ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Περιορισμός της παραπληροφόρησης στις πλατφόρμες κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ηλιάνας Μαρίας Γ.
Ξύγκου

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Περιορισμός της παραπληροφόρησης στις πλατφόρμες κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ηλιάνας Μαρίας Γ.
Ξύγκου

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την $17^{η}$ Ιουνίου 2022.

Συμεών Παπαβασιλείου	Θεοδώρα Βαρβαρίγου	Ιωάννα Ρουσσάκη
Καθηγητής Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.	Επίκουρη Καθηγήτρια Ε.Μ.Π.

Αθήνα, Ιούνιος 2022

.....

Ηλιάνα Μαρία Γ. Ξύγκου
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Σκοπός της παρούσας Διπλωματικής Εργασίας είναι η μελέτη ενός νέου προβλήματος σχετικά με τον περιορισμό της παραπληροφόρησης σε μια πλατφόρμα κοινωνικής δικτύωσης σε συνδυασμό με την ταυτόχρονη αποφυγή της διαταραχής της διάδοσης της αληθούς πληροφορίας, και η ανάπτυξη ενός αποδοτικού αλγορίθμου επίλυσής του.

Το πρόβλημα που μελετάται στην παρούσα Εργασία και αναφέρεται ως Cautious Misinformation Minimization (CMM) ορίζεται ως η ελαχιστοποίηση της διάδοσης της ψευδούς πληροφορίας με την ταυτόχρονη ελάχιστη μείωση της διάδοσης της αληθούς περιορίζοντας τις αλληλεπιδράσεις μεταξύ των χρηστών, δηλαδή αφαιρώντας ακμές περιορισμένου πλήθους στο γράφο που αναπαριστά το υπό μελέτη δίκτυο. Επιθυμώντας την ενσωμάτωση γνωρισμάτων του χρήστη στην εξέλιξη της διάχυσης της πληροφορίας στην παρούσα Εργασία τροποποιούνται τα γνωστά μοντέλα διάδοσης Independent Cascade (IC) και Deterministic Linear Threshold (DLT), ώστε να λαμβάνεται υπόψη η εξειδίκευση του χρήστη στη θεματική κατηγορία στην οποία ανήκει η διαδιδόμενη πληροφορία. Υπό τα εν λόγω μοντέλα αλλά και το πιθανοτικό μοντέλο Linear Threshold (LT), το πρόβλημα CMM αποδεικνύεται NP-Hard. Έτσι, για την επίλυσή του υπό τα μοντέλα LT και DLT επιστρατεύονται άπληστοι επαναληπτικοί αλγόριθμοι των οποίων κριτήριο για την επιλογή της προς αφαίρεση ακμής σε κάθε επανάληψη, είναι η μέγιστη μείωση του αθροίσματος της διαφοράς της τρέχουσας διάδοσης της αληθούς πληροφορίας από την αντίστοιχη αρχική διάδοση, και της τρέχουσας διάδοσης της ψευδούς πληροφορίας.

Η πειραματική αξιολόγηση των προτεινόμενων αλγορίθμων διεξάγεται σε πραγματικά κοινωνικά δίκτυα. Τα αποτελέσματα αυτών συγκρίνονται με τα αντίστοιχα μεθόδων που αξιοποιούν κατά κύριο λόγο τοπολογικά χαρακτηριστικά και εν μέρει τη δυναμική εξέλιξη της διάδοσης της πληροφορίας. Με βάση αυτά, αναδεικνύεται η υπεροχή των προτεινόμενων μεθόδων που επιτυγχάνουν με την αφαίρεση μικρού πλήθους ακμών να μειώσουν σημαντικά την εξάπλωση της παραπληροφόρησης χωρίς να επηρεάσουν σε μεγάλο βαθμό τη διάδοση της αληθούς πληροφορίας.

Λέξεις-Κλειδιά: Ανάλυση Σύνθετων/Κοινωνικών Δικτύων, Κοινωνικά Δίκτυα, Διάχυση Πληροφορίας, Περιορισμός Παραπληροφόρησης

Abstract

The purpose of this Diploma Thesis is to study a new problem related to the limitation of misinformation spreading in a social network platform in combination with the simultaneous avoidance of the disturbance of the dissemination of true information, and the development of an efficient algorithm for its solution.

The problem under consideration in this Thesis is the one of Cautious Misinformation Minimization (CMM) which is defined as minimizing the spread of false information while minimizing the decrement of the spread of true information by limiting the interactions between the users, i.e., by removing a limited number of edges in the graph representing the network. Desiring the integration of user features in the evolution of information diffusion, the known Independent Cascade (IC) and Deterministic Linear Threshold (DLT) models are modified in this Thesis, in order to take into account the user's specialization in the thematic category to which the disseminated information belongs. Under these models as well as the probabilistic Linear Threshold model (LT), the CMM problem is proved to be NP-Hard. Thus, to solve it under the LT and DLT models, greedy iterative algorithms are employed whose criterion for selecting the edge to be removed in each iteration is the maximum reduction of the sum of the difference between true information's current spread and its initial one, and false information's current spread.

The experimental evaluation of the proposed algorithms is carried out using real social networks. Their results are compared with the ones of the methods that utilize mainly topological features and partly the dynamic evolution of information dissemination. Based on these, the superiority of the proposed methods is highlighted since they achieve through the removal of a small number of edges the significant reduction of the spread of misinformation without greatly affecting the dissemination of true information.

Keywords: Complex/Social Network Analysis, Social Networks, Information Diffusion, Misinformation Containment

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Συμεών Παπαβασιλείου για την επίβλεψη της παρούσας Διπλωματικής Εργασίας, καθώς και για το ενδιαφέρον, τις συμβουλές και τις ευκαιρίες που προσέφερε στο πλαίσιο της ακαδημαϊκής πορείας μου.

Επιπλέον, οφείλω να εκφράσω την ευγνωμοσύνη μου στη Δρ. Μαργαρίτα Βιτοροπούλου για τη συνεχή καθοδήγηση, τις ερευνητικές προτάσεις και τον πολύτιμο χρόνο που διέθεσε κατά την εκπόνηση της Εργασίας.

Τέλος, ευχαριστώ από καρδιάς την οικογένειά μου και τους φίλους μου για τη αμέριστη στήριξη που μου παρείχαν καθ'όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Σύνθετα Δίκτυα

Η οκονομική και κοινωνικοπολιτική ευημερία κατά το δεύτερο ήμισυ του $20^{ου}$ αιώνα έθεσε το πλαίσιο για τη ραγδαία ανάπτυξη της τεχνολογίας. Ένας εκ των τομέων που γνώρισε μεγάλη πρόοδο αποτελεί αυτός των δικτύων, όπως αποτυπώνεται στην πλέον απανταχού παρουσία τους με διάφορες μορφές, όπως τα Δίκτυα Επικοινωνιών, το Διαδίκτυο, τα Κοινωνικά Δίκτυα, τα Έξυπνα Ηλεκτρικά Δίκτυα (smart grids), κ.α.. Αυτά τα δίκτυα υπάγονται στην κατηγορία των Σύνθετων Δικτύων (Complex Networks) και αποτελούν το αντικείμενο μελέτης της επιστήμης της "Ανάλυσης Σύνθετων (ή πιο συγκεκριμένα Κοινωνικών) Δικτύων" (Social Network Analysis), η οποία επικεντρώνεται στην ανάπτυξη και τη δομή των δικτύων, την αλληλεπίδραση των χρηστών που τα απαρτίζουν, καθώς και τη διάδοση της πληροφορίας εντός αυτών. Η τελευταία αποτελεί σημείο εξαιρετικού ενδιαφέροντος την τελευταία δεκαετία λόγω της συμμετοχής ταχύτατα αυξανόμενου πλήθους ανθρώπων στις πλατφόρμες κοινωνικής δικτύωσης. Χαρακτηριστικά, τον Άπριλιο του 2022 περίπου το 58.7%, δηλαδή η πλειοψηφία, του παγκόσμιου πληθυσμού είναι χρήστες κοινωνικών πλατφορμών [1].

Είναι εμφανές ότι υπό αυτές τις συνθήκες μια πληροφορία μπορεί να διαχυθεί με ταχύτητα σε μεγάλο πλήθος ανθρώπων. Αν και ευνοϊκό για την άρση των οποιωνδήποτε εμποδίων απέναντι στην ενημέρωση των ατόμων, το φαινόμενο αυτό παρέχει τη δυνατότητα διάδοσης και ψευδών ειδήσεων ή αλλιώς της παραπληροφόρησης με πιθανές αρνητικές συνέπειες σε οποιαδήποτε έκφανση της ανθρώπινης κοινωνίας, όπως είναι η πολιτική και η οικονομική σταθερότητα [2]. Επομένως, κρίνεται αναγκαία η καταπολέμηση της παραπληροφόρησης, γεγονός το οποίο μπορεί να υλοποιηθεί σε δύο στάδια, τον εντοπισμό των ψευδών ειδήσεων και την ανακοπή της διάδοσής του [3]. Το πρώτο αξιοποιεί ιδέες και εργαλεία της Μηχανικής Μάθησης για τη διάκριση της αναλήθειας στις διακινούμενες πληροφορίες, ενώ το δεύτερο επιστρατεύει μεθόδους μεταβολής της δομής του δικτύου ή/και επιστράτευσης χρηστών για τη διάδοση της αντικρουόμενης στην ψευδή αληθούς πληροφορίας. Υφίσταται σημαντικός όγκος βιβλιογραφίας σχετικά με το δεύτερο βήμα, ο οποίος συγκεντρώνεται εν συντομία με κόμψο τρόπο στο [4].

Ένα γνώριμο και εύχρηστο τρόπο αναπαράστασης ενός σύνθετου δικτύου αποτελεί ο γράφος, στον οποίο οι κόμβοι και οι ακμές αναπαριστούν τους χρήστες και τις μεταξύ τους σχέσεις αντίστοιχα. Με αυτόν τον τρόπο, το πρόβλημα της παραπληροφόρησης μπορεί να αντιμετωπισθεί με αλγορίθμους αφαίρεσης κόμβων/ακμών για τη φραγή των διόδων της διάδοσης της ψευδούς είδησης ή αλγορίθμους επιλογής κατάλληλων κόμβων για την καμπάνια αλήθειας. Βέβαια, καθώς γίνεται λόγος για πλήθος κόμβων (χρηστών) που υπερβαίνει τις εκατοντάδες χιλιάδες, κρίνεται επιτακτική η εύρεση αλγορίθμων που είναι χρονικά αποδοτικοί και κλιμακώσιμοι, δηλαδή επιτυγχάνουν τον περιορισμό της παραπληροφόρησης εντός εύλογου χρονικού διαστήματος, διότι ειδάλλως η διάχυση των πληροφοριών θα έχει ολοκληρωθεί προτού επιδράσει αποτελεσματικά ο εφαρμοζόμενος αλγόριθμος.

1.2 Αντικείμενο Διπλωματικής και Συνεισφορά

Στα πλαίσια της Διπλωματικής Εργασίας προτάθηκε και μελετήθηκε ένα πρωτότυπο από όσο γνωρίζουμε πρόβλημα, κατά το οποίο αφαιρώντας ακμές (συνδέσεις) περιορισμένου πλήθους είναι επιθυμητή η μέγιστη μείωση της παραπληροφόρησης σε ένα γράφο κοινωνικής δικτύωσης σε συνδυασμό με την ελάχιστη μείωση της διάδοσης της αληθούς είδησης. Στο πρόβλημα Cautious Misinformation Minimization (CMM), όπως ονομάστηκε, διαδίδονται στο δίκτυο πληροφορίες δύο κλάσεων, I_T και I_F ανάλογα με το αν οι πληροφορίες είναι αληθείς ή ψευδείς αντίστοιχα. Ωστόσο, δεν πρόκειται για αντικρουόμενες ειδήσεις, αλλά για ανεξάρτητες, δηλαδή η υιοθέτηση μίας αληθούς πληροφορίας ούτε αναιρεί ούτε εμποδίζει την υιοθέτηση μιας ψευδούς, και αντίστροφα. Για τον ακριβή ορισμό και τη μελέτη του προβλήματος CMM χρησιμοποιήθηκαν τα δύο πιο γνωστά μοντέλα διάδοσης Independent Cascade (IC) και Linear Threshold (LT), καθώς και η ντετερμινιστική εκδοχή του δεύτερου, Deterministic Linear Threshold (DLT). Στο πρώτο και το τρίτο εξ αυτών έγινε ενσωμάτωση μιας ιδιότητας των χρηστών, της εξειδίκευσής τους στη θεματική κατηγορία στην οποία ανήκουν οι διαδιδόμενες πληροφορίες, καθώς είναι κατανοητό ότι ένας γνώστης θα αποδεχθεί μια ψευδή είδηση με μεγαλύτερη δυσκολία από έναν αδαή χρήστη.

Στη συνέχεια, αποδείχθηκε η δυσκολία (NP-Hardness) του προβλήματος υπό και τα 3 μοντέλα, και προτάθηκαν άπληστοι επαναληπτικοί αλγόριθμοι για την επίλυσή του υπό τα μοντέλα LT και DLT. Η βασική ιδέα των αλγορίθμων είναι η αφαίρεση σε κάθε επανάληψη της ακμής που συνεπάγεται τη μεγαλύτερη μείωση της αντικειμενικής συνάρτησης. Με στόχο την κλιμακωσιμότητα του αλγορίθμου υπό το μοντέλο LT αξιοποιήθηκαν οι live-edge γράφοι και τα δένδρα επιρροής που προτείνονται στα [5] και [6] αντίστοιχα, ούτως ώστε να είναι εφικτός ο (προσεγγιστικός) υπολογισμός της έκτασης της διάδοσης μιας πληροφορίας.

Τέλος, έγινε αξιολόγηση των προτεινόμενων άπληστων μεθόδων συγκρίνοντας την απόδοσή τους με αυτή άλλων 3 μεθόδων που αξιοποιούν κυρίως τα τοπολογικά χαρακτηριστικά του δικτύου αντί της δυναμικής εξέλιξης της διάδοσης της είδησης. Ως δεδομένα εισόδου χρησιμοποιήθηκαν 3 πραγματικά κοινωνικά δίκτυα με πλήθος κόμβων της τάξεως των χιλιάδων λόγω της περιορισμένης διαθέσιμης υπολογιστικής ισχύος. Τα αποτελέσματα είναι θετικά υπέρ των προτεινόμενων άπληστων αλγορίθμων και παρουσιάζονται υπό τη μορφή γραφικών παραστάσεων.

1.3 Διάρθρωση κειμένου

Η Διπλωματική Εργασία απαρτίζεται από 8 Κεφάλαια. Στα Κεφάλαια 2 έως 4 παρέχονται οι απαιτούμενες θεωρητικές γνώσεις για την κατανόηση των εννοιών και των τεχνικών που χρησιμοποιούνται στη συνέχεια. Στο Κεφάλαιο 5 γίνεται μια παρουσίαση της υπάρχουσας βιβλιογραφίας σχετικά με το γενικότερο θέμα της Παραπληροφόρησης σε πλατφόρμες κοινωνικής δικτύωσης, ενώ στο Κεφάλαιο 6 ορίζεται και αναλύεται το μελετούμενο πρόβλημα, και παρατίθενται οι προτεινόμενοι αλγόριθμοι επίλυσής του. Τέλος, στο Κεφάλαιο 7 παρουσιάζονται τα αποτελέσματα της πειραματικής εκτέλεσης των παραπάνω αλγορίθμων. Αναλυτικότερα:

- Στο Κεφάλαιο 2 αναφέρονται βασικές έννοιες της Θεωρίας Γραφημάτων, του μαθηματικού εργαλείου για την αναπαράσταση και τη μελέτη των Σύνθετων Δικτύων, συνοδευόμενες από τις αντίστοιχες ορολογίες και τους αντίστοιχους μαθηματικούς συμβολισμούς.

- Στο Κεφάλαιο 3 παρουσιάζονται μοντέλα κατασκευής γράφων που προσομοιώνουν τα Σύνθετα Δίκτυα, με ιδιαίτερη έμφαση στα Δίκτυα Ελεύθερης Κλίμακας, στα οποία αντιστοιχούν τα κοινωνικά δίκτυα. Επιπλέον, παρουσιάζονται οι βασικές κεντρικότητες κόμβου και η κεντρικότητα ενδιαμεσικότητας ακμής.

- Στο Κεφάλαιο 4 παρουσιάζονται αναλυτικά τα κατά κόρον χρησιμοποιούμενα στη βιβλιογραφία μοντέλα διάχυσης της πληροφορίας. Επίσης, γίνεται αναφορά στην επιρροή ενός κόμβου και τη δυσκολία του ακριβούς υπολογισμού του.

- Στο Κεφάλαιο 5 περιγράφεται το γενικότερο πρόβλημα της παραπληροφόρησης εντός των κοινωνικών δικτύων και παρατίθεται υπάρχουσα βιβλιογραφία σχετικά με τις μεθόδους αντιμετώπισής της.

- Στο Κεφάλαιο 6 ορίζεται το πρόβλημα Cautious Misinformation Minimization υπό τα μοντέλα διάδοσης IC, LT και DLT. Επιπλέον, αποδεικνύεται η δυσκολία επίλυσής του υπό αυτά τα μοντέλα, και προτείνονται αλγόριθμοι για την αντιμετώπισή του υπό τα μοντέλα LT και DLT.

- Στο Κεφάλαιο 7 παρέχονται και σχολιάζονται τα αποτελέσματα των πειραμάτων από την εφαρμογή των προτεινόμενων αλγορίθμων και άλλων 3 συγκριτικών μεθόδων πάνω σε πραγματικά κοινωνικά δίκτυα.

- Στο Κεφάλαιο 8 συνοψίζονται τα συμπεράσματα της Διπλωματικής Εργασίας και προτείνονται ιδέες για μελλοντική ερευνητική μελέτη.

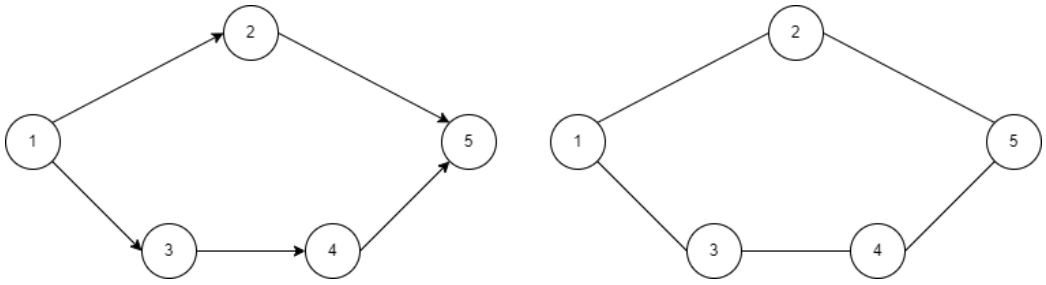
2 Στοιχεία της Θεωρίας Γραφημάτων

Η μελέτη των σύνθετων δικτύων, στα οποία υπάγονται και τα κοινωνικά δίκτυα, απαιτεί ένα μαθηματικό εργαλείο που θα επιτυγχάνει να αναπαριστά τα στοιχεία και τις μεταξύ τους αλληλεπιδράσεις. Η Θεωρία Γραφημάτων παρέχει αυτό το πλαίσιο συνοδευόμενο με πληθώρα αξιωμάτων και αλγορίθμων [7]. Παρακάτω παρουσιάζονται οι θεωρητικοί πυλώνες του συγκεκριμένου μαθηματικού τομέα καθώς και πιο εξειδικευμένες έννοιες σχετικές με το αντικείμενο της παρούσας διπλωματικής εργασίας.

2.1 Θεμελιώδεις έννοιες

Ένας γράφος (graph) ορίζεται ως ένα διατεταγμένο ζεύγος συνόλων $G = (V, E)$, όπου V το σύνολο των κόμβων (vertices) και E το σύνολο των ακμών (edges) του γράφου. Ως ακμή εννοείται ένα διατεταγμένο ζεύγος (u, v) με $u, v \in V$ υποδηλώνοντας τη σύνδεση των δύο αυτών κόμβων. Αποτελεί άμεση συνέπεια ότι $E \subseteq V^2$.

