica A 389 179 (2010)