Οι γράφοι κατατάσσονται σε δύο κατηγορίες, κατευθυνόμενους (directed) ή μη κατευθυνόμενους (undirected), ανάλογα με το γεγονός αν οι ακμές έχουν ορισμένη φορά ή όχι αντίστοιχα. Για παράδειγμα, σε ένα κοινωνικό δίκτυο όπως η πλατφόρμα του Facebook όπου μεταξύ των κόμβων συνάπτονται "φιλίες" είναι σκόπιμη η αναπαράστασή του με μη κατευθυνόμενο γράφο, καθώς οι σχέσεις μεταξύ των στοιχείων είναι αμφίδρομες. Αντίθετα, σε ένα διαφορετικό κοινωνικό δίκτυο όπως το Twitter όπου ορίζονται σχέσεις ακόλουθου-ακολουθούμενου είναι αναγκαία η χρήση κατευθυνόμενων γράφων ώστε να καθίσταται σαφής η διαδρομή της διάδοσης πληροφορίας μέσα στο δίκτυο. Στο Σχήμα 1 παρουσιάζονται ένας κατευθυνόμενος και ένας μη κατευθυνόμενος γράφος, όπου οπτικά ο πρώτος ξεχωρίζει από το δεύτερο με τη χρήση βέλους στο τέλος της ακμής.



Σχήμα 1: Αριστερά ένας κατευθυνόμενος γράφος και δεξιά ο αντίστοιχος μη κατευθυνόμενος γράφος

Δύο κόμβοι u, v αποκαλούνται γείτονες (neighbors) αν προσπίπτουν σε ίδια ακμή, δηλαδή αν $(u, v) \in E$ ή $(v, u) \in E$. Στους μη κατευθυνόμενους γράφους το σύνολο των γειτόνων ενός κόμβου u συμβολίζεται ως $N(u)$, ενώ η πληθικότητα του εν λόγω συνόλου ονομάζεται βαθμός κόμβου (degree) και συμβολίζεται ως $deg(u) = |N(u)|$. Εύκολα παρατηρεί κανείς ότι $\sum_{u \in V} deg(u) = 2|E|$. Ανάλογα, στους κατευθυνόμενους γράφους υφίστανται δύο είδη βαθμών κόμβου, ο προς-τα-έσω βαθμός

(in-degreee) $deg_{in}(u)$ και ο προς-τα-έξω βαθμός (out-degree) $deg_{out}(u)$, οι οποίοι εκφράζουν το πλήθος των εισερχόμενων ακμών (ή εισερχόμενων γειτόνων $N_{in}(u)$) και το πλήθος των εξερχόμενων ακμών (ή εξερχόμενων γειτόνων $N_{out}(u)$) αντίστοιχα. Είναι εύκολα επαγόμενο ότι ισχύει $\sum_{u \in V} deg_{in}(u) = \sum_{u \in V} deg_{out}(u) = |E|$.

Μια άλλη κατηγοριοποίηση των γράφων έγκειται στην απόδοση ή μη μετρήσιμων ποσοτήτων στις ακμές του γράφου. Αυτές οι ποσότητες ονομάζονται βάρη (weights) και συμβολίζονται ως $w : E \to \mathbb{R}$. Σκοπός τους αποτελεί η έκφραση των ιδιαίτερων χαρακτηριστικών μιας ακμής, όπως είναι η "απόσταση" μεταξύ δύο κόμβων ή η "ισχύς" της σύνδεσης δύο κόμβων. Τέτοια στοιχεία διατελούν σημαντικό ρόλο στην εκτέλεση αλγορίθμων όπως είναι η εύρεση συντομότερων μονοπατιών μεταξύ κόμβων ή η διάδοση μιας πληροφορίας στο κοινωνικό δίκτυο. Σε αυτή την περίπτωση, ο γράφος αποκαλείται γράφος με βάρη (weighted graph) και συμβολίζεται ως $G = (V, E, w)$, ενώ σε αντίθετη περίπτωση ονομάζεται γράφος χωρίς βάρη (unweighted graph) και συμβολίζεται κατά τα γνωστά ως $G = (V, E)$.

Προκειμένου να δύναται κανείς να αξιοποιήσει αξιώματα και τεχνικές της Γραμμικής Άλγεβρας για την ανάλυση ενός γραφήματος, αυτό αναπαρίσταται με τη χρήση πίνακα γειτνίασης A (adjacency matrix), όπου:

$$A[u, v] = \begin{cases} 1, & \text{αν } (u, v) \in E \\ 0, & \text{αλλιώς} \end{cases}$$

στην περίπτωση ενός γράφου χωρίς βάρη, ή

$$A[u, v] = \begin{cases} w(u, v), & \text{αν } (u, v) \in E \\ 0, & \text{αλλιώς} \end{cases}$$

στην περίπτωση ενός γράφου με βάρη. Άμεσα συνεπάγεται ότι ο πίνακας ενός μη κατευθυνόμενου γράφου είναι συμμετρικός, ενώ στη συνήθη περίπτωση όπου ο γράφος δεν περιέχει ανακυκλώσεις (self-loops), δηλαδή δεν περιέχει ακμή με αρχή και τέλος τον ίδιο κόμβο, η διαγώνιος του αντίστοιχου πίνακα γειτνίασης αποτελείται μόνο από μηδενικά. Παρακάτω στο Σχήμα 2 φαίνονται οι πίνακες γειτνίασης των γράφων του Σχήματος 1.

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Σχήμα 2: Αριστερά ο πίνακας γειτνίασης του μη κατευθυνόμενου γράφου και δεξιά ο αντίστοιχος του κατευθυνόμενου γράφου.

2.2 Διαδρομή, μονοπάτια και αποστάσεις

Σε ένα γράφο $G(V, E)$ ορίζεται ως διαδρομή ή περίπατος (walk) μια ακολουθία ακμών $(e_1, ..., e_k)$ όπου για κάθε $1 \leq i \leq k-1$ ισχύει ότι το τέλος της ακμής e_i ταυτίζεται με την αρχή της ακμής e_{i+1}. Το πλήθος των ακμών αυτής της διαδρομής ονομάζεται μήκος διαδρομής, ενώ αν πρόκειται για γράφο με βάρη το μήκος της διαδρομής μπορεί να οριστεί εναλλακτικά και ανάλογα με τους σκοπούς της μελέτης ως $\sum_{1 \leq i \leq k} w(e_i)$. Επιπλέον, αν ισχύει ότι $e_i \neq e_j$, $\forall i \neq j$, δηλαδή δε γίνεται επανάληψη ακμής, τότε η διαδρομή χαρακτηρίζεται ως μονοκονδυλιά (trail), ενώ αν ισχύει $\forall e_i = (u_i, v_i), e_j = (u_j, v_j)$ με $i \neq j$ ότι $v_i \neq v_j$, δηλαδή δε γίνεται επανάληψη κορυφής, τότε η διαδρομή χαρακτηρίζεται ως μονοπάτι (path).

Αν ο αρχικός και ο τελικός κόμβος μιας διαδρομής συμπίπτουν, τότε αυτή αποκαλείται κλειστή διαδρομή. Ως επέκταση των προαναφερθέντων, μια κλειστή μονοκονδυλιά και ένα κλειστό μονοπάτι ονομάζονται κύκλος (cycle) και απλός κύκλος (simple cycle) αντίστοιχα.

Σε ένα μη κατευθυνόμενο γράφο $G(V, E)$, η ύπαρξη μονοπατιού μεταξύ δύο οποιωνδήποτε κόμβων $u, v \in V$ καθιστά το γράφο συνδεδεμένο (connected). Αντίστοιχα, σε ένα κατευθυνόμενο γράφο $G(V, E)$ η ύπαρξη μη κατευθυνόμενου μονοπατιού μεταξύ δύο οποιωνδήποτε κόμβων $u, v \in V$ συνεπάγεται ότι ο γράφος είναι αδύναμα συνδεδεμένος (weekly connected), ενώ η ύπαρξη κατευθυνόμενων μονοπατιών από τον κόμβο u στον κόμβο v και αντιστρόφως καθιστά το γράφο ισχυρά συνδεδεμένο (strongly connected).

Ως απόσταση (distance) ενός κόμβου u προς ένα κόμβο v σε ένα γράφο $G(V, E)$ καλείται το μήκος του συντομότερου μονοπατιού μεταξύ τους και συμβολίζεται ως $d(u, v)$. Εάν πρόκειται για συνδεδεμένο γράφο χωρίς βάρη αρνητικών τιμών, η απόσταση ως μετρική διέπεται από τους ακόλουθους κανόνες για $\forall u, v, w \in V$ με $u \neq v \neq w$:

1. $d(u, u) = 0$
2. $d(u, v) > 0$
3. $d(u, v) = d(v, u)$ (μόνο στην περίπτωση ενός μη κατευθυνόμενου γράφου)
4. $d(u, w) + d(w, v) \geq d(u, v)$ (τριγωνική ανισότητα)

Για τον υπολογισμό αποστάσεων μεταξύ των κόμβων ενός γράφου χρησιμοποιούνται κυρίως 3 αλγόριθμοι ανάλογα με τα χαρακτηριστικά του γράφου και τα ζητούμενα αποτελέσματα:

1. Αλγόριθμος Djikstra [8, 9], για την εύρεση συντομότερων μονοπατιών από έναν κόμβο προς όλους τους υπόλοιπους μόνο αν ο γράφος έχει θετικά βάρη. Χρονική πολυπλοκότητα $O((|V| + |E|)log|V|)$.

2. Αλγόριθμος Bellman-Ford [10, 11], για την εύρεση συντομότερων μονοπατιών από έναν κόμβο προς όλους τους υπόλοιπους ακόμα κι αν ο γράφος έχει αρνητικά βάρη. Χρονική πολυπλοκότητα $O(|E||V|)$.

3. Αλγόριθμος Floyd-Warshall [12, 13, 14], για την εύρεση συντομότερων μονοπατιών μεταξύ όλων των κόμβων. Χρονική πολυπλοκότητα $O(|V|^3)$.

Αναλυτικότερα, ο αλγόριθμος Dijkstra, του οποίου γίνεται χρήση σε επόμενο κεφάλαιο, παρουσιάζεται παρακάτω:

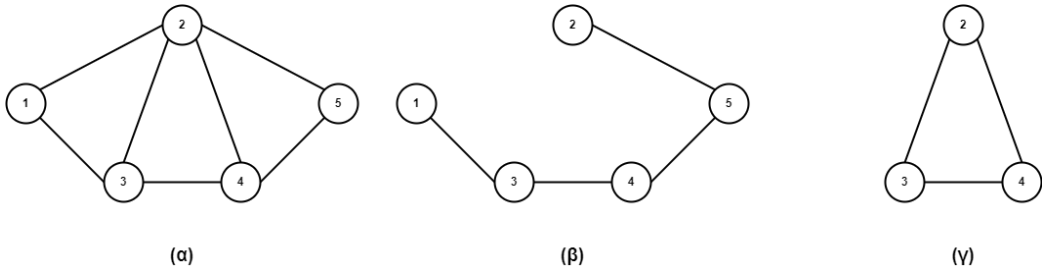
```
Input: graph G(V,E,w), vertex s
Variables: priority queue Q with keys distance[·]
Output: distances distance, paths previous

for each vertex v ∈ V:
    distance[v] = ∞
    previous[v] = NULL
distance[s] = 0
Q.enqueue(s)
while Q is not empty:
    u = Q.dequeue() // minimum distance[·]
    for each vertex v ∈ N_out(u):
        if distance[v] > distance[u] + w(u,v):
            distance[v] = distance[u] + w(u,v)
            previous[v] = u
            Q.enqueue(v) with new key distance[v]
```

2.3 Υπογράφοι

Υπογράφος $G'(V', E')$ (subgraph) ενός γράφου $G(V, E)$ ονομάζεται αυτός για τον οποίο ισχύει $V' \subseteq V$ και $E' \subseteq E$. Διακρίνονται 2 σημαντικές περιπτώσεις υπογράφων, οι οποίες παρουσιάζονται στο Σχήμα 3:

1. Αν ισχύει $V' = V$, τότε ο υπογράφος ονομάζεται επικαλύπτων (spanning).
2. Αν ισχύει $V' \subset V$ και $E' = \{(u, v) \in E : u, v \in V'\}$, τότε ο υπογράφος ονομάζεται επαγόμενος (induced).



Σχήμα 3: (α) Αρχικός γράφος $G(V, E)$, (β) Επικαλύπτων υπογράφος $G'(V, E')$, (γ) Επαγόμενος υπογράφος $G''(V'', E'')$

2.4 Διάσχιση γράφων

Ως διάσχιση γράφου (graph traversal) νοείται η εξερεύνηση της δομής ενός γράφου μέσω της επίσκεψης κόμβων ακολουθώντας ορισμένους κανόνες. Προς τούτο υφίστανται δύο προσεγγίσεις:

- Αναζήτηση κατά πλάτος (Breadth-First Search (BFS)) με χρονική πολυπλοκότητα $O(|V| + |E|)$. Τα βήματα του αλγορίθμου είναι τα ακόλουθα:

  1. Επίλεξε τυχαία ένα κόμβο $u \in V$ που δεν είναι στο σύνολο των προσπελαυμένων κόμβων *visited* και τοποθέτησέ τον στο τέλος μιας ουράς Q.
  2. Αφαίρεσε το πρώτο στοιχείο v της ουράς Q και πρόσθεσέ το στο σύνολο των προσπελαυμένων κόμβων *visited*.
  3. Για κάθε γείτονα του κόμβου v, $w \in N(v)$ (ή $N_{out}(v)$ αν εξετάζεται κατευθυνόμενος γράφος) έλεγξε αν ανήκει στο σύνολο των προσπελαυμένων κόμβων *visited* κι αν δεν ανήκει, τοποθέτησέ τον στο τέλος της ουράς Q.
  4. Επανάλαβε τα βήματα 2 και 3 έως ότου αδειάσει η ουρά Q.
  5. Επανάλαβε τα βήματα $1, 2$ και 3 έως ότου κάθε κόμβος $u \in V$ να ανήκει στο σύνολο των προσπελαυμένων κόμβων *visited*.

  Αυτό το είδος αναζήτησης βασίζεται στην διάσχιση του γράφου κατά επίπεδα, δηλαδή πρώτα προσπελαύνονται οι κόμβοι σε απόσταση 1, μετά αυτοί σε απόσταση 2 κ.ο.κ., και αποδεικνύεται εξαιρετικά χρήσιμο στην περίπτωση των κοινωνικών δικτύων όπου είναι επιθυμητή η εύρεση ατόμων που είναι "φίλοι" ή "φίλοι φίλων" κ.ο.κ. του αρχικού κόμβου, δηλαδή άτομα με πιθανά κοινά ενδιαφέροντα και χαρακτηριστικά με τον αρχικό κόμβο.

- Αναζήτηση κατά βάθος (Depth-First Search (DFS)) με χρονική πολυπλοκότητα $O(|V| + |E|)$. Τα βήματα του αλγορίθμου είναι τα ακόλουθα:

  1. Επίλεξε τυχαία ένα κόμβο $u \in V$ που δεν είναι στο σύνολο των προσπελαυμένων κόμβων *visited* και τοποθέτησέ τον στην κορυφή μιας στοίβας S.
  2. Αφαίρεσε το κορυφαίο στοιχείο v της στοίβας S και πρόσθεσέ το στο σύνολο των προσπελαυμένων κόμβων *visited*.
  3. Για κάθε γείτονα του κόμβου v, $w \in N(v)$ (ή $N_{out}(v)$ αν εξετάζεται κατευθυνόμενος γράφος) έλεγξε αν ανήκει στο σύνολο των προσπελαυμένων κόμβων *visited* κι αν δεν ανήκει, τοποθέτησέ τον στην κορυφή της στοίβας S.
  4. Επανάλαβε τα βήματα 2 και 3 έως ότου αδειάσει η στοίβα S.

5. Επανάλαβε τα βήματα 1, 2 και 3 έως ότου κάθε κόμβος $u \in V$ να ανήκει στο σύνολο των προσπελαυμένων κόμβων *visited*.

Το συγκεκριμένο είδος αναζήτησης θέτει ως προτεραιότητα την όσο πιο βαθιά διείσδυση στο γράφο, και προσομοιώνει σε μεγάλο βαθμό παιχνίδια αποφάσεων, όπως το σκάκι, κατά τα οποία κάθε κίνηση ξεκλειδώνει μερικές επόμενες κινήσεις, η καθεμία με τη δικιά τους ακολουθία επόμενων δυνατών κινήσεων.

2.5 Δένδρα

Ένα γράφημα που δεν περιέχει κύκλους ονομάζεται δάσος (forest), ενώ αν επιπλέον είναι και συνδεδεμένο καλείται δένδρο (tree). Το τελευταίο αποτελεί την πιο κοινή και χρήσιμη δομή γράφου λόγω της ιεραρχικής μορφής του. Παρακάτω παρατίθενται ισοδύναμοι ορισμοί ενός μη-κατευθυνόμενου γράφου $G = (V, E)$ ως δένδρου:

1. Δύο οποιοιδήποτε κόμβοι $u, v \in V$ συνδέονται με ένα και μοναδικό μονοπάτι.
2. Κάθε ακμή $e \in E$ είναι γέφυρα, δηλαδή η αφαίρεσή της επιφέρει την απώλεια της συνεκτικότητας του γράφου.
3. Ο γράφος G είναι συνδεδεμένος και $|E| = |V| - 1$.
4. Ο γράφος G είναι ακυκλικός και $|E| = |V| - 1$.
5. Η προσθήκη μιας ακμής δημιουργεί κύκλο στο γράφο.

Σε ένα δένδρο, οι κόμβοι με βαθμό 1 ονομάζονται φύλλα (leaves) και οι υπόλοιποι κόμβοι (βαθμός ≥ 2) ονομάζονται εσωτερικοί κόμβοι (internal nodes). Επίσης, μπορεί ένας κόμβος να οριστεί ως ρίζα (root) καθιστώντας το γράφο δένδρο με ρίζα (rooted tree). Ιδιαίτερη σημασία έχει ο ορισμός της ρίζας όταν πρόκειται για κατευθυνόμενο δένδρο με ρίζα (directed rooted tree) (δηλαδή κατευθυνόμενο ακυκλικό γράφημα με δένδρο ως υποκείμενο μη κατευθυνόμενο γράφημα), όπου ως ρίζα νοείται ο κόμβος $u \in V$ και είτε όλες οι ακμές έχουν κατεύθυνση μακριά από αυτόν (arborescence ή out-tree) ή όλες οι ακμές έχουν κατεύθυνση προς αυτόν (anti-arborescence ή in-tree). Πατέρας ενός κόμβου u ονομάζεται ο εισερχόμενος γείτονάς του και πρόγονος ενός κόμβου u καλείται οποιοσδήποτε κόμβος v συμμετέχει στο μονοπάτι από τη ρίζα προς τον κόμβο u. Αντίστοιχα, παιδί ενός κόμβου u ονομάζεται ένας εξερχόμενος γείτονάς του και απόγονος ενός κόμβου u καλείται οποιοσδήποτε κόμβος v του οποίου πρόγονος είναι ο κόμβος u.

2.6 Λοιπές κατηγορίες γράφων

Πέραν των δένδρων, υφίσταται πληθώρα γραφημάτων με ιδιαίτερα χαρακτηριστικά, όπως τα διμερή (bipartite), τα επίπεδα (planar), κ.α.. Άξιος αναφοράς είναι ο πλήρης γράφος (complete graph) ή κλίκα (clique) με συμβολισμό K_n με $n = |V|$, όπου όλοι οι κόμβοι συνδέονται με όλους τους υπόλοιπους κόμβους με μία ακμή, δηλαδή

$|E| = \frac{|V|\cdot(|V|-1)}{2}$ (μη κατευθυνόμενος γράφος) ή κάθε ζευγάρι κόμβων συνδέεται με δύο αντίθετης κατεύθυνσης ακμές, δηλαδή $|E| = |V| \cdot (|V| - 1)$ (κατευθυνόμενος γράφος). Επιπλέον, στην περίπτωση που όλοι οι κόμβοι έχουν το ίδιο βαθμό k ο γράφος ανήκει στην κατηγορία των k-κανονικών γράφων (k-regular graphs), με πληθώρα εφαρμογών σε τομείς όπως τη βιολογία ή την αεροδιαστημική μηχανική χάρη στα ιδιαίτερα τοπολογικά χαρακτηριστικά τους [15].

3 Στοιχεία των Σύνθετων Δικτύων

Ως Σύνθετα Δίκτυα νοούνται αυτά που συναντώνται πλέον κατά κόρον στην καθημερινότητα με τη μορφή δικτύων επικοινωνιών, κοινωνικών δικτύων, βιολογικών δικτύων, κ.α.. Προκειμένου να διευκολυνθεί η μελέτη τους με τη χρήση εργαλείων της Επιστήμης και Ανάλυσης Δικτύων προσομοιώνονται με τεχνητά μοντέλα που έχουν ως στόχο τη διατήρηση των χαρακτηριστικών και των ιδιοτήτων των πραγματικών δικτύων. Στα πλαίσια της μελέτης τους, γίνεται χρήση μετρικών για την κατηγοριοποίησή τους, ενώ ορίζονται και μετρικές με στόχο την εύρεση κυρίαρχων κόμβων ή/και ακμών στο δίκτυο.

Παρακάτω θα γίνει αναφορά στα επικρατέστερα μοντέλα σύνθετων δικτύων και με ποια κριτήρια ένα τυχαίο δίκτυο μπορεί να αντιστοιχισθεί σε κάποιο από αυτά. Επιπλέον, θα παρουσιαστούν οι σημαντικότερες κεντρικότητες κόμβου και ακμής, οι οποίες βασίζονται στην τοπολογία του προκύπτοντος γράφου.

3.1 Μοντέλα Προσομοίωσης Σύνθετων Δικτύων

Τα μοντέλα κατασκευής τεχνητών δικτύων χωρίζονται σε δύο κύριες κατηγορίες, στα σχεσιακά και στα χωρικά, όπου στα πρώτα οι ακμές μεταξύ των κόμβων προκύπτουν με βάση τις τοπολογικές ιδιότητές τους και στα δεύτερα με βάση τη θέση τους σε ένα ορισμένο γεωμετρικό χώρο. Ενδεικτικά, στην πρώτη κατηγορία βρίσκονται τα Δίκτυα Ελεύθερης Κλίμακας (Scale-free networks) [16], τα οποία θα αναλυθούν εκτενέστερα παρακάτω, και τα Δίκτυα Μικρού Κόσμου (Small-world networks) [17] (η απόσταση των κόμβων μεταξύ τους είναι κατά μέσο όρο σχετικά μικρή), ενώ στη δεύτερη κατηγορία ανήκουν τα Κανονικά Δίκτυα (Regular networks) [18] (όλοι ή σχεδόν όλοι οι κόμβοι έχουν τον ίδιο βαθμό) και οι Τυχαίοι Γεωμετρικοί Γράφοι (Random Geometric Graphs) [19] (οι κόμβοι τοποθετούνται τυχαία πάνω στο επίπεδο και συνδέονται μόνο με όσους βρίσκονται σε ευκλείδια απόσταση μικρότερη μιας ακτίνας ρ). Αξίζει να σημειωθεί η χρησιμότητα των Τυχαίων Γράφων (Random Graphs) [20], που ανήκουν κι αυτοί στην κατηγορία των σχεσιακών μοντέλων, για την εξαγωγή γενικών συμπερασμάτων κατά τη μελέτη της δυναμικής εξέλιξης των δικτύων, κατά τα οποία οι ακμές προκύπτουν τυχαία ανάμεσα σε οποιουσδήποτε δύο κόμβους.

3.1.1 Δίκτυα Ελεύθερης Κλίμακας

Τα Δίκτυα Ελεύθερης Κλίμακας επιτυγχάνουν σε μεγάλο βαθμό την ακριβέστερη προσομοίωση των κοινωνικών δικτύων. Προς τεκμηρίωση αυτού, γίνεται χρήση τριών βασικών μετρικών που χρησιμοποιούνται στην Ανάλυση Δικτύων [21]:

1. Κατανομή βαθμού κορυφής, $P(k) = \frac{\#\text{κόμβοι με } deg(\cdot)=k}{\#\text{κόμβοι}}$.

2. Μέσο μήκος συντομότερου μονοπατιού, $l_g = \frac{2 \cdot \sum_{u,v \in V} d(u,v)}{|V|(|V|-1)}$ στην περίπτωση ενός μη κατευθυνόμενου γράφου, ειδάλλως πρέπει να διαιρεθεί το κλάσμα με

2. Σε κάθε περίπτωση, αν ο γράφος είναι με βάρη τα οποία φέρουν σημασία για τον ορισμό της απόστασης μεταξύ των κόμβων συμπεριλαμβάνονται στον υπολογισμό της μετρικής l_g μέσω του $d(u, v)$ όπως ορίστηκε στην παράγραφο 2.2.

3. Συντελεστής Ομαδοποίησης, που δείχνει κατά πόσο τείνουν οι κόμβοι να δημιουργούν κλίκες μεταξύ τους. Υπάρχουν 3 εναλλακτικοί ορισμοί αυτού ανάλογα με τη σκοπιά μελέτης:

   (α′) Ολικός Συντελεστής Ομαδοποίησης:

   $$C = \frac{3 \times \#\text{τρίγωνα}}{\#\text{συνδεδεμένες τριάδες κόμβων}} \tag{1}$$

   (β′) Τοπικός Συντελεστής Ομαδοποίησης ενός κόμβου $u \in V$:

   $$C(u) = \frac{2 \cdot |\{(v, w) : v, w \in N(u), (v, w) \in E\}|}{deg(u) \cdot (deg(u) - 1)} \tag{2}$$

   Στην περίπτωση κατευθυνόμενου γράφου δε χρειάζεται το 2 στον αριθμητή.

   (γ′) Μέσος συντελεστής ομαδοποίησης δικτύου:

   $$\bar{C} = \frac{\sum_{u \in V} C(u)}{|V|} \tag{3}$$

Με βάση τα παραπάνω εργαλεία, ένα Δίκτυο Ελεύθερης Κλίμακας χαρακτηρίζεται από τις παρακάτω ιδιότητες, οι οποίες κατά κόρον οφείλονται στην ύπαρξη κόμβων-επικέντρων (hubs) που συγκεντρώνουν υψηλό αριθμό γειτόνων:

- Η κατανομή του βαθμού κόμβου είναι $P(k) \sim k^{-\gamma}$, όπου ο εκθέτης γ υπολογίζεται εμπειρικά και συνήθως έγκειται στο διάστημα $(2, 3)$. Με άλλα λόγια, η εν λόγω κατανομή είναι νόμου-δύναμης (power-law distribution), όπου λίγοι κόμβοι-επίκεντρα έχουν υψηλό πλήθος γειτόνων, ενώ οι περισσότεροι κόμβοι που πιθανότατα συνδέονται με τους προαναφερθέντες κόμβους με ακμή έχουν μικρό βαθμό. Αυτό αντικατοπτρίζει πολλά σενάρια των πραγματικών κοινωνικών δικτύων, αφού υπάρχουν ορισμένα λίγα διάσημα άτομα με πολλές συνδέσεις και ακολούθους, σε αντίθεση με τη πλειοψηφία του πληθυσμού που έχει λίγες συνδέσεις με οικεία άτομα.

- Το μέσο μήκος μονοπατιού είναι μικρό χάρη στην ύπαρξη των κόμβων-επικέντρων που λειτουργούν ως ενδιάμεσος σταθμός μεταξύ πολλών διαφορετικών κόμβων.

- Ο συντελεστής ομαδοποίησης είναι σχετικά υψηλός αφού οι ίδιοι οι κόμβοι-επίκεντρα που είναι γείτονες της πλειοψηφίας των κόμβων συνδέονται και μεταξύ τους δημιουργώντας κλίκες.

Όπως αναφέρθηκε προηγουμένως, τα δίκτυα ελεύθερης κλίμακας είναι σχεσιακά, δηλαδή ένας κόμβος συνδέεται με κάποιον άλλο ανάλογα με τα τοπολογικά χαρακτηριστικά τους. Στην προκειμένη περίπτωση, το χαρακτηριστικό που λαμβάνεται υπόψη είναι ο βαθμός κόμβου. Πιο αναλυτικά, σύμφωνα με το μαθηματικό μοντέλο των Barabási-Albert [16] η κατασκευή ενός τέτοιου δικτύου επιτυγχάνεται με βάση ορισμένες αρχικές συνθήκες και δύο βασικές αρχές:

1. Αρχικές Συνθήκες: Τη χρονική στιγμή $t = 1$ το δίκτυο αποτελείται από m_0 κόμβους.

2. Εξέλιξη (Growth): Σε $\forall t > 1$ ένας νέος κόμβος προστίθεται στο δίκτυο και συνδέεται με $m \leq m_0$ διαφορετικούς κόμβους.

3. Επιλεκτική σύνδεση (Preferential Attachment): Ο νεοεισερχόμενος κόμβος w συνδέεται με τον προϋπάρχοντα κόμβο u με πιθανότητα:

$$\Pi(u) = \frac{deg(u)}{\sum_{v \in V} deg(v)} \tag{4}$$

   Αυτό έχει ως αποτέλεσμα τη δημιουργία κόμβων-επικέντρων, αφού εμφανώς ένας κόμβος με πολλούς γείτονες είναι πιο πιθανό να αποκτήσει ένα νέο γείτονα σε σύγκριση με κάποιον με λιγότερους γείτονες από αυτόν.

Εφαρμόζοντας το παραπάνω μοντέλο καταλήγει κανείς μετά από χρονικά βήματα t σε ένα δίκτυο $G = (V, E)$ με $|V| = m_0 + t, |E| = m \cdot t$ και με κατανομή βαθμού κόμβου $P(k) \sim k^{-\gamma}$ με $\gamma = 2.9 \pm 0.1$.

3.2 Κεντρικότητες

Οι κεντρικότητες αποτελούν μετρικές που επιχειρούν να ποσοτικοποιήσουν τη σπουδαιότητα ή την επιρροή ενός κόμβου ή μιας ακμής μέσα στο δίκτυο. Σε αυτή την ενότητα θα παρουσιαστούν συνοπτικά οι σημαντικότερες κεντρικότητες κόμβων, η καθεμία εκ των οποίων αποδίδει βάρος σε διαφορετικά χαρακτηριστικά του κόμβου [22]. Επιπλέον, θα αναλυθεί η ιδέα και ο τρόπος υπολογισμού της κεντρικότητας ενδιαμεσικότητας ακμής.

3.2.1 Κεντρικότητες κόμβων

Κάτωθι παρατίθενται εν συντομία οι κυριότερες κεντρικότητες κόμβων:

1. Κεντρικότητα βαθμού (Degree Centrality) ενός κόμβου u:

$$C_D(u) = \frac{deg(u)}{|V| - 1} \tag{5}$$

   Σημαντικός θεωρείται ένας κόμβος με πολλούς γείτονες (ή "φίλους"), αφού έχει τη δυνατότητα να επηρεάσει μεγάλο πλήθος κόμβων άμεσα.

2. Κεντρικότητα εγγύτητας (Closeness Centrality) ενός κόμβου u [23]:

$$C_P(u) = \frac{|V| - 1}{\sum_{v \in V, v \neq u} d(u, v)} \tag{6}$$

Σημαντικός θεωρείται ένας κόμβος που βρίσκεται σε μικρή απόσταση από τους υπόλοιπους κόμβους, οπότε δίνεται έμφαση στη χωρική κυριαρχία του κόμβου και την ισχύ του να διαδώσει γρήγορα κάποια πληροφορία στο υπόλοιπο δίκτυο.

3. Κεντρικότητα ενδιαμεσικότητας (Betweenness Centrality) ενός κόμβου u [24]:

$$C_B(u) = \frac{2 \cdot \sum_{s \neq u \neq t} \frac{\sigma_{st}(u)}{\sigma_{st}}}{(|V| - 1)(|V| - 2)}, \tag{7}$$

όπου σ_{st} και $\sigma_{st}(u)$ το πλήθος των συντομότερων μονοπατιών μεταξύ των κόμβων s και t, και το πρότερο στα οποία συμμετέχει ο κόμβος u αντίστοιχα. Αν πρόκειται για κατευθυνόμενο γράφο, το 2 στον αριθμητή παραλείπεται.

Εν προκειμένω, η σπουδαιότητα ενός κόμβου κρίνεται από την παρουσία του στα συντομότερα μονοπάτια μεταξύ των κόμβων ενός γράφου και κατά συνέπεια, τη δυνατότητα του να ελέγχει τη διάδοση μιας πληροφορίας εάν αυτή γίνεται μέσω της συντομότερης διαδρομής.

4. Κεντρικότητα ιδιοδιανύσματος (Eigenvector Centrality) ενός κόμβου u [25]:

$$v(u) = \frac{1}{\lambda} \sum_{k \in V} A[u, k] v(k), \tag{8}$$

όπου A ο πίνακας γειτνίασης του γράφου, λ η μεγαλύτερη ιδιοτιμή αυτού και v το αντίστοιχο ιδιοδιάνυσμα. Σε αυτή την περίπτωση, ένας κόμβος είναι σημαντικός αν οι γείτονές του είναι σημαντικοί.

3.2.2 Κεντρικότητα ενδιαμεσικότητας ακμής

Σε απόλυτη αντιστοιχία με αυτή του κόμβου, η κεντρικότητα ενδιαμεσικότητας (Edge Betweenness Centrality) μιας ακμής e υπολογίζεται ως:

$$C_B(e) = \frac{2 \cdot \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}}{|V|(|V| - 1)} \tag{9}$$

Ο ορισμός αυτός δόθηκε στα πλαίσια της έρευνας εντοπισμού κοινοτήτων εντός ενός κοινωνικού δικτύου από τους Girvan-Newman [26] με στόχο τον εντοπισμό ακμών που με μεγάλη πιθανότητα είναι γέφυρες, η αφαίρεση των οποίων θα επέφερε τη διάσπαση του δικτύου σε 2 συνδεδεμένες συνιστώσες ή τουλάχιστον θα δυσχαίρενε την επικοινωνία μεταξύ των κόμβων του γράφου. Για τον υπολογισμό της εν λόγω μετρικής εφαρμόζεται ο αλγόριθμος του Brandes [27, 28], ο οποίος είναι ο πλέον αποδοτικός με χρονική πολυπλοκότητα $O(|V| \cdot |E|)$ και $O(|V| \cdot |E| + |V|^2 log|V|)$

για γράφους χωρίς και με βάρη αντίστοιχα. Παρακάτω, παρουσιάζεται σε μορφή ψευδοκώδικα ο εν λόγω αλγόριθμος για κατευθυνόμενους γράφους με βάρη που θα χρησιμοποιηθεί και στη συνέχεια της παρούσας εργασίας:

```
Input: Graph G(V, E, w)
Variables: Priority queue Q with keys dist[·], stack S,
   distance from source dist[v], list of predecessors on
   shortest paths from source pred[v], number of shortest
   paths from source to v σ[v], dependency of source on v δ[v]
Output: Betweenness centrality ∀u ∈ V ∪ E

for each vertex s ∈ V:
    // find single-source shortest paths
    for each vertex w ∈ V:
        pred[w] = []
    for each vertex t ∈ V:
        dist[t] = ∞
        σ[t] = 0
    dist[s] = 0
    σ[s] = 1
    Q.enqueue(s)
    while Q is not empty:
        v = Q.dequeue() // minimum dist[·]
        S.push(v)
        for each vertex z ∈ N_out(v):
            if dist[z] > dist[v] + w(v, z):
                dist[z] = dist[v] + w(v, z)
                Q.enqueue(z) with new key dist[z]
                σ[z] = 0
                pred[z] = []
            if dist[z] = dist[v] + w(v, z):
                σ[z] = σ[z] + σ[v]
                pred[z].append(v)
    // collect paths
    for each vertex v ∈ V:
        δ[v] = 0
        c_B[v] = 0
    for each edge e ∈ E:
        c_B[e] = 0

    while S is not empty:
        z = S.pop()
        for each vertex v ∈ pred[z]:
            c = (σ[v]/σ[z]) · (1 + δ[z])
            c_B[(v, z)] = c_B[(v, z)] + c
            δ[v] = δ[v] + c
        if z ≠ s:
            c_B[z] = c_B[z] + δ[z]
```

4 Διάδοση Πληροφορίας

Με τη συμμετοχή δισεκατομμυρίων ανθρώπων στα κοινωνικά δίκτυα, αυτά αποτελούν το πλέον σημαντικό πλαίσιο για τη διάδοση ιδεών, πληροφοριών, συμπεριφορών αλλά και δυσάρεστων στοιχείων, όπως ιών. Επομένως, κρίνεται επιτακτική η μελέτη και κυρίως η πρόβλεψη του τρόπου διάχυσης μιας πληροφορίας στο δίκτυο, με στόχο είτε τον εντοπισμό των σημαντικών κόμβων που ασκούν μεγάλη επιρροή στους υπολοίπους είτε την έγκαιρη αποτροπή εξάπλωσης ενός προβλήματος. Αυτή η ανάλυση καθίσταται εφικτή μέσω μοντέλων που προσομοιώνουν τη διάδοση στοιχείων στο δίκτυο.

Σε αυτό το κεφάλαιο θα παρουσιαστούν τα δύο επικρατέστερα μοντέλα διάχυσης πληροφορίας και μετρικές που αποσκοπούν στην ποσοτικοποίηση της ισχύος-επιρροής ενός κόμβου.

4.1 Μοντέλα Διάχυσης

Τα μοντέλα διάχυσης αποτελούν το εργαλείο μελέτης και πρόβλεψης της εξάπλωσης μιας πληροφορίας δοθέντος ενός γράφου $G(V, E)$ που αναπαριστά το κοινωνικό δίκτυο, και κάποιων αρχικών ενστερνιστών κόμβων (seed nodes) μιας ιδέας που θα αποτελέσουν την αρχή της διάδοσης (όπως θα ήταν αντίστοιχα ο ασθενής Μηδέν σε μια επιδημία). Τα δύο κυριότερα εξ αυτών που συναντώνται κατά κόρον στη βιβλιογραφία είναι το Independent Cascade Model (IC) και το Linear Threshold Model (LT) τα οποία πρώτα προτάθηκαν στα [29, 30] και [31, 32] αντίστοιχα, και μελετήθηκαν στο επιθυμητό πλαίσιο στο [5]. Πάνω σε αυτά έχουν γίνει αρκετές επεκτάσεις και εξειδικεύσεις ανάλογα με τον τύπο του μελετούμενου δικτύου και είδους πληροφορίας, και τον επιδιωκόμενο στόχο. Εκτός αυτών, δύνανται να χρησιμοποιηθούν και επιδημιολογικά μοντέλα για την προσομοίωση αυτής της διαδικασίας, αντιμετωπίζοντας την πληροφορία ή ιδέα ως έναν ιό.

4.1.1 Independent Cascade Model

Σύμφωνα με αυτό το μοντέλο, ένα κοινωνικό δίκτυο αναπαρίσταται με τη μορφή ενός κατευθυνόμενου γράφου με βάρη $G(V, E, p)$. Τα βάρη εκφράζουν την πιθανότητα και το βαθμό επιρροής ενός κόμβου σε έναν άλλο. Έτσι, μαθηματικά ορίζονται ως $p : E \to [0, 1]$. Κατά τη διάρκεια της εξέλιξης της διαδικασίας οι κόμβοι διαχωρίζονται σε δύο κατηγορίες, τους ενεργούς και τους ανενεργούς, με τους πρώτους να είναι αυτοί που έχουν υιοθετήσει τη διαδιδόμενη ιδέα και θα παραμείνουν ες αεί σε αυτή την κατάσταση, και τους δεύτερους να είναι αυτοί που δεν έχουν προσβληθεί από αυτή αλλά είναι ευάλωτοι να τη λάβουν από κάποιον ενεργό γείτονά τους. Το μοντέλο συνοψίζεται ως εξής: Θεωρείται ένα σύνολο αρχικών ενστερνιστών A (seed set). Θεωρώντας διακριτά χρονικά βήματα $t \in \mathbb{Z}$, σε κάθε χρονική στιγμή t κάθε κόμβος v που γίνεται για πρώτη φορά ενεργός επιχειρεί για μία και μοναδική φορά να ενεργοποιήσει κάθε εξερχόμενο γείτονά του w. Η προσπάθειά του είναι επιτυχής με ανεξάρτητη πιθανότητα $p(v, w)$, ενώ αν ο κόμβος w έχει περισσότερους από ένα πρώτη

φορά ενεργούς εισερχόμενους γείτονες, οι προσπάθειες ενεργοποίησής του γίνονται από αυτούς με αυθαίρετη σειρά. Σε περίπτωση επιτυχούς ενεργοποίησης του κόμβου w, αυτός θα μεταβεί σε ενεργή κατάσταση τη χρονική στιγμή $t+1$. Η διαδικασία σταματάει τη χρονική στιγμή t', κατά την οποία κανένας κόμβος δε μεταβαίνει σε ενεργή κατάσταση.

4.1.2 Linear Threshold Model

Σύμφωνα με αυτό το μοντέλο, ένα κοινωνικό δίκτυο αναπαρίσταται με τη μορφή ενός κατευθυνόμενου γράφου με βάρη $G(V, E, b)$. Τα βάρη εκφράζουν το βαθμό επιρροής ενός κόμβου σε έναν άλλο, μαθηματικά συμβολίζονται ως $b : E \to [0, 1]$ και υπόκεινται στον περιορισμό $\sum_{u \in N_{in}(v)} b(u, v) \leq 1$ για $\forall v \in V$, όπου $N_{in}(v) = \{u : (u, v) \in E\}$. Με άλλα λόγια, το άθροισμα των βαρών των εισερχόμενων ακμών σε ένα κόμβο είναι το πολύ 1 για λόγους που θα φανούν παρακάτω. Όπως και στο προηγούμενο μοντέλο, οι κόμβοι διακρίνονται σε 2 κατηγορίες, ενεργούς και ανενεργούς με τα ίδια προηγούμενα χαρακτηριστικά. Επιπλέον, κάθε κόμβος u επιλέγει τυχαία σύμφωνα με την ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ ένα κατώφλι $\theta(u)$, το οποίο εκφράζει το δισταγμό του κόμβου να υιοθετήσει μια ιδέα των γειτόνων του. Το μοντέλο συνοψίζεται ως ακολούθως: Θεωρείται ένα σύνολο αρχικών ενστερνιστών A (seed set) και μια τυχαία επιλογή κατωφλίων $\theta(u) : u \in V$. Θεωρώντας διακριτά χρονικά βήματα $t \in \mathbb{Z}$ σε κάθε χρονική στιγμή t κάθε κόμβος u ενεργοποιείται αν ισχύει το ακόλουθο:

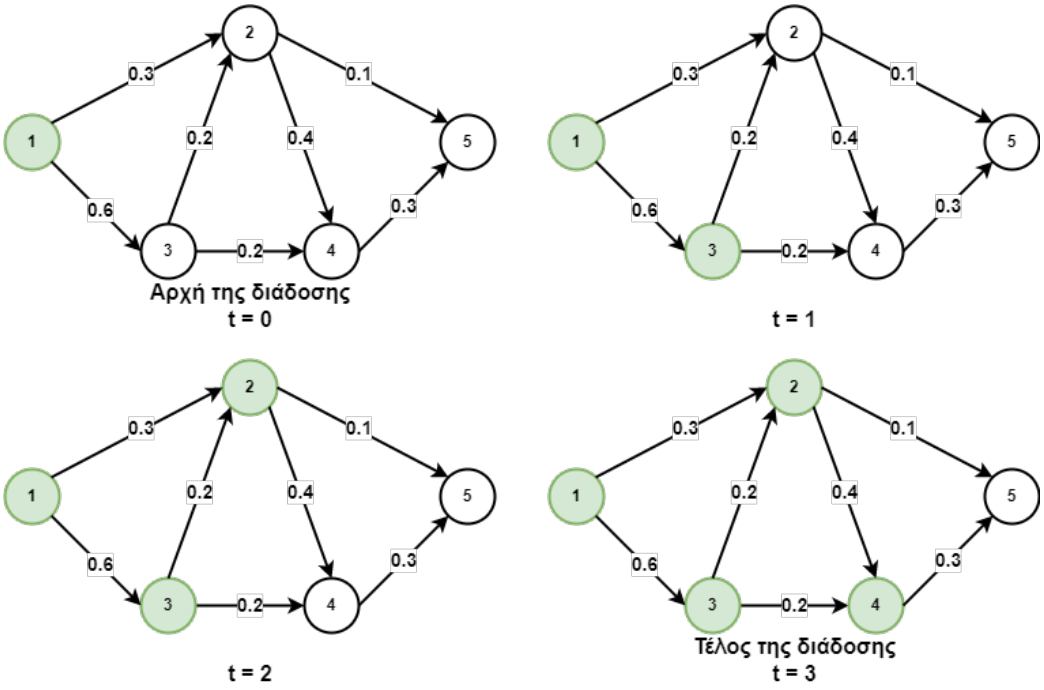
$$\sum_{w \in N_{in}(u)} b(w, u) \geq \theta(u) \tag{10}$$

Η διαδικασία σταματάει τη χρονική στιγμή t', κατά την οποία κανένας κόμβος δε μεταβαίνει σε ενεργή κατάσταση. Η τυχαία επιλογή των τιμών των κατωφλίων αποδίδεται στην απουσία a priori γνώσης για τους ενδοιασμούς ενός κόμβου, ενώ παρέχει το πλαίσιο και για την εξαγωγή γενικότερων συμπερασμάτων που θα αναφερθούν σε επόμενο κεφάλαιο.

Στο Σχήμα 4 φαίνεται ένα παράδειγμα διάδοσης πληροφορίας σύμφωνα με το μοντέλο LT, όπου $\theta(v) = 0.5, \forall v \in V$. Πάνω στις ακμές έχουν τοποθετηθεί τα αντίστοιχα βάρη, ενώ με πράσινο χρωματίζονται οι κόμβοι που είναι ενεργοί τη συγκεκριμένη χρονική στιγμή.

4.1.3 Λοιπά σχετικά μοντέλα

Πέραν των προαναφερθέντων και πολυμελετημένων μοντέλων, έχουν προταθεί και ερευνηθεί επεκτάσεις αυτών. Για παράδειγμα, στο [33] προτείνεται το μοντέλο Multiple Factors-Aware Diffusion, το οποίο βασιζόμενο στο μοντέλο IC προσθέτει ένα πιθανοτικό ταξινομητή $f_v(x)$ για κάθε κόμβο $v \in V$ με x το διάνυσμα χαρακτηριστικών της έκθεσης στην είδηση, έτσι ώστε να λαμβάνονται υπόψη χαρακτηριστικά του κόμβου που συνθέτουν την τάση του να αποδεχθεί μια πληροφορία. Έτσι, η μόνη αλλαγή στην εξέλιξη της εξάπλωσης σε σχέση με το μοντέλο IC είναι ότι η



Σχήμα 4: Εξέλιξη της διάδοσης μιας πληροφορίας σύμφωνα με το μοντέλο LT, όπου $\theta(u) = 0.5, \forall u \in V$.

πιθανότητα του πρόσφατα ενεργοποιημένου κόμβου u να μεταβάλει την κατάσταση ενός ανενεργού εξερχόμενου γείτονά του v είναι πλέον $p(u, v) \cdot f_v(x)$. Με αυτό τον τρόπο, ενσωματώνονται στη μελέτη και τα ιδιαίτερα χαρακτηριστικά κάθε κόμβου. Παράλληλα, στο [34] μελετάται το Ντετερμινιστικό LT μοντέλο, κατά το οποίο το κατώφλι ενός κόμβου u, $\theta(u)$ πλέον δεν επιλέγεται τυχαία, αλλά με βάση κάποια πρότερη γνώση. Μια πιο εξελιγμένη εκδοχή του μοντέλου LT παρουσιάζεται στο [35] με όνομα Δυναμικό LT μοντέλο (DLT), σύμφωνα με το οποίο στο δίκτυο διαδίδονται αντικρουόμενες ιδέες και η πιθανότητα επιρροής $p(\cdot, \cdot)$ και το κατώφλι $\theta(\cdot)$ ενός κόμβου μετά την υιοθέτηση μιας ιδέας είναι χρονικά μεταβλητά. Επιπλέον, είναι δυνατή η ανάκληση μιας ιδέας ανάλογα με την ασκούμενη επιρροή των ενεργών εισερχόμενων γειτόνων.

Άξια αναφοράς είναι η χρήση επιδημιολογικών μοντέλων για την προσομοίωση της διάδοσης μιας πληροφορίας αντιμετωπίζοντας την ιδέα ως ιό. Σύμφωνα με αυτά, ένας κόμβος-άτομο μπορεί να βρίσκεται σε μια εκ των παρακάτω καταστάσεων:

- Susceptible (S), επιρρεπής στην αποδοχή της πληροφορίας,
- Infected (I), ενστερνιστής της πληροφορίας,
- Removed (R), ενστερνιστής της ιδέας που την ανακαλεί και προσωρινά αφαιρείται από τη διάδοση της πληροφορίας,

- Dead (D), έχει απόλυτη ανοσία απέναντι στην πληροφορία και αφαιρείται από το δίκτυο,
- Passive Immunity (M), έχει προσωρινή ανοσία στην πληροφορία (αντίστοιχα έχει γεννηθεί με αντισώματα απέναντι στον ιό),
- Exposed (E), έχει εκτεθεί στην πληροφορία αλλά δεν μπορεί ακόμα να τη μεταδώσει.

Σε αυτά τα μοντέλα η μελέτη γίνεται με χρήση διαφορικών εξισώσεων, και οριακών και αρχικών συνθηκών. Στο Σχήμα 5 παρουσιάζεται το γενικό επιδημιολογικό μοντέλο MSEIR με τις κάτωθι εξισώσεις:

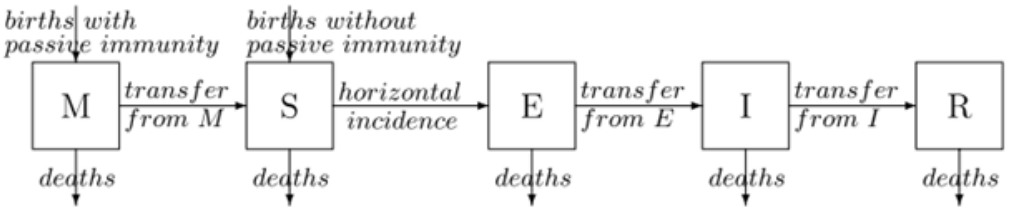
$$\begin{cases} \dfrac{dM}{dt} = b(N - S(t)) - (\delta + b)M(t), \\ \dfrac{dS}{dt} = bS(t) + \delta M(t) - \dfrac{\beta}{N}S(t)I(t) - bS(t), \\ \dfrac{dE}{dt} = \dfrac{\beta}{N}S(t)I(t) - (\epsilon + b)E(t), \\ \dfrac{dI}{dt} = \epsilon E(t) - (\gamma + b)I(t), \\ \dfrac{dR}{dt} = \gamma I(t) - bR(t), \end{cases} \tag{11}$$

όπου

- b, ο ρυθμός γέννησεων και ο ρυθμός θανάτων,
- $N = M(t) + S(t) + E(t) + I(t) + R(t)$, ο αριθμός πληθυσμού που παραμένει σταθερός,
- δ, ο ρυθμός απώλειας της παθητικής ανοσίας,
- β, ο ρυθμός μετάδοσης της πληροφορίας,
- ϵ, ο ρυθμός καταστάλαξης της πληροφορίας,
- γ, ο ρυθμός ανάκλησης της πληροφορίας.

4.2 Ισχύς και επιρροή

Όπως είναι φυσικό, λόγω τόσο της τοπολογίας του δικτύου όσο και των ιδιαίτερων χαρακτηριστικών των κόμβων και των σχέσεων μεταξύ τους, δεν ασκούν όλοι οι κόμβοι την ίδια δύναμη στο υπόλοιπο δίκτυο. Ο εντοπισμός των ισχυρών κόμβων αποτελεί σημαντικό κομμάτι της έρευνας του τμήματος μάρκετινγκ σε πολλές επιχειρήσεις για την επιτυχή προώθηση προϊόντων, αλλά και σε άλλους τομείς όπως την πολιτική, την κοινωνία, κ.α., για τη διακίνηση ιδεών και ενεργειών στο μέγιστο δυνατό



Σχήμα 5: Γενικό επιδημιολογικό μοντέλο MSEIR [36]

πλήθος ατόμων. Ιδιαίτερα στα κοινωνικά δίκτυα, όπως το Facebook και το Instagram, αυτό γίνεται καταφανές όταν εταιρίες τεχνολογικών προιόντων ή προϊόντων προσωπικής φροντίδας χρηματοδοτούν άτομα με πολλούς εικονικούς ακολούθους ή φίλους και γενικότερα ευρύ δίκτυο επιρροής, αλλά και με χαρακτηριστικά με τα οποία πολλοί άνθρωποι μπορούν να ταυτιστούν ή επιδιώκουν να αποκτήσουν. Έτσι, η διαφήμιση προϊόντων μέσω αυτών των ατόμων αποσκοπεί στο να φτάσει η σχετική πληροφορία σε μεγάλο μέρος του πληθυσμού, το οποίο εν δυνάμει θα αγοράσει τα προϊόντα και θα αποφέρει κέρδος στις επιχειρήσεις. Οπότε, είναι εμφανής η ανάγκη μετρικών που ποσοτικοποιούν αυτή τη δύναμη που διαθέτουν κάποια άτομα, ώστε να είναι δυνατή η κατάλληλη επιλογή τους με το μικρότερο δυνατό κόστος. Σε ένα κόσμο, όπου η διαφήμιση θα ήταν δωρεάν θα μπορούσε μια επιχείρηση να προβάλλει τη διαφήμιση σε όλο τον πληθυσμό. Βέβαια, και αυτή η τακτική θα συναντούσε κωλύματα, λόγω του βομβαρδισμού των διαφημίσεων που θα λάμβανε ο μέσος άνθρωπος, τις οποίες δε θα μπορούσε να επεξεργαστεί και να λάβει μια σχετική κερδοφόρα για τις επιχειρήσεις απόφαση [37]. Ωστόσο, στην τωρινή πραγματικότητα οι διαφημίσεις κοστίζουν και ως εκ τούτου οι εταιρίες επιδιώκουν να ελαχιστοποιήσουν το κόστος της διαφήμισης εντοπίζοντας περιορισμένο αριθμό ανθρώπων με τη μεγαλύτερη ισχύ στο δίκτυο.

Σε αυτό το κεφάλαιο, θα παρουσιαστούν ο ορισμός και τα συμπεράσματα περί του υπολογισμού της επιρροής ενός κόμβου σε ένα κοινωνικό δίκτυο.

4.2.1 Επιρροή κόμβου

Ως επιρροή ενός κόμβου $u \in V$ ορίζεται το αναμενόμενο πλήθος κόμβων που θα βρίσκονται εν τέλει σε ενεργή κατάσταση δεδομένου ότι ο κόμβος u είναι ο αρχικός ενστερνιστής της ιδέας, και μαθηματικά συμβολίζεται ως $\sigma(u, G)$ [5]. Βέβαια, μπορεί να έχει ως όρισμα και σύνολο αρχικών ενστερνιστών A με τον ίδιο ακριβώς ορισμό, όπως προηγουμένως. Στα [38] και [39] αποδεικνύεται ότι ο ακριβής υπολογισμός της επιρροής $\sigma(\cdot, G)$ είναι #P-Δύσκολος για τα μοντέλα IC και LT, δηλαδή όλα τα προβλήματα που ανήκουν στην κλάση #P ανάγονται σε αυτό. Η κλάση αυτή περιέχει τα προβλήματα καταμέτρησης των διαφορετικών εφικτών λύσεων των προβλημάτων απόφασης που ανήκουν στην κλάση NP. Είναι λοιπόν προφανές ότι δεν μπορεί να υπολογιστεί σε πολυωνυμικό χρόνο. Στην περίπτωση του μοντέλου IC στο [38] η αναγωγή έγινε από το #P-Πλήρες πρόβλημα μέτρησης των επαγόμενων υπογράφων ενός γράφου $G(V, E)$ στους οποίους υπάρχει μονοπάτι μεταξύ των κόμβων $s, t \in V$

(s-t connectness problem). Αντίστοιχα, στην περίπτωση του μοντέλου LT στο [39] η αναγωγή έγινε από το #P-Δύσκολο πρόβλημα καταμέτρησης των απλών μονοπατιών σε κατευθυνόμενους γράφους (simple path counting problem).

Συνεπώς, γίνεται κατανοητό ότι είναι αναγκαία η επιστράτευση κάποιας προσεγγιστικής τεχνικής για τον υπολογισμό της επιρροής ενός κόμβου ή ενός συνόλου κόμβων.

Όπως φάνηκε παραπάνω, είναι δύσκολος ο υπολογισμός της επιρροής ενός κόμβου και ως εκ τούτου ο εντοπισμός των σημαντικών κόμβων, δηλαδή αυτών με τη μεγαλύτερη επιρροή. Έτσι, καταφεύγει κανείς στην αξιοποίηση της τοπολογίας του γράφου, προκειμένου να εντοπίσει με μικρό χρονικό κόστος αλλά με κίνδυνο σημαντικού λάθους τους κόμβους με τη μεγαλύτερη επίδραση στο δίκτυο. Όπως είναι αναμενόμενο, εξέχουσα θέση σε αυτή την απόπειρα κατέχουν οι κεντρικότητες κόμβων που αναφέρθηκαν στην Ενότητα 3.2.1, δηλαδή η κεντρικότητα βαθμού, κεντρικότητα εγγύτητας, κεντρικότητα ενδιαμεσικότητας και ιδιοδιανύσματος, κ.α..

5 Παραπληροφόρηση στα κοινωνικά δίκτυα

Η ραγδαία τεχνολογική ανάπτυξη και η μαζική ίδρυση ανταγωνιστικών μεταξύ τους τεχνολογικών εταιριών επέφερε αναπόφευκτα τη διάθεση σχετικών προϊόντων σε χαμηλές τιμές, με αποτέλεσμα όλο και περισσότεροι άνθρωποι να βρίσκονται στη θέση να δύνανται να αποκτήσουν ηλεκτρονικά προϊόντα με πρόσβαση στο διαδίκτυο. Με αυτό τον τρόπο, τέθηκε ευνοϊκό πλαίσιο για τη δημιουργία κοινωνικών πλατφορμών, όπου άνθρωποι από κάθε γωνία της υφηλίου θα μπορούσαν να επικοινωνήσουν μεταξύ τους και να ανταλλάξουν κοινωνικά και πολιτιστικά στοιχεία. Πολύ γρήγορα αυτές οι πλατφόρμες απαριθμούσαν υπέρογκο αριθμό χρηστών, οι οποίοι πλέον κατέφευγαν σε αυτές όχι μόνο για τους προαναφερθέντες σκοπούς, αλλά και για την ενημέρωσή τους σχετικά με γεγονότα πάσας χροιάς (κοινωνικά, πολιτικά, κ.α.). Χαρακτηριστικά σύμφωνα με έρευνα του Κέντρου Έρευνας Pew (Pew Research Centre) [40] που διεξήχθη κατά τη χρονική περίοδο 31/08/2020 - 07/09/2020 και στην οποία έλαβαν μέρος 9220 ενήλικοι πολίτες των Η.Π.Α., σχεδόν οι μισοί (53%) απάντησαν ότι ενημερώνονται συχνά ή μερικές φορές για την επικαιρότητα μέσω των κοινωνικών δικτύων. Κυρίαρχο φαίνεται να είναι το Facebook, το οποίο χρησιμοποιείται από το 68% των ερωτηθέντων, εκ των οποίων το 36% το αντιμετωπίζει ως πηγή ενημέρωσης.

Είναι λοιπόν εμφανές ότι μια ιδέα ή μια είδηση μπορεί να διαδοθεί εξαιρετικά γρήγορα σε μεγάλο πλήθος ανθρώπων μέσω των κοινωνικών δικτύων, γεγονός το οποίο με μια πρώτη ματιά φαίνεται σπουδαίο, αφού πλέον η γνώση δεν περιορίζεται στην προεπεξεργασμένη αντίστοιχη που δύνανται να παρέχουν τα παροδοσιακά Μ.Μ.Ε., αλλά γίνεται προϊόν διακίνησης όλων των χρηστών εντός μιας εικονικής κοινότητας. Βέβαια, θα πρέπει να σημειωθεί ότι σε επίπεδο κρατών ή περιοχών είναι δυνατή η επιβολή φίλτρων με τη χρήση εργαλείων επιπέδου δικτύου (π.χ., με τον περιορισμό εισερχόμενων πακέτων από συγκεκριμένες διευθύνσεις IP) που να αποκόβουν τη διάχυση πληροφορίας εντός ή εκτός ορισμένων ορίων αναιρώντας τη δυνατότητα πρόσβασης σε οποιαδήποτε διακινούμενη πληροφορία. Φυσικά, αυτή η δυνατότητα μπορεί να χρησιμοποιηθεί επιτηδευμένα ή μη για κακοήθεις σκοπούς προωθώντας λανθασμένες πληροφορίες. Τότε, γίνεται λόγος για παραπληροφόρηση και αποτελεί ένα από τα σημαντικότερα προβλήματα της σημερινής εποχής της πληροφορίας, καθώς μπορεί να έχει καταστροφικές συνέπειες σε πολλαπλά επίπεδα ανάλογα με το περιεχόμενο και το επιδιωκόμενο αποδέκτη της είδησης. Χαρακτηριστικό παράδειγμα αποτελεί μια ψευδή ανάρτηση από τον επίσημο λογαριασμό του Associated Press (κυρίαρχο ειδησεογραφικό πρακτορείο των Η.Π.Α.) στην πλατφόρμα Twitter στις 23/04/2013 [2], όπου γινόταν λόγος για δύο βομβαρδισμούς στο Λευκό Οίκο και τον τραυματισμό του τότε προέδρου των Η.Π.Α., Barack Obama. Λόγω της τεταμένης κατάστασης εκείνης της εποχής και του συνεχούς φόβου πιθανών τρομοκρατικών επιθέσεων, μια τέτοια ανάρτηση κατάφερε να παραλύσει προσωρινά την αγορά και να προκαλέσει την απώλεια εκατοντάδων δισεκατομμυρίων δολαρίων μέσω της πτώσης δεικτών του χρηματιστηρίου. Η κατάσταση απετράπη επιτυχώς με την ανακοίνωση περί ψευδούς ανάρτησης που προέκυψε από την παραβίαση του εν λόγω λογαριασμού.

Με στόχο την αποτροπή τέτοιων δυσάρεστων γεγονότων με δυνητικά οικουμενι-

κές ολέθριες συνέπειες κρίνεται επιτακτική η εύρεση μεθόδων πρόβλεψης και καταστολής της παραπληροφόρησης μέσω του εντοπισμού ψευδών ειδήσεων και την ανακοπή της εξάπλωσής τους αντίστοιχα. Στη συνέχεια, θα παρουσιαστεί συνοπτικά η ιδέα της αναγνώρισης μιας ψευδούς είδησης και θα γίνει μια εκτενεστερη παράθεση της βιβλιογραφίας σχετικά με τους τρόπους καταπολέμησής της.

5.1 Εντοπισμός της παραπληροφόρησης

Το πρόβλημα εντοπισμού της παραπληροφόρησης μπορεί να αντιμετωπισθεί ως πρόβλημα ταξινόμησης στο πλαίσιο της Μηχανικής Μάθησης [3], κατά το οποίο τελικός στόχος είναι δεδομένου μιας πληροφορίας η αντίστοιχισή της είτε στην κατηγορία "αληθής" είτε στην κατηγορία "ψευδής". Για την επίλυση τέτοιων προβλημάτων εφαρμόζονται δύο αρχές:

1. Εξαγωγή χαρακτηριστικών, κατά την οποία επιδιώκεται η αναπαράσταση των διαφόρων χαρακτηριστικών μιας είδησης σε μια αποτελεσματική και διαχειρίσιμη μορφή (συνήθως σε πίνακες).

2. Κατασκευή μοντέλων, τα οποία επεξεργάζονται τα προηγούμενα χαρακτηριστικά με στόχο την εκπαίδευσή τους και την απόφανσή τους σχετικά με την κατηγορία της πληροφορίας.

Στο πλαίσιο της αναζήτησης χρήσιμων χαρακτηριστικών μιας είδησης και της κατάλληλης αναπαράστασής τους, μπορεί κανείς να λάβει υπόψη, μεταξύ άλλων, τα ακόλουθα:

- Το περιεχόμενο της πληροφορίας (εμφανιζόμενες λέξεις) [41, 42].

- Την αξιοπιστία του χρήστη μέσω της ανάλυσης των προηγούμενων δημοσιεύσεών του [43], αφού ένας αξιόπιστος χρήστης θα προωθήσει μια ψευδή πληροφορία με μικρότερη πιθανότητα από έναν αναξιόπιστο χρήστη.

- Τη μεροληψία του χρήστη (π.χ. πολιτική) μέσω της διασταύρωσης της πληροφορίας με ποικίλες ιστοσελίδες, καθώς ένας προδιατεθειμένος χρήστης προς μια ιδέα είναι πολύ πιθανό να δημοσιεύσει πληροφορίες που θα την υποστηρίζουν [44].

- Τη συχνότητα και τα ενδιάμεσα χρονικά διαστήματα προώθησης της πληροφορίας από τους χρήστες με στόχο τη μελέτη της χρονικής εξέλιξης της εξάπλωσής της [45, 46, 47]. Για αυτό το σκοπό προτείνονται μοντέλα Επαναλαμβανόμενων Νευρωνικών Δικτύων (RNN) που επιτρέπουν τη χρήση επαναλήψεων εντός του νευρωνικού δικτύου για τη μοντελοποίηση των δεδομένων [46].

- Την τοπολογία του δικτύου σε επίπεδο γειτόνων, κοινοτήτων, κ.α. [48, 49, 50, 51], καθώς η διάδοση ψευδών ειδήσεων υποδεικνύει την ύπαρξη πολωμένων ομάδων [45, 52].

- Την αξιολόγηση της υποκείμενης ιδέας με βάση την αξιοπιστία των σχετικών αναρτήσεων [53].

- Την ύπαρξη αντικρουόμενων ή φίλα προσκείμενων δημοσιεύσεων ή αντιδράσεων (π.χ. στην πλατφόρμα του Facebook με τη μορφή του "like" [54]) σχετικά με την ίδια ιδέα, αφού όσο περισσότερες αναρτήσεις με την ίδια σκοπιά πάνω στην πληροφορία υφίστανται τόσο μεγαλύτερη είναι η αξιοπιστία αυτής της θέσης.

5.2 Περιορισμός της παραπληροφόρησης

Εφόσον έχει επιτευχθεί η αναγνώριση μιας ψευδούς πληροφορίας, επόμενο βήμα αποτελεί η ανακοπή της διάχυσής της εντός του κοινωνικού δικτύου [4]. Αυτή μοντελοποιείται εν γένει ως πρόβλημα ελαχιστοποίησης του πλήθους των χρηστών που θα λάβουν την παραπληροφόρηση και μαθηματικά εκφράζεται ως εξής:

$$S^* = \arg\min_S \sigma_M^S(A, G), \text{ υπό ορισμένες συνθήκες,} \tag{12}$$

όπου S μια εφαρμόσιμη στρατηγική καταπολέμησης της διάδοσης της παραπληροφόρησης και σ το πλήθος των προσβεβλημένων χρηστών από τη ψευδή πληροφορία στο τέλος της διάδοσής της στο δίκτυο αναπαριστώμενο με το γράφο $G = (V, E)$ με αρχικούς ενστερνιστές το σύνολο $A \subset V$ και σύμφωνα με το μοντέλο διάδοσης M.

Οι μέθοδοι που δύνανται να ακολουθηθούν με στόχο τη μείωση της εξάπλωσης μιας ψευδούς πληροφορίας χωρίζονται σε δύο μεγάλες κατηγορίες:

- Στρατηγική αφαίρεσης, όπου μεταβάλλεται το δίκτυο με την προσωρινή αφαίρεση κόμβων ή ακμών με στόχο την ανακοπή της ροής της πληροφορίας.

- Στρατηγική διαφώτισης, όπου διαδίδεται η αντίστοιχη αληθής πληροφορία με στόχο την επαγρύπνηση των χρηστών και την κριτική στάση απέναντι στην ψευδή είδηση.

Φυσικά αμφότερες στρατηγικές δεν είναι ιδανικές, καθώς στην περίπτωση της πρώτης είναι πιθανή η δυσφορία των χρηστών, οι οποίοι μπορεί να την αντιμετωπίσουν ως παραβίαση της ελευθερίας της έκφρασης [55] ή μπορεί να εντοπίσουν αλλαγές στην εμπειρία τους εντός του δικτύου με τελικό αποτέλεσμα την αποδοκιμασία ή και την εγκατάλειψή του [56]. Στη δεύτερη περίπτωση, η παροχή της αλήθειας στο χρήστη, αφότου αυτός έχει δεχθεί την παραπληροφόρηση ίσως να μην είναι αποτελεσματική λόγω του ενδοιασμού του να αλλάξει στάση εκ των υστέρων [55].

Σε κάθε περίπτωση, είναι αναγκαίος ο υπολογισμός της επιρροής του αρχικού συνόλου A (ή εναλλακτικά της εξάπλωσης της πληροφορίας) $\sigma(A)$. Προς τούτο, γίνεται χρήση των παρακάτω μεθόδων:

- **Κεντρικότητες** κόμβων [57], όπως κεντρικότητα βαθμού, κεντρικότητα εγγύτητας και κεντρικότητα ενδιαμεσικότητας, οι οποίες περιγράφονται στην ενότητα 3.2.1.

- **Προσομοίωση** του μελετούμενου μοντέλου διάδοσης πολλαπλές φορές και υπολογισμός του μέσου όρου των αποτελεσμάτων.

- **Μονοπάτια**, όπου κυριαρχεί το μοντέλο Maximum Influence Arborescence (MIA) [58], κατά το οποίο υπολογίζονται τα συντομότερα μονοπάτια μεταξύ όλων των κόμβων και λαμβάνονται υπόψη μόνο όσα έχουν πιθανότητα εμφάνισης μεγαλύτερη από την τιμή ενός κατωφλίου θ. Έτσι, η επιρροή ενός κόμβου περιορίζεται σε μια τοπική περιοχή. Χρησιμοποιώντας τα υπολογισμένα μονοπάτια συντίθενται για κάθε κόμβο δύο δέντρα με ρίζα τον ίδιο, τα Maximum Influence In-Arborescence (MIIA) και Maximum Influence Out-Arborescence (MIOA) για την υπόδειξη των κόμβων που ασκούν και υφίστανται επιρροή από τον κόμβο αντίστοιχα.

- **Δειγματοληψία**, όπου κυριαρχεί το μοντέλο Reverse Influence Sampling (RIS) [59, 60]. Σύμφωνα με αυτό παράγονται τυχαία θ δείγματα του γράφου, σε καθένα εκ των οποίων επιλέγεται ένας τυχαίος κόμβος και υπολογίζεται το σύνολο των κόμβων (reverse reachable set) από τους οποίους υπάρχει μονοπάτι προς αυτόν. Η εμφάνιση του ίδιου συνόλου σε πολλά δείγματα υποδεικνύει την ισχύ της επιρροής του.

Στο συγκεκριμένο κεφάλαιο, θα παρουσιαστεί η περαιτέρω κατηγοριοποίηση των μεθόδων αντιμετώπισης της παραπληροφόρησης, καθώς και η υπάρχουσα σχετική βιβλιογραφία.

5.2.1 Μέθοδοι αφαίρεσης κόμβων/ακμών

Οι μέθοδοι αφαίρεσης χρησιμοποιούν ως εργαλείο ενάντια στη διάδοση της παραπληροφόρησης την προσωρινή αφαίρεση κόμβων ή ακμών ούτως ώστε η προκύπτουσα μορφή του δικτύου να μην παρέχει διόδους για τη διάχυση της ψευδούς πληροφορίας. Είναι κατανοητό ότι η αφαίρεση (ή απομόνωση) κόμβων έχει πιο ισχυρή επίδραση πάνω στην τοπολογία του δικτύου αφού επιφέρει και την αφαίρεση των προσπίπτουσων ακμών στους εν λόγω κόμβους. Με αυτό τον τρόπο, ωστόσο, ένας χρήστης αποκόβεται πλήρως από το δίκτυο, γεγονός το οποίο, ιδιαίτερα αν έχει μεγάλη διάρκεια, ενδέχεται να οδηγήσει στην απεγγραφή του ίδιου ή και άλλων, είτε συμπάσχοντων είτε διαφωνούντων με τέτοιες τακτικές, από την πλατφόρμα. Κάτι τέτοιο θα ήταν οικονομικά επιβλαβές για μια τέτοια εταιρία, της οποίας τα κέρδη βασίζονται κατά κύριο λόγο στις διαφημίσεις που προβάλλονται στους χρήστες όταν είναι συνδεδεμένοι στο κοινωνικό δίκτυο. Επομένως, η αφαίρεση ακμών φαίνεται να είναι πιο κομψή λύση καθώς δεν επηρεάζει σε τόσο μεγάλο βαθμό την εμπειρία του χρήστη και είναι λιγότερο κοστοβόρα.

5.2.1.1 Μέθοδοι αφαίρεσης κόμβων

Σε αυτή την υποκατηγορία μεθόδων επίλυσης του προβλήματος παραπληροφόρησης, είναι επιθυμητή η εύρεση ενός συνόλου κόμβων NS, η αφαίρεση (ή απομόνωση)

των οποίων από το γράφο (μαζί με τις προσπίπτουσες σε αυτούς ακμές) θα επιφέρει τη μείωση του τελικού πλήθους των αποδεκτών της λανθασμένης πληροφορίας. Αυτή η επιλογή και αναστολή κόμβων μπορεί να γίνει είτε στατικά στην αρχή της διάδοσης της πληροφορίας είτε δυναμικά προσαρμοζόμενη στην εξέλιξη της εξάπλωσης της είδησης.

Στην πρώτη περίπτωση, προκειμένου να βρεθούν οι πιο κυρίαρχοι κόμβοι του δικτύου προς αφαίρεση γίνεται χρήση των κεντρικοτήτων που αναφέρθηκαν στην ενότητα 3.2.1 αγνοώντας το σύνολο των αρχικών ενστερνιστών της πληροφορίας A [62, 63, 64]. Κανείς μπορεί να λάβει αυτό το σύνολο υπόψη και να θέσει κάποιους περιορισμούς στο πλήθος των προς αφαίρεση κόμβων ή στο συνολικό κόστος που επιφέρει η αφαίρεσή τους δεδομένου ότι η απομόνωση κάθε κόμβου έχει κάποιο κόστος [65, 66, 67, 68]. Επιπλέον, έχει προταθεί η αφαίρεση των κόμβων οι οποίοι έχουν τη μεγαλύτερη πιθανότητα μόλυνσης και προώθησης της πληροφορίας και έτσι αποτελούν διόδους της διάδοσής της [69], αλλά και η απομάκρυνση του ελάχιστου αριθμού κόμβων που θα επιφέρουν μείωση στο πλήθος των τελικών προσβεβλημένων κόμβων μεγαλύτερη από κάποιο ορισμένο κατώφλι [70, 71]. Το τελευταίο εμπλουτίζεται και με την επιθυμία για περιορισμό της παραπληροφόρησης εντός της κοινότητας στην οποία ανήκουν οι αρχικοί ενστερνιστές A [72].

Στη δεύτερη περίπτωση, είναι προτιμητέα η παρατήρηση της διάδοσης της πληροφορίας και η λήψη μέτρων για την αφαίρεση κόμβων κατά τη διάρκειά της. Είναι προφανές ότι στους σχετικούς αλγορίθμους λαμβάνεται υπόψη το στοιχείο του χρόνου και σε κάθε διακριτό βήμα αυτού επιλέγεται ένας κόμβος ή ένα σύνολο κόμβων που μεγιστοποιούν τη μείωση της εξάπλωσης της πληροφορίας [73, 74, 75]. Εκτός αυτής της άπληστης προσέγγισης, δύναται κανείς να λάβει υπόψη τη δημοφιλία της είδησης και την εμπειρία του χρήστη, ώστε να θέσει ως χρονικό περιορισμό το διάστημα ανοχής της απομόνωσης του χρήστη [56].

Στον Πίνακα 1, παρουσιάζονται συνοπτικά οι περισσότερες εκ των υπάρχουσων μεθόδων που επιστρατεύουν την αφαίρεση κόμβων για τον περιορισμό της παραπληροφόρησης.

Αν κληθεί κανείς να επιλέξει μεταξύ μιας στατικής ή μιας δυναμικής μεθόδου αφαίρεσης κόμβων οφείλει να λάβει υπόψη του ότι ναι μεν οι στατικές μέθοδοι είναι υπολογιστικά απλές και φθηνές, αλλά στερούνται ακρίβειας αφού αγνοούν την πραγματική εξέλιξη της διάδοσης της πληροφορίας. Αντίστοιχα, οι δυναμικές μέθοδοι λαμβάνοντας αυτό υπόψη επιτυγχάνουν καλύτερα αποτελέσματα ακόμα και με την αφαίρεση λιγότερων κόμβων, αλλά με μεγάλο υπολογιστικό κόστος λόγω της μελέτης του τρόπου διάδοσης.

5.2.1.2 Μέθοδοι αφαίρεσης ακμών

Σε αυτή την υποκατηγορία μεθόδων επίλυσης του προβλήματος της παραπληροφόρησης, είναι επιθυμητή η εύρεση ενός συνόλου ακμών πληθικότητας k, η αφαίρεση των οποίων από το γράφο θα επιφέρει τη μείωση του τελικού πλήθους των αποδεκτών της λανθασμένης πληροφορίας. Αυτή η επιλογή και αναστολή συνδέσεων μπορεί να γίνει λαμβάνοντας ή όχι υπόψη το αρχικό σύνολο ενστερνιστών A. Είναι εμφανές

Πίνακας 1: Κατηγοριοποίηση αλγορίθμων αφαίρεσης κόμβων (ως SIR νοείται περίπτωση του γενικού μοντέλου MSEIR)

Δημοσίευση Αλγορίθμου	**Κατηγορία**	**Γράφος**		**Μοντέλο Διάχυσης**	**Υπολογισμός** σ
		Κατευθυνόμενος	**Με βάρη**		
[62]	Στατικός	Όχι	Όχι	-	Κεντρικότητες
[63]	Στατικός	Όχι	Όχι	SIR	Κεντρικότητες
[64]	Στατικός	Όχι	Όχι	-	Κεντρικότητες
[65]	Στατικός	Ναι	Όχι	IC	Προσομοίωση
[66]	Στατικός	Ναι	Ναι	LT	Δειγματοληψία
[67]	Στατικός	Ναι	Ναι	IC	Κεντρικότητες
[68]	Στατικός	Ναι	Ναι	LT	Μονοπάτια
[69]	Στατικός	Όχι	Ναι	SIR	Προσομοίωση
[76]	Στατικός	Ναι	Ναι	LT	Κεντρικότητες
[70]	Στατικός	Ναι	Ναι	LT	Προσομοίωση
[71]	Στατικός	Ναι	Ναι	LT, IC	Δειγματοληψία
[72]	Στατικός	Ναι	Ναι	IC	Προσομοίωση
[73]	Δυναμικός	Ναι	Ναι	IC	Μονοπάτια
[74]	Δυναμικός	Ναι	Ναι	LT	Δειγματοληψία
[75]	Δυναμικός	Ναι	Ναι	-	Προσομοίωση
[56]	Δυναμικός	Ναι	Ναι	IC	Κεντρικότητες
[77]	Δυναμικός	Όχι	Ναι	SIR	Κεντρικότητες

ότι χωρίς τη γνώση των σημείων εκκίνησης της διάδοσης στόχος αποτελεί ο γενικός περιορισμός της ροής της πληροφορίας στο δίκτυο.

Στην περίπτωση έλλειψης πληροφορίας για το αρχικό σύνολο ενστερνιστών A μπορεί να τεθεί ως σκοπός η μείωση του μέσου (ή αναμενόμενου) βαθμού μόλυνσης του δικτύου, ο οποίος ορίζεται ως ο μέσος όρος των επιρροών των κόμβων του γράφου [78]. Μια εναλλακτική του αναμενόμενου βαθμού μόλυνσης είναι ο χείριστος βαθμός μόλυνσης, δηλαδή η μέγιστη επιρροή των κόμβων του δικτύου [79]. Όταν το μοντέλο διάδοσης είναι το LT εφαρμόζεται η μέθοδος Bond Percolation [80] για τον αποδοτικό κατά προσέγγιση υπολογισμό της επιρροής του κάθε κόμβου και ως αλγόριθμος επίλυσης προτείνεται μια επαναλαπτική άπληστη στρατηγική, σε κάθε επανάληψη της οποίας αφαιρείται η ακμή που προκαλεί τη μεγαλύτερη μείωση στην τιμή του βαθμού μόλυνσης του δικτύου. Αντίστοιχα, η ίδια αντιμετώπιση υιοθετείται και στην περίπτωση του IC ως μοντέλου διάδοσης [81].

Όμοια, ως μετρική μπορεί να οριστεί η ευαισθησία ενός δικτύου απέναντι στη διάδοση μιας πληροφορίας ως το άθροισμα των επιρροών όλων των κόμβων του γράφου [82]. Μελετώντας το μοντέλο διάδοσης LT εφαρμόζεται η μέθοδος των live-edge γράφων για τον υπολογισμό της μετρικής, της οποίας η μείωση επιτυγχάνεται με την επιστράτευση ενός άπληστου αλγορίθμου που εξασφαλίζει προσέγγιση της βέλτιστης λύσης της τάξης $1 - \frac{1}{\epsilon}$. Αντί του υπολογισμού της ευαισθησίας του δικτύου, δύναται να χρησιμοποιηθεί ως ένδειξη αυτής η μεγαλύτερη ιδιοτιμή του πίνακα γειτνίασης του γράφου, και ως εκ τούτου αφαιρώντας τις k ακμές με τη μεγαλύτερη βαθμολογία υπολογισμένη με βάση τις ιδιοτιμές του πίνακα, επιδιώκεται η ελαχιστοποίηση της εν λόγω

μεγαλύτερης ιδιοτιμής [83]. Επιπλέον, ένας εναλλακτικός ορισμός της ευπάθειας ενός δικτύου είναι το άθροισμα των μεγεθών των μεγαλύτερων συνδεδεμένων συνιστωσών του γράφου, όπου ως εργαλείο ελαχιστοποίησής του επιλέγεται η κεντρικότητα ενδιαμεσικότητας των ακμών [63]. Αυτή η μετρική αξιοποιείται και στην περίπτωση όπου το πρόβλημα εκφράζεται με τη χρήση τυχαίων περιπάτων [64].

Στην περίπτωση που οι αρχικοί ενστερνιστές A της ψευδούς είδησης είναι γνωστοί, ως στόχος τίθεται η μείωση των τελικών προσβληθέντων από την είδηση, όπως αυτή αρχικά διαδόθηκε από το σύνολο κόμβων A. Έτσι, επιστρατεύεται η γνωστή άπληστη στρατηγική, κατά την οποία σε κάθε επανάληψη αφαιρείται η ακμή που προκαλεί τη μέγιστη μείωση της επιρροής του συνόλου A [84]. Προκειμένου να γίνει ο υπολογισμός της εν λόγω επιρροής πιο αποδοτικά χρησιμοποιείται η μέθοδος των live-edge γράφων, η οποία ευνοεί την κατασκευή δένδρων απογόνων των κόμβων του συνόλου A και κατά συνέπεια την ανανέωση της τιμής του κέρδους αφαίρεσης κάθε ακμής σε κάθε επανάληψη [6]. Επιπλέον, είναι δυνατή η επιλογή των προς αφαίρεση k ακμών από ένα συγκεκριμένο υποψήφιο σύνολο με στόχο τη μείωση του αθροίσματος των πιθανοτήτων μόλυνσης των κόμβων του δικτύου [85]. Προς τούτο, προτείνεται ένας άπληστος αλγόριθμος που επιλέγει σε κάθε επανάληψη την ακμή με τη μεγαλύτερη μείωση της προαναφερθείσας μετρικής και ανανεώνει αντίστοιχα τις πιθανότητες ενεργοποίησης των κόμβων. Φυσικά, ως περιορισμός στην επίλυση του προβλήματος μπορεί να τεθεί αντί της πληθικότητας του συνόλου των αφαιρεμένων ακμών η μη υπέρβαση ενός προϋπολογισμού δεδομένου ότι η αφαίρεση κάθε ακμής αντιστοιχεί σε κάποιο κόστος [86].

Το πρόβλημα περιορισμού της παραπληροφόρησης μπορεί να συγκεκριμενοποιηθεί θέτοντας ως στόχο τη μείωση της παραπληροφόρησης απέναντι σε ένα ορισμένο σύνολο κόμβων, που πρέπει να προστατευθεί από αυτή [87]. Σε αυτή την εκδοχή του ζητήματος, μπορεί να επιβληθεί ή όχι περιορισμός στο πλήθος των ακμών που δύνανται να αφαιρεθούν, με αποτέλεσμα να λύνεται είτε με τη χρήση επαναληπτικού άπληστου αλγορίθμου βασιζόμενου σε δειγματοληψία είτε με την αξιοποίηση του προβλήματος της ελάχιστης τομής [88] αντίστοιχα. Μια επέκταση του που θεωρεί μοντέλο διαδοχικής διάδοσης επιλύεται μέσω μαθηματικού προγραμματισμού [89].

Στον Πίνακα 2, παρουσιάζονται συνοπτικά οι περισσότερες εκ των υπάρχουσων μεθόδων που επιστρατεύουν την αφαίρεση ακμών για τον περιορισμό της παραπληροφόρησης.

Είναι φυσικό επόμενο ότι οι μέθοδοι αφαίρεσης ακμών που αξιοποιούν τη γνώση του αρχικού συνόλου A είναι πιο αποτελεσματικές στον περιορισμό της παραπληροφόρησης από αυτές που δρουν εν αγνοία αυτού. Ωστόσο, προϋποθέτουν ότι αυτή η γνώση δύναται να αποκτηθεί με ταχύτητα και ακρίβεια, ειδάλλως η προσοχή καταναλώνεται στον εντοπισμό των αρχικών ενστερνιστών της είδησης και στρέφεται μακριά από το υπό μελέτη πρόβλημα της αφαίρεσης ακμών.

5.2.2 Μέθοδοι διαφώτισης

Οι μέθοδοι διαφώτισης (clarification-based methods) προβάλλουν ως εργαλείο καταπολέμησης της διάδοσης της παραπληροφόρησης την ταυτόχρονη διάδοση της α-

Πίνακας 2: Κατηγοριοποίηση αλγορίθμων αφαίρεσης ακμών

Δημοσίευση Αλγορίθμου	**Αρχικοί Ενστερνιστές**	**Γράφος**		**Μοντέλο Διάχυσης**	**Υπολογισμός σ**
		Κατευθυνόμενος	**Με βάρη**		
[78]	Άγνωστοι	Ναι	Ναι	LT	Προσομοίωση
[81]	Άγνωστοι	Ναι	Όχι	IC	Προσομοίωση
[79]	Άγνωστοι	Ναι	Όχι	IC	Προσομοίωση
[82]	Άγνωστοι	Ναι	Ναι	LT	Προσομοίωση
[83]	Άγνωστοι	Ναι	Όχι	-	Κεντρικότητες
[63]	Άγνωστοι	Όχι	Όχι	SIR	Κεντρικότητες
[64]	Άγνωστοι	Όχι	Όχι	-	Κεντρικότητες
[84]	Γνωστοί	Ναι	Ναι	IC	Προσομοίωση
[6]	Γνωστοί	Ναι	Ναι	LT	Προσομοίωση
[85]	Γνωστοί	Ναι	Ναι	IC	Μονοπάτια
[86]	Γνωστοί	Ναι	Ναι	LT	Κεντρικότητες
[87]	Γνωστοί	Ναι	Ναι	LT	Δειγματοληψία
[89]	Γνωστοί	Ναι	Όχι	-	Κεντρικότητες

ντίστοιχης αληθούς πληροφορίας από ένα αρχικό σύνολο διαφωτιστών B με στόχο τη μείωση των τελικών κόμβων που θα αποδεχθούν την ψευδή είδηση. Αυτές μπορούν να χωρισθούν σε δύο κατηγορίες, των οποίων ειδοποιός διαφορά αποτελεί ο τύπος περιορισμού που επιβάλλεται στη λύση του προβλήματος. Έτσι, οι μέθοδοι εκστρατείας αποσκοπούν στη μείωση της διάδοσης της παραπληροφόρησης με το πλήθος των αρχικών διαφωτιστών να είναι μικρότερο ενός ορισμένου κατωφλίου, ενώ οι μέθοδοι προστασίας στοχεύουν στην προφύλαξη ενός ποσοστού των κόμβων του γράφου από την ψευδή πληροφορία με το ελάχιστο δυνατό πλήθος αρχικών διαφωτιστών. Φυσικά, η επιλογή και παρότρυνση ενός συνόλου κόμβων B να διαδώσουν πρώτοι την αληθή είδηση είναι κοστοβόρα. Είναι κατανοητό ότι στη δεύτερη κατηγορία μεθόδων είναι θεμιτή η "θυσία" ενός ορισμένου ποσοστού των κόμβων προκειμένου να γίνει εξοικονόμηση στο πλήθος των κόμβων που πρέπει να εκκινήσουν την καμπάνια αλήθειας. Από άλλη σκοπιά, φαίνεται πως ακόμα κι αν αυτή συνεπάγεται μεγαλύτερο κόστος, είναι επιτακτική η προστασία ενός μέρους του πληθυσμού, ίσως γιατί θεωρείται ότι είναι το μικρότερο δυνατό χωρίς καταστροφικές συνέπειες εντός κι εκτός κοινωνικού δικτύου. Για παράδειγμα, αν ένα μόνο μικρό ποσοστό χρηστών πιστέψει ότι γίνεται πόλεμος σε γειτονική χώρα εξαιτίας ψευδούς είδησης, αυτό δε θα προκαλέσει τόσο μεγάλη αναταραχή όσο αν το πίστευε ένα μεγαλύτερο πλήθος ανθρώπων. Σε άλλη τροχιά, η πρώτη κατηγορία μεθόδων καλείται να επιλύσει το συνηθισμένο πρόβλημα ελαχιστοποίησης της ζημίας έχοντας ένα συγκεκριμένο προϋπολογισμό.

5.2.2.1 Μέθοδοι εκστρατείας

Σε αυτή την υποκατηγορία των μεθόδων διαφώτισης, είναι επιθυμητή η εύρεση ενός συνόλου κόμβων B πληθικότητας το πολύ k, οι οποίοι θα εκκινήσουν τη διάδοση της αντίστοιχης αληθούς είδησης με στόχο την ελαχιστοποίηση του τελικού πλήθους των κόμβων που θα προσβληθούν από τη διάδοση της παραπληροφόρησης. Για αυτό

το σκοπό, το σύνολο των αρχικών ενστερνιστών της αλήθειας B μπορεί να επιλεχθεί με βάση είτε την τοπολογία του δικτύου είτε σε συνδυασμό με τη δομή του γράφου, την ατομική συμπεριφορά των χρηστών (προτιμήσεις, προσωπικό όφελος, τοποθεσία, κ.α.).

Έχοντας ως εργαλείο μόνο την τοπολογία του γράφου έχουν προταθεί επεκτάσεις του μοντέλου διάδοσης LT που εισάγουν το στοιχείο του ανταγωνισμού μεταξύ των διαδιδόμενων πληροφοριών, ούτως ώστε να ενσωματώνεται πληροφορία τόσο για τη διάδοση της αληθούς και της ψευδούς είδησης όσο και για την αλληλεπίδρασή τους στους προσβαλλόμενους κόμβους. Σε κάθενα εξ αυτών η ανταγωνιστική φύση της σχέσης των δύο ειδήσεων αποτυπώνεται στις δύο διαφορετικές τιμές του κατωφλίου θ για κάθε κόμβο ή/και του βάρους της ακμής που εκφράζει την πιθανότητα της διάδοσης στην εν λόγω ακμή, ενώ ορίζεται και ποια πληροφορία υπερισχύει σε περίπτωση ταυτόχρονης προσβολής ενός κόμβου [90, 91, 92]. Προκειμένου να επιλεχθεί το κατάλληλο σύνολο αρχικών ενστερνιστών της αλήθειας μπορούν να εφαρμοσθούν επαναληπτικοί αλγόριθμοι επιλέγοντας σε κάθε επανάληψη τον κόμβο που προκαλεί τη μεγαλύτερη μείωση των συνεπειών της παραπληροφόρησης [90, 92]. Εναλλακτικά, μπορούν να επιλεχθούν οι κόμβοι, οι οποίοι θα προλάβουν να διαδώσουν την αληθή είδηση πριν την ψευδή σε σημαντικούς κόμβους με μεγάλη επιρροή [91]. Επιπλέον, μπορεί να ληφθεί υπόψη στο μοντέλο LT το ενδεχόμενο ανάκλησης άποψης από τους χρήστες, όπου για την επίλυση του προβλήματος υιοθετείται μια άπληστη μέθοδος και μια μέθοδος που αξιοποιεί την κεντρικότητα Page Rank [93, 94, 95].

Όπως για το μοντέλο LT προτείνονται και για το μοντέλο IC αντίστοιχες επεκτάσεις. Σε αυτές γίνεται χρήση ευριστικών μεθόδων με βάση τους κόμβους που έχουν μεγάλο βαθμό, ή που προσβάλλονται νωρίς ή εμφανίζουν τη μεγαλύτερη πιθανότητα να προσβληθούν από την παραπληροφόρηση [96]. Επίσης, μπορεί να εφαρμοσθεί άπληστος επαναληπτικός αλγόριθμος επιλέγοντας κάθε φορά τον κόμβο που θα αποφέρει τη μεγαλύτερη μείωση της διάδοσης της ψευδούς πληροφορίας [97], ή μπορούν να αξιοποιηθούν οι κεντρικότητες βαθμού, ενδιαμεσικότητας και εγγύτητας [98]. Αν κανείς λάβει υπόψη τις κοινότητες εντός ενός δικτύου, είναι σκόπιμο να αφιερώσει ποσοστό του συνόλου των προς επιλογή κόμβων σε κάθε κοινότητα ανάλογα με το εντός αυτής πλήθος των αρχικών ενστερνιστών της ψευδούς είδησης [99]. Βέβαια, είναι δυνατό να αξιοποιηθούν μέθοδοι δειγματοληψίας προκειμένου να επιλεχθούν οι κατάλληλοι κόμβοι για την εκστρατεία της αλήθειας [100, 101]. Αντίστοιχα με προηγουμένως, μπορεί να θεωρήσει κανείς δυνατή την ανάκληση μιας άποψης αφότου λάβει την αληθή είδηση και να το αντιμετωπίσει με την κλασική άπληστη μέθοδο επιλογής κόμβων με βάση τη μέγιστη μείωση της διάδοσης της ψευδούς είδησης [102], ή επιπλέον να γίνει εκκίνηση περισσότερων από μιας εκστρατειών αλήθειας επιστρατεύοντας αρχές της Θεωρίας Παιγνίων [103].

Λαμβάνοντας υπόψη συμπληρωματικά με τη δομή του δικτύου, και τα χαρακτηριστικά των χρηστών, μπορεί να γίνει μια πιο λεπτομερής προσέγγιση του μελετούμενου προβλήματος. Για παράδειγμα μπορεί να τεθεί ως περιορισμός ένα χρονικό διάστημα T, στα όρια του οποίου είναι επιθυμητή η μείωση της διάδοσης της παραπληροφόρησης, και να συμπεριληφθεί στη μελέτη ο χρόνος καθυστέρησης της ανταλλαγής πληροφο-

ρίας μεταξύ δύο κόμβων. Προς τούτο, μπορεί να γίνει χρήση των αλγορίθμων διάσχισης DFS και BFS προς εύρεση των κόμβων με τη μεγαλύτερη επιρροή στη διάδοση της αλήθειας ενάντια της ψευδούς πληροφορίας [104], ή εναλλακτικά αξιοποιώντας και το προσωπικό ενδιαφέρον του χρήστη στην παραπληροφόρηση και την αλήθεια μπορεί να εφαρμοσθεί ο γνωστός επαναληπτικός αλγόριθμος επιλογής κόμβων με βάση τη μέγιστη μείωση της διάδοσης της ψευδούς είδησης [105]. Επιπλέον, είναι δυνατό να θεωρήσει κανείς ότι ένας χρήστης θα αλλάξει γνώμη μετά την επιρροή και λήψη άλλων ειδήσεων, όπου σε επέκταση του μοντέλου LT συμπεριλαμβάνονται ο βαθμός αξιοπιστίας και το κατώφλι ανάκλησης άποψης για κάθε κόμβο [35]. Σε παρόμοια λογική, και με αξιοποίηση της πρότερης γνώσης, του ενδεχόμενου ενδοιασμού και της προσωρινής απώλειας μνήμης περί της υπό μελέτη είδησης των χρηστών προτείνεται ο συνήθης άπληστος αλγόριθμος με επιλογή κόμβων με βάση τη μέγιστη μείωση της διάδοσης της ψευδούς είδησης [106, 107]. Πέραν αυτών, μπορεί να τεθεί ως στόχος η προστασία από την παραπληροφόρηση κάποιων ακμών υψηλού κόστους [108] ή να θεωρηθεί ότι άνω της μίας εκστρατείες αλήθειας διενεργούνται, όπου η αποδοχή έκαστης πληροφορίας από τον κόμβο γίνεται ανάλογα με την υπόληψη της πηγής της, της προσωπικής πεποίθησης του κόμβου και της αξιοπιστίας του μηνύματος [109].

Αντικείμενο μελέτης έχει υπάρξει και η γεωγραφική θέση των χρηστών του κοινωνικού δικτύου. Ενδέχεται να είναι θεμιτός ο περιορισμός της παραπληροφόρησης εντός μιας ορισμένης γεωγραφικής περιοχής, για την επίτευξη του οποίου εντοπίζονται άπληστα οι κόμβοι με τη μεγαλύτερη επιρροή [110]. Εναλλακτικά, μπορεί να τεθεί ως περιορισμός του προβλήματος η επιλογή κόμβων εντός μιας συγκεκριμένης περιοχής [111]. Φυσικά, είναι λογικό να ληφθεί υπόψη και η κινητικότητα των χρηστών ανά περιοχές, οπότε επιλέγονται κινητοί κόμβοι με στόχο τη διάδοση της αλήθειας [112].

Είναι εύκολα κατανοητό ότι οι μέθοδοι που αξιοποιούν τα ιδιαίτερα χαρακτηριστικά των κόμβων για τον περιορισμό της παραπληροφόρησης είναι πιο αποτελεσματικοί από αυτές που περιορίζονται στη δομή του δικτύου. Βέβαια, η απόκτηση γνώσης σχετικά με αυτά τα χαρακτηριστικά απαιτεί συχνά τη συναίνεση των χρηστών ή ακόμα δεν είναι καν διαθέσιμη στα περισσότερα πραγματικά δίκτυα.

5.2.2.2 Μέθοδοι προστασίας

Σε αυτή την υποκατηγορία των μεθόδων διαφώτισης, είναι επιθυμητή η εύρεση ενός συνόλου κόμβων B ελάχιστης δυνατής πληθικότητας, οι οποίοι θα εκκινήσουν τη διάδοση της αντίστοιχης αληθούς είδησης με στόχο ένα ορισμένο ποσοστό χρηστών να μην προσβληθεί από την ψευδή είδηση.

Το εν λόγω πρόβλημα μπορεί να αναλυθεί σε τέσσερις διαφορετικές παραλλαγές θέτοντας περιορισμό ή όχι ως προς τη γνώση του συνόλου των αρχικών ενστερνιστών της ψευδούς πληροφορίας A, ή το πλήθος των χρονικών διακριτών βημάτων T από την πηγή της διάδοσης στα οποία σταματάει η τελευταία [113, 114]. Στην περίπτωση άγνοιας του συνόλου A, το πρόβλημα αντιμετωπίζεται ως πρόβλημα μεγιστοποίησης της επιρροής των κόμβων που θα διαδώσουν την αληθή είδηση, και για αυτό με τη μεταβλητή T είτε ορισμένη είτε όχι, υιοθετείται επαναληπτικός άπληστος αλγόριθμος

όπου επιλέγεται κάθε φορά ο κόμβος με τη μεγαλύτερη αύξηση της επιρροής του συνόλου των ενστερνιστών της αλήθειας μέχρις ότου αυτή η επιρροή να ικανοποιεί το επιθυμητό προς προστασία ποσοστό των χρηστών. Αντίστοιχα, αν το σύνολο A είναι γνωστό και ανεξάρτητα από τον περιορισμό ή μη στη μεταβλητή T, ελέγχεται αν η επιρροή του συνόλου A παραβιάζει το επιθυμητό ποσοστό προστασίας και ακολούθως επιλέγονται, είτε από όλο το δίκτυο, είτε ανά κοινότητα επαναληπτικά και άπληστα, κόμβοι με μεγάλη επιρροή μέχρι να ικανοποιηθεί η ζητούμενη συνθήκη. Επιπλέον, μπορεί να οριστεί ότι το σύνολο A βρίσκεται εξ ολοκλήρου εντός μίας κοινότητας και έτσι να στοχεύσει κανείς στον περιορισμό της παραπληροφόρησης εντός αυτής εντοπίζοντας κόμβους-γέφυρες, των οποίων η προστασία κατά ένα ποσοστό θα επιτευχθεί με την άπληστη επιλογή του ελάχιστου δυνατού συνόλου υποστηρικτών της αληθούς είδησης [115]. Τέλος, είναι εύλογο να θεωρήσει κανείς τη διάδοση της πληροφορίας σε περισσότερα του ενός κοινωνικά δίκτυα μέσω χρηστών που δε συμμετέχουν αποκλειστικά σε ένα, οπότε κρίνεται σκόπιμη η εφαρμογή ενός άπληστου επαναληπτικού αλγορίθμου επιλογής κόμβων με βάση τη μεγαλύτερη μείωση της διάδοσης της παραπληροφόρησης, οι οποίοι ουσιαστικά θα είναι οι ελάχιστοι δυνατοί με την απαιτούμενη επιρροή πάνω στους κόμβους που ανήκουν σε πολλαπλά δίκτυα [116].

Στον Πίνακα 3, παρουσιάζονται συνοπτικά οι περισσότερες εκ των υπάρχουσων μεθόδων που επιστρατεύουν την εκκίνηση της διάδοσης της αληθούς πληροφορίας επιλέγοντας ένα σύνολο αρχικών ενστερνιστών B με στόχο τη μείωση της εξάπλωσης της αντίστοιχης ψευδούς είδησης.

Πίνακας 3: Κατηγοριοποίηση αλγορίθμων διαφώτισης

Δημοσίευση Αλγορίθμου	**Κατηγορία**	**Γράφος**		**Μοντέλο Διάχυσης**	**Υπολογισμός σ**
		Κατευθυνόμενος	**Με βάρη**		
[90]	Εκστρατεία-Δομή	Ναι	Ναι	LT	Μονοπάτια
[91]	Εκστρατεία-Δομή	Ναι	Ναι	LT	Προσομοίωση
[92]	Εκστρατεία-Δομή	Ναι	Ναι	LT	Κεντρικότητες
[94]	Εκστρατεία-Δομή	Ναι	Ναι	LT	Κεντρικότητες, Προσομοίωση
[95]	Εκστρατεία-Δομή	Ναι	Ναι	LT	Κεντρικότητες, Προσομοίωση
[96]	Εκστρατεία-Δομή	Ναι	Ναι	IC	Κεντρικότητες, Προσομοίωση
[97]	Εκστρατεία-Δομή	Ναι	Ναι	IC	Μονοπάτια
[98]	Εκστρατεία-Δομή	Όχι	Όχι	IC	Κεντρικότητες
[99]	Εκστρατεία-Δομή	Ναι	Ναι	IC	Κεντρικότητες, Μονοπάτια
[100]	Εκστρατεία-Δομή	Ναι	Ναι	IC	Δειγματοληψία
[101]	Εκστρατεία-Δομή	Ναι	Ναι	IC	Δειγματοληψία
[102]	Εκστρατεία-Δομή	Ναι	Ναι	LT, IC	Κεντρικότητες
[103]	Εκστρατεία-Δομή	Ναι	Ναι	IC	Δειγματοληψία
[104]	Εκστρατεία-Χρήστες	Ναι	Ναι	IC	Δειγματοληψία
[105]	Εκστρατεία-Χρήστες	Ναι	Ναι	LT, IC	Προσομοίωση
[35]	Εκστρατεία-Χρήστες	Ναι	Ναι	LT	Προσομοίωση
[106]	Εκστρατεία-Χρήστες	Ναι	Ναι	-	Κεντρικότητες
[108]	Εκστρατεία-Χρήστες	Ναι	Ναι	IC	Δειγματοληψία
[109]	Εκστρατεία-Χρήστες	Ναι	Όχι	IC	Προσομοίωση
[110]	Εκστρατεία-Χρήστες	Ναι	Ναι	IC	Μονοπάτια
[111]	Εκστρατεία-Χρήστες	Ναι	Ναι	IC	Μονοπάτια
[112]	Εκστρατεία-Χρήστες	Ναι	Ναι	SIR	Κεντρικότητες
[113]	Προστασία	Ναι	Ναι	LT, IC	Δειγματοληψία
[114]	Προστασία	Ναι	Ναι	LT, IC	Δειγματοληψία
[115]	Προστασία	Ναι	Όχι	IC	Προσομοίωση
[116]	Προστασία	Ναι	Όχι	IC	Προσομοίωση

6 Πρόβλημα της από κοινού βελτιστοποίησης της διάδοσης της ψευδούς και της αληθούς πληροφορίας

6.1 Μοντέλο Συστήματος

6.1.1 Κοινωνικό Δίκτυο

Το κοινωνικό δίκτυο αναπαρίσταται ως ένας κατευθυνόμενος γράφος με βάρη χωρίς πολλαπλές ακμές $G = (V, E, \mathbf{w})$, όπου $V = \{v_1, v_2, ..., v_n\}$ το σύνολο των χρηστών και $(u, v) \in E$ η σχέση μεταξύ των χρηστών κατά την οποία ο χρήστης u ασκεί επιρροή στο χρήστη v μέσω αναρτήσεων, μηνυμάτων, κ.ο.κ., εντός της κοινωνικής πλατφόρμας. Η ισχύς της εν λόγω επιρροής εκφράζεται με τη συνάρτηση $w : V^2 \to [0, 1]$. Συνεπώς, για $\forall(u, v) \in E$, η τιμή $w(u, v)$ είναι η πιθανότητα να επηρεασθεί ο χρήστης v από το χρήστη u, ενώ για $\forall(u, v) \notin E$ είναι προφανές ότι $w(u, v) = 0$.

6.1.2 Διάδοση πληροφορίας

Θεωρούνται δύο κλάσεις πληροφορίας I_T και I_F σχετικές με μια θεματική ενότητα I. Σε καθεμία εξ αυτών περιλαμβάνονται αναρτήσεις, μηνύματα, κ.α., ανάλογα με το αν το περιεχόμενό τους σχετικά με το θέμα I είναι αληθές ή ψευδές αντίστοιχα. Βέβαια, αξίζει να σημειωθεί ότι αντικείμενο μελέτης αποτελούν οι ειδήσεις που δεν είναι μεταξύ τους αντικρουόμενες. Για παράδειγμα, αν οριστεί ως θεματική κατηγορία "Πόλεμος στην Ουκρανία", οι ειδήσεις "Κατάρριψη των πύργων τηλεπικοινωνιών της χώρας" και "Κατάληψη του Κιέβου από τη Ρωσία" δεν είναι αντικρουόμενες, καθώς είναι δυνατή η ταυτόχρονη πραγματοποίηση του περιεχομένου τους. Έτσι, ακόμα κι αν θεωρηθεί η πρώτη αληθής και η δεύτερη ψευδής, η διάδοση έκαστης γίνεται ανεξάρτητα μέσα στο δίκτυο, και έτσι ένας χρήστης μπορεί να υιοθετήσει τη μία εκ των δύο ή και τις δύο ή καμία εξ αυτών.

Επιπλέον, ορίζεται μια μετρική της γνώσης ή εξοικείωσης του χρήστη με τη θεματική κατηγορία I και συμβολίζεται ως $e : V \to [0, 1]$, όπου η τιμή 0 εκφράζει την πλήρη άγνοια ενώ η τιμή 1 την αυθεντία.

Θα μελετηθούν 3 διαφορετικά μοντέλα διάδοσης: Independent Cascade Model, Linear Threshold Model και Ντετερμινιστικό Linear Threshold Model.

6.1.2.1 Independent Cascade Model

Το μοντέλο IC παρουσιάστηκε αναλυτικά στην ενότητα 4.1.1. Στην παρούσα εργασία θα μελετηθεί μια επέκτασή του, ώστε να συμπεριλαμβάνεται στη διάδοση της πληροφορίας και η a priori γνώση του χρήστη. Συνεπώς, διαχωρίζεται η διάδοση της πληροφορίας σε ένα χρήστη $u \in V$ σε δύο στάδια, τη μετάδοση μέσω μιας σύνδεσης $(\cdot, u) \in E$ και την αποδοχή από τον κόμβο u ανάλογα με το επίπεδο της εξειδίκευσής

του, $e(u)$. Το πρώτο στάδιο διενεργείται ακριβώς όπως περιγράφηκε στην ενότητα 4.1.1, ενώ το δεύτερο λαμβάνει χώρα μετά την επιτυχία του πρώτου αλλά ανεξάρτητα από αυτό, και διέπεται από τις παρακάτω αρχές:

- Η X_u για κάθε $u \in V$ είναι μια τυχαία μεταβλητή για την οποία ισχύει:

$$X_u = \begin{cases} 1 & \text{αν ο χρήστης } u \text{ υιοθετήσει την είδηση.} \\ 0 & \text{αλλιώς.} \end{cases} \tag{13}$$

- Η πιθανότητα ο χρήστης u να υιοθετήσει μια είδηση i της κλάσης I_T εξαρτάται από τη γνώση του σχετικά με το θέμα I σύμφωνα με την παρακάτω εξίσωση:

$$p_T(u) = Pr\{X_u = 1 | i \in I_T\} = \max\{e(u), 1 - e(u)\} \tag{14}$$

  Με αυτό τον τρόπο, γίνεται φανερό ότι τόσο οι γνώστες όσο και οι αδαείς χρήστες σχετικά με το θέμα I έχουν προδιάθεση να υιοθετήσουν μια σχετική αληθή είδηση. Ωστόσο, υπάρχει πιθανότητα να την απορρίψουν, $q_T(u) = 1 - p_T(u)$, η οποία μπορεί να αποδοθεί σε τυχόντα δισταγμό του χρήστη να αποδεχθεί ακόμα και την αλήθεια.

- Η πιθανότητα ο χρήστης u να υιοθετήσει μια είδηση i της κλάσης I_F εξαρτάται από τη γνώση του σχετικά με το θέμα I σύμφωνα με την παρακάτω εξίσωση:

$$p_F(u) = Pr\{X_u = 1 | i \in I_F\} = 1 - e(u) \tag{15}$$

  Επομένως, γίνεται φανερό ότι οι γνώστες του θέματος I θα αποδεχθούν τη σχετική ψευδή είδηση με μικρή πιθανότητα εν αντιθέσει με τους αδαείς που εμφανίζουν μεγαλύτερη τάση να την υιοθετήσουν λόγω της έλλειψης μηχανισμών ή/και γνώσεων που θα τους προέτρεπαν να διακρίνουν την αναλήθειά της.

Τα βήματα Μετάδοσης και Αποδοχής της πληροφορίας μπορούν μαθηματικά να συμπτυχθούν σε ένα με τον ακόλουθο τρόπο. Έστω $C \in \{T, F\}$ η κλάση ειδήσεων στην οποία ανήκει η είδηση που διαδίδεται στο δίκτυο. Εφόσον, τα δύο προαναφερθέντα βήματα είναι ανεξάρτητα μεταξύ τους, η πιθανότητα ενεργοποίησης ενός κόμβου u τη χρονική στιγμή $t+1$ από ένα για πρώτη φορά τη χρονική στιγμή t ενεργό εισερχόμενο γείτονα $v \in N_{in}(u)$ εκφράζεται κάτωθι:

$$p_C(v, u) = w(v, u) \cdot p_C(u) \tag{16}$$

Δηλαδή με αυτόν τον τρόπο, θα μπορούσε κανείς να προσομοιώσει τις διαδόσεις ειδήσεων που ανήκουν στις κλάσεις I_T και I_F ορίζοντας δύο νέους γράφους $G_T = (V, E, p_T)$ και $G_F = (V, E, p_F)$ αντίστοιχα και εφαρμόζοντας τις αρχές του κλασικού μοντέλου IC. Το γεγονός ότι πλέον η διάδοση μελετάται σε διαφορετικούς γράφους δε φέρει σημασία αφού ούτως ή άλλως οι μεταδιδόμενες ειδήσεις δεν είναι ανταγωνιστικές μεταξύ τους.

6.1.2.2 Linear Threshold Model

Καθώς στόχο της παρούσας εργασίας και των ακολούθων πειραμάτων συναποτελεί η μελέτη του πιθανοτικού μοντέλου LT που παρουσιάστηκε στην ενότητα 4.1.2, στη συγκεκριμένη περίπτωση δε θα εισαχθεί στο μοντέλο διάδοσης η έννοια της εξειδίκευσης του χρήστη και επομένως θα ακολουθηθούν οι αρχές ακριβώς όπως αυτές περιγράφησαν στην ενότητα 4.1.2. Δηλαδή, η εξέλιξη της εξάπλωσης μιας πληροφορίας στο δίκτυο θα είναι ως προς τη δυναμική ίδια ανεξάρτητα από το αν ανήκει στην κλάση I_T ή στην κλάση I_F, αλλά φυσικά το τελικό αποτέλεσμά της θα διαμορφωθεί ανάλογα με το σύνολο των αρχικών ενστερνιστών A της εν λόγω ιδέας.

6.1.2.3 Ντετερμινιστικό Linear Threshold Model

Το ντετερμινιστικό Linear Threshold Model (ή εν συντομία DLT) διαφέρει από αυτό που αναφέρθηκε στην ενότητα 4.1.2 ως προς το γεγονός ότι τα κατώφλια των κόμβων $\theta(\cdot)$ επιλέγονται ντετερμινιστικά σύμφωνα με κάποια a priori γνώση ή εμπειρικά μέσω ερευνών και τεχνικών εξόρυξης δεδομένων (data mining) [117], κι όχι τυχαία σύμφωνα με την ομοιόμορφη κατανομή από το διάστημα $[0, 1]$. Προκειμένου να ενσωματωθεί σε αυτό το μοντέλο η γνώση ενός χρήστη σχετικά με τη θεματική κατηγορία I, κάθε κόμβος $u \in V$ διαθέτει δύο κατώφλια θ_T και θ_F ανάλογα με την κλάση στην οποία ανήκει η διαδιδόμενη είδηση. Συνεπώς, για κάθε κόμβο $u \in V$ ορίζονται τα δύο κάτωθι:

- Το κατώφλι για την αποδοχή της είδησης $i \in I_T$:

$$\theta_T(u) = \min\{e(u), 1 - e(u)\} \tag{17}$$

  Γίνεται εμφανές ότι όταν πρόκειται για πληροφορία που ανήκει στην κλάση αλήθειας I_T, τόσο ο γνώστης όσο και ο αδαής θα την αποδεχθούν με μεγαλύτερη ευκολία, με αυτό να αποτυπώνεται στην ανάγκη είτε για λιγότερους ενεργούς εισερχόμενους γείτονες είτε για μικρότερες πιθανότητες μετάδοσης από έκαστο ενεργό εισερχόμενο γείτονα, αφού το κατώφλι θα είναι σχετικά χαμηλό. Φυσικά, υφίσταται η πιθανότητα να απορριφθεί η είδηση, όταν το κατώφλι δεν είναι μηδενικό, και εκφράζει το δισταγμό του χρήστη να υιοθετήσει την οποιαδήποτε πληροφορία, ακόμα κι αν είναι ακριβής.

- Το κατώφλι για την αποδοχή της είδησης $i \in I_F$:

$$\theta_F(u) = e(u) \tag{18}$$

  Είναι επόμενο ότι για ένα χρήστη με υψηλή γνώση επί του θέματος I θα απαιτηθεί πολλή επιρροή από τους ενεργούς εισερχόμενους γείτονες του ούτως ώστε να αποδεχθεί την αναληθή είδηση i. Αντίθετα, ο χρήστης με έλλειψη σχετικής εξειδίκευσης μπορεί εύκολα να αποδεχθεί την ψευδή πληροφορία αφού δεν κατέχει το κατάλληλο γνωστικό υπόβαθρο ώστε να κρίνει την ορθότητά της.

Αξιοποιώντας τα παραπάνω, θα μπορούσε κανείς να προσομοιώσει τις διαδόσεις ειδήσεων που ανήκουν στις κλάσεις I_T και I_F ορίζοντας δύο νέους γράφους $G_T = (V, E, \mathbf{w})$ με κατώφλι $\theta_T(u), \forall u \in V$ και $G_F = (V, E, \mathbf{w})$ με κατώφλι $\theta_F(u), \forall u \in V$ αντίστοιχα, και εφαρμόζοντας τις αρχές διάδοσης του κλασικού μοντέλου LT. Όπως και στο μοντέλο IC, το γεγονός ότι πλέον η διάδοση μελετάται σε διαφορετικούς γράφους δεν επηρεάζει την περαιτέρω μελέτη χάρη στη μη ανταγωνιστική σχέση των διαδιδόμενων ειδήσεων.

6.2 Ορισμός του προβλήματος

Εν γένει το μελετούμενο πρόβλημα με όνομα Cautious Misinformation Minimization (CMM) ορίζεται για πρώτη φορά από όσο γνωρίζουμε ως η αφαίρεση συνδέσεων-ακμών πλήθους το πολύ k με στόχο τον περιορισμό της παραπληροφόρησης και ταυτόχρονα την ελάχιστη δυνατή μείωση της διάδοσης της ορθής πληροφορίας στο κοινωνικό δίκτυο. Ως συνέχεια του προηγούμενου παραδείγματος με θεματική ενότητα I = "Πόλεμος στην Ουκρανία", αν θεωρηθεί ότι η είδηση i_1 = "Κατάρριψη των πύργων τηλεπικοινωνιών της χώρας" είναι αληθής (δηλ. $i_1 \in I_T$) και η είδηση i_2 = "Κατάληψη του Κιέβου από τη Ρωσία" είναι ψευδής (δηλ. $i_2 \in I_F$), τότε είναι επιθυμητή η αφαίρεση ενός συνόλου ακμών E' με $|E'| \leq k$ που θα έχει ως αποτέλεσμα μεν να μειωθεί κατά το μέγιστο δυνατό το πλήθος των αποδεκτών της είδησης i_2 δε να μειωθεί κατά το ελάχιστο δυνατό το πλήθος των αποδεκτών της είδησης i_1. Η ψευδής πληροφορία i_2 θα επέφερε πανικό τόσο σε τοπική όσο και σε παγκόσμια κλίμακα λόγω της αποσταθεροποίησης ενός ολόκληρου κράτους, ενώ η έλλειψη ενημέρωσης σχετικά με την αληθή πληροφορία i_1 θα προκαλούσε ανησυχία και αναστάτωση σε όσους προσπαθούσαν να επικοινωνήσουν με άτομα εντός της Ουκρανίας. Παρακάτω θα παρουσιαστούν τρεις παραλλαγές του συγκεκριμένου προβλήματος ανάλογα με το υιοθετούμενο μοντέλο διάδοσης:

- Υπό το μοντέλο διάδοσης LT:
  Ως ευπάθεια ενός δικτύου $G = (V, E)$ στη διάδοση πληροφοριών από ένα σύνολο αρχικών υποστηρικτών A ορίζεται το άθροισμα των αναμενόμενων τιμών των επιρροών $\sigma(u, G), u \in A$ [6]:
  $$\sum_{u \in A} \sigma(u, G) \tag{19}$$
  Δοθέντων ενός δικτύου $G = (V, E, \mathbf{w})$, των δύο κλάσεων πληροφορίας I_T, I_F και των αντίστοιχων τρόπων διάδοσής τους υπό το μοντέλο LT (ενότητα 6.1.2.2), τα σύνολα των αρχικών ενστερνιστών της κλάσης I_T και της κλάσης I_F, S_T και S_F αντίστοιχα, και ενός θετικού ακέραιου αριθμού k, στόχος είναι η εύρεση ενός συνόλου ακμών $E' \subseteq E, |E'| \leq k$ των οποίων η αφαίρεση θα επιφέρει τη μέγιστη δυνατή μείωση της διάδοσης της κλάσης I_F με την ταυτόχρονη ελάχιστη δυνατή μείωση της διάδοσης της κλάσης I_T. Το εν λόγω πρόβλημα ποσοτικοποιώντας τη διάδοση μέσω της ευπάθειας του δικτύου εκφράζεται

μαθηματικά κάτωθι:

$$E' = \underset{E', E' \subseteq E, |E'| \leq k}{\arg\min} \{ \sum_{a \in S_T} \sigma(a, (V, E)) - \sum_{a \in S_T} \sigma(a, (V, E \backslash E')) + \sum_{b \in S_F} \sigma(b, (V, E \backslash E')) \} \tag{20}$$

Η αντικειμενική συνάρτηση του προβλήματος, της οποίας επιθυμητή είναι η ελαχιστοποίηση, αποτυπώνεται μαθηματικά σύμφωνα με τα παραπάνω ως:

$$f_{LT}(E') = \sum_{a \in S_T} \sigma(a, (V, E)) - \sum_{a \in S_T} \sigma(a, (V, E \setminus E')) + \sum_{b \in S_F} \sigma(b, (V, E \setminus E')) \tag{21}$$

- Υπό τα μοντέλα διάδοσης IC και DLT:
  Όπως αναφέρθηκε στην Ενότητα 4.2.1, η επιρροή ενός συνόλου αρχικών υποστηρικτών A σε ένα δίκτυο $G = (V, E)$ ορίζεται ως το αναμενόμενο πλήθος των ενεργών κόμβων στο τέλος της διάδοσης της πληροφορίας και συμβολίζεται ως $\sigma(A, G)$. Φυσικά, στην περίπτωση του μοντέλου DLT το εν λόγω πλήθος υπολογίζεται με ακρίβεια ντετερμινιστικά.

  Δοθέντων ενός δικτύου $G = (V, E, \mathbf{w})$, των δύο κλάσεων πληροφορίας I_T, I_F και των αντίστοιχων τρόπων διάδοσής τους υπό το μοντέλο IC (ενότητα 6.1.2.1) ή το μοντέλο DLT (ενότητα 6.1.2.3), τα σύνολα των αρχικών ενστερνιστών της κλάσης I_T και της κλάσης I_F, S_T και S_F αντίστοιχα, και ενός θετικού ακέραιου αριθμού k, στόχος είναι η εύρεση ενός συνόλου ακμών $E' \subseteq E, |E'| \leq k$ των οποίων η αφαίρεση θα επιφέρει τη μέγιστη δυνατή μείωση της διάδοσης της κλάσης I_F με την ταυτόχρονη ελάχιστη δυνατή μείωση της διάδοσης της κλάσης I_T. Το εν λόγω πρόβλημα ποσοτικοποιώντας τη διάδοση μέσω της επιρροής ενός συνόλου εκφράζεται μαθηματικά κάτωθι:

  $$E' = \underset{E', E' \subseteq E, |E'| \leq k}{\arg\min} \{ \sigma(S_T, (V, E)) - \sigma(S_T, (V, E \setminus E')) + \sigma(S_F, (V, E \setminus E')) \} \tag{22}$$

  Η αντικειμενική συνάρτηση του προβλήματος, της οποίας επιθυμητή είναι η ελαχιστοποίηση, αποτυπώνεται μαθηματικά σύμφωνα με τα παραπάνω ως:

  $$f_{IC/DLT}(E') = \sigma(S_T, (V, E)) - \sigma(S_T, (V, E \setminus E')) + \sigma(S_F, (V, E \setminus E')) \tag{23}$$

6.3 Δυσκολία επίλυσης του προβλήματος

6.3.1 Δυσκολία προβλήματος υπό το μοντέλο IC

Στη συγκεκριμένη ενότητα, θα αναλυθεί η δυσκολία επίλυσης του προβλήματος CMM υπό το μοντέλο IC μέσω μιας σειράς αναγωγών από NP-Complete προβλήματα.

Ορισμός 1. *Το πρόβλημα Max Cut ορίζεται ως:*
Δοθέντος ενός μη-κατευθυνόμενου γράφου $G = (V, E)$ και ενός θετικού ακέραιου αριθμού k, ζητείται αν υφίσταται διαμερισμός των κόμβων του γράφου σε δύο σύνολα S

και T, έτσι ώστε το σύνολο ακμών $C = \{(u, v) \in E : u \in S, v \in T$ ή $v \in S, u \in T\}$ να έχει πληθικό αριθμό τουλάχιστον k, δηλαδή $|C| \geq k$.

Ορισμός 2. *Το πρόβλημα Max Directed Cut ορίζεται ως:*
Δοθέντος ενός κατευθυνόμενου γράφου $G = (V, E)$ και ενός θετικού ακέραιου αριθμού k, ζητείται αν υφίσταται διαμερισμός των κόμβων του γράφου σε δύο σύνολα S και T, έτσι ώστε το σύνολο ακμών $C = \{(u, v) \in E : u \in S, v \in T\}$ να έχει πληθικό αριθμό τουλάχιστον k, δηλαδή $|C| \geq k$.

Θεώρημα 1. *Το πρόβλημα Max Directed Cut είναι NP-Complete.*

Απόδειξη. Θα γίνει αναγωγή από το πρόβλημα Max Cut, το οποίο είναι NP-Complete [118]. Δοθέντος ενός μη κατευθυνόμενου γράφου $G = (V, E)$ παράγεται σε πολυωνυμικό χρόνο ο αντίστοιχος κατευθυνόμενος γράφος $G' = (V, E'), E' = \{(u, v), (v, u) : (u, v) \in E\}$, δηλαδή κάθε μη κατευθυνόμενη ακμή αντικαθίσταται από δύο αντίθετης κατεύθυνσης κατευθυνόμενες ακμές. Αν υπάρχει διαμερισμός S, T ως λύση για το πρόβλημα Max Cut στο γράφο G με $|C| \geq k$, ο ίδιος διαμερισμός αποτελεί λύση και για το πρόβλημα Max Directed Cut για το γράφο G', αφού κάθε μη κατευθυνόμενη ακμή του γράφου G $(u, v) \in E, u \in S, v \in T$ ή $v \in S, u \in T$ αντιστοιχεί σε δύο κατευθυνόμενες ακμές στο γράφο G', από τις οποίες όμως μόνο η μία ακμή $(u, v), u \in S, v \in T$ προσμετράται στο σύνολο C. Αντίστροφα, αν υπάρχει διαμερισμός S, T ως λύση για το πρόβλημα Max Directed Cut στο γράφο G' με $|C| \geq k$, ο ίδιος διαμερισμός αποτελεί λύση και για το πρόβλημα Max Cut για το γράφο G, αφού κάθε κατευθυνόμενη ακμή $(u, v), u \in S, v \in T$ αντιστοιχεί σε μία μη-κατευθυνόμενη ακμή στο γράφο G, (u, v) ή (v, u). Επομένως, το σύνολο των μετρούμενων ακμών μεταξύ των συνόλων S, T είναι ίδιας πληθικότητας και στους δύο γράφους. Επιπλέον, δοθείσης μιας πιθανής λύσης του προβλήματος Max Directed Cut απαιτείται πολυωνυμικός χρόνος για τον έλεγχο της ορθότητάς της, αφού αρκεί ο έλεγχος της πληθικότητας του συνόλου των ακμών μεταξύ των συνόλων S και T. Επομένως, το πρόβλημα Max Directed Cut είναι NP-Complete. □

Ορισμός 3. *Το πρόβλημα Max Bisection ορίζεται ως:*
Δοθέντος ενός κατευθυνόμενου γράφου $G = (V, E)$ και ενός θετικού ακέραιου αριθμού k, ζητείται αν υφίσταται διαμερισμός των κόμβων του γράφου σε δύο σύνολα S και T, έτσι ώστε $|S| = |T| = \frac{|V|}{2}$ και το σύνολο ακμών $C = \{(u, v) \in E : u \in S, v \in T\}$ να έχει πληθικό αριθμό τουλάχιστον k, δηλαδή $|C| \geq k$.

Θεώρημα 2. *Το πρόβλημα Max Bisection είναι NP-Complete.*

Απόδειξη. Είναι δυνατή η αναγωγή στο πρόβλημα Max Bisection από το πρόβλημα Max Directed Cut προσθέτοντας σε πολυωνυμικό χρόνο $|V|$ απομονωμένους κόμβους, παράγοντας έτσι ένα νέο γράφο $G' = (V', E)$. Είναι προφανές ότι $|V'| = 2|V|$ κι ότι εφόσον οι νέοι κόμβοι είναι απομονωμένοι, το αποτέλεσμα της εύρεσης του ζητούμενου διαμερισμού των κόμβων στα σύνολα S, T με βάση τις μεταξύ τους ακμές είναι το ίδιο είτε στο γράφο G είτε στο γράφο G'. Έτσι, έστω ότι βρίσκεται η λύση

του προβλήματος για το γράφο G με τα σύνολα S, T, τότε η λύση του προβλήματος στο γράφο G' επάγεται από τα σύνολα S, T μοιράζοντας τους νέους απομονωμένους κόμβους του G' στα S, T με τέτοιο τρόπο, ώστε $|S| = |T| = |V|$. Αντίστροφα, αν υπάρχει η λύση του προβλήματος για το γράφο G' με τα σύνολα S, T, τότε κανείς βρίσκει τη λύση του προβλήματος για το γράφο G αφαιρώντας από τα S, T τους νέους απομονωμένους κόμβους του G'. Επίσης, δοθείσης μιας πιθανής λύσης του προβλήματος Max Bisection απαιτείται πολυωνυμικός χρόνος για τον έλεγχο της ορθότητάς της, αφού αρκεί ο έλεγχος της πληθικότητας των δοθέντων συνόλων S, T και του συνόλου των ακμών μεταξύ αυτών. Επομένως, το πρόβλημα Max Bisection είναι NP-Complete. □

Ορισμός 4. *Το πρόβλημα Bisection Width ορίζεται ως:*
Δοθέντος ενός κατευθυνόμενου γράφου $G = (V, E)$ και ενός θετικού ακέραιου αριθμού k, ζητείται αν υφίσταται διαμερισμός των κόμβων του γράφου σε δύο σύνολα S και T, έτσι ώστε $|S| = |T| = \frac{|V|}{2}$ και το σύνολο ακμών $C = \{(u, v) \in E : u \in S, v \in T\}$ να έχει πληθικό αριθμό το πολύ k, δηλαδή $|C| \leq k$.

Θεώρημα 3. *Το πρόβλημα Bisection Width είναι NP-Complete.*

Απόδειξη. Είναι δυνατή η αναγωγή του προβλήματος Bisection Width από το πρόβλημα Max Bisection με τον παρακάτω τρόπο. Έστω ότι στο πρόβλημα Max Bisection ορίζεται ο γράφος $G = (V, E)$ όπου $|V|$ άρτιος αριθμός. Τότε, για το πρόβλημα Bisection Width εξάγεται σε πολυωνυμικό χρόνο ο συμπληρωματικός γράφος του G, $\overline{G} = (V, E'), E' = \{(u, v) : u, v \in V, (u, v) \notin E\}$ και ορίζεται ως νέο $k' = (\frac{|V|}{2})^2 - k$. Σε αυτό το σημείο πρέπει να γίνει αντιληπτό το γεγονός ότι αν υπάρχει διαμερισμός των κόμβων του γράφου G σε ισοπληθή σύνολα S, T με $|C| = k$, τότε για τον αντίστοιχο διαμερισμό του γράφου $\overline{G}$ στα σύνολα S, T ισχύει $|C'| = (\frac{|V|}{2})^2 - k$. Αυτό οφείλεται στο γεγονός ότι στο γράφο G, από το σύνολο S εξέρχονται k ακμές που έχουν το τέλος τους στο σύνολο T, με αποτέλεσμα στο γράφο $\overline{G}$ να εξέρχονται από το σύνολο S ακμές με τέλος στο σύνολο T, πλήθους $(\frac{|V|}{2})^2$(το μέγιστο δυνατό πλήθος τέτοιων ακμών) $- k$ (το πλήθος των ακμών μεταξύ των συνόλων S, T στο γράφο G που δεν υπάρχουν στο γράφο $\overline{G}$). Συνεπώς, αν το πρόβλημα Max Bisection έχει λύση για το γράφο G τα σύνολα S, T με $|C| \geq k$, τότε το πρόβλημα Bisection Width έχει λύση για το γράφο $\overline{G}$ τα σύνολα S, T με $|C'| \leq (\frac{|V|}{2})^2 - k$, και αντιστρόφως. Επίσης, δοθείσης μιας πιθανής λύσης του προβλήματος Bisection Width απαιτείται πολυωνυμικός χρόνος για τον έλεγχο της ορθότητάς της, αφού αρκεί ο έλεγχος της πληθικότητας των δοθέντων συνόλων S, T και του συνόλου των ακμών μεταξύ αυτών. Επομένως, το πρόβλημα Bisection Width είναι NP-Complete. □

Θεώρημα 4. *Το πρόβλημα CMM είναι NP-Hard υπό το μοντέλο διάδοσης IC.*

Απόδειξη. Θα γίνει αναγωγή από το πρόβλημα Bisection Width στο πρόβλημα Cautious Misinformation Minimization:

Έστω ο γράφος $G = (V, E), |V| = n$ (άρτιος αριθμός) για τον οποίο ζητείται να λυθεί το πρόβλημα Bisection Width. Είναι προφανές ότι $k \leq (\frac{n}{2})^2$, το πλήθος όλων των δυνατών κατευθυνόμενων ακμών από ένα σύνολο S σε σύνολο T, με $|S| = |T|$. Τότε, για το πρόβλημα CMM προστίθενται σε πολυωνυμικό χρόνο στο γράφο G, $(\frac{n}{2})^2 + 1$ νέοι κόμβοι, $A = \{s_1, s_2, ..., s_{(\frac{n}{2})^2+1}\}$, έκαστος εκ των οποίων συνδέεται με κάθε κόμβο $u \in V$, και ως εκ τούτου προκύπτει νέος γράφος $G' = (V', E'') = (V \cup A, E \cup \{(s, u) : s \in A, u \in V\})$. Πάνω σε αυτό το γράφο ορίζεται στιγμιότυπο του προβλήματος CMM ως εξής: Έστω το κοινωνικό δίκτυο $G' = (V', E'', w)$, για το οποίο ισχύει $w(u, v) = 1, \forall (u, v) \in E''$ και $e(u) = 0, \forall u \in V'$. Έτσι, σύμφωνα με τις εξισώσεις 14, 15, 16 ισχύει ότι $p_T(u, v) = p_F(u, v) = 1, \forall (u, v) \in E''$, με αποτέλεσμα η διάδοση πληροφοριών τόσο της κλάσης I_T όσο και της κλάσης I_F να είναι ντετερμινιστική. Επιπλέον, ορίζεται ότι τα σύνολα των αρχικών υποστηρικτών των κλάσεων I_T και I_F είναι $S_T = \emptyset$ και $S_F = A$, αντίστοιχα. Με αυτό τον τρόπο, η εξίσωση 22 απλοποιείται σε:

$$E' = \underset{E', E' \subseteq E'', |E'| \leq k'}{\arg\min} \{\sigma(A, (V', E'' \setminus E'))\} \tag{24}$$

Εφόσον ισχύει ότι ένα πρόβλημα βελτιστοποίησης (στη συγκεκριμένη περίπτωση ελαχιστοποίησης) είναι NP-Hard αν το αντίστοιχο πρόβλημα απόφασης (θέτοντας περιορισμό $\leq B$ στην αντικειμενική συνάρτηση) είναι NP-Hard, θα συνεχιστεί η απόδειξη με το πρόβλημα απόφασης CMM, δηλαδή ορίζοντας ως στόχο:

$$\sigma(A, (V', E'' \setminus E')) \leq B \tag{25}$$

Τότε, στο στιγμιότυπο του προβλήματος CMM ορίζονται επιπλέον $B = \frac{n}{2} + (\frac{n}{2})^2 + 1$ και $k' = k + \frac{n}{2} \cdot ((\frac{n}{2})^2 + 1)$. Πρακτικά, η λύση του εν λόγω προβλήματος απαιτεί το διαμερισμό του συνόλου των κόμβων V' σε δύο σύνολα S και T, τα οποία είναι το σύνολο των κόμβων που δέχθηκαν την είδηση της κλάσης I_F και το σύνολο των κόμβων που δεν τη δέχθηκαν αντίστοιχα. Είναι προφανές ότι οι κόμβοι A ως εκκινητές της διάδοσης της είδησης ανήκουν στο σύνολο S, για το οποίο ισχύει $|S| \leq B$. Προκειμένου να είναι δυνατός αυτός ο διαμερισμός, δηλαδή να μην έχει μολυνθεί ένας κόμβος v του συνόλου T είναι απαραίτητο να μην υπάρχει εισερχόμενος γείτονας u από το σύνολο S, διότι τότε λόγω του $p_F(u, v) = 1$ ο κόμβος v θα δεχόταν την ψευδή πληροφορία. Επομένως, απαιτείται η αφαίρεση κάθε ακμής (u, v) για την οποία ισχύει $u \in S, v \in T$. Φυσικά, θα πρέπει το πλήθος αυτών των ακμών να είναι το πολύ k'. Θα αποδειχθεί ότι υπάρχει λύση στο πρόβλημα Bisection Width αν και μόνο αν υπάρχει λύση στο πρόβλημα CMM. Έστω ότι υπάρχει λύση (S, T) για το πρόβλημα CMM που ικανοποιεί τις απαιτήσεις B και k'. Τότε, ισχύει $|S| = B$, αφού αν $|S| < B$, τότε λόγω του γεγονότος ότι οι κόμβοι $A, |A| = (\frac{n}{2})^2 + 1$ συνδέονται με όλους τους κόμβους V, το πλήθος των αφαιρεμένων ακμών θα ήταν τουλάχιστον:

$$\begin{aligned}(n + (\frac{n}{2})^2 + 1 - |S|) \cdot ((\frac{n}{2})^2 + 1) \geq (n + (\frac{n}{2})^2 + 1 - \frac{n}{2} - (\frac{n}{2})^2) \cdot ((\frac{n}{2})^2 + 1) \geq \\ \frac{n}{2} \cdot ((\frac{n}{2})^2 + 1) + ((\frac{n}{2})^2 + 1) > \frac{n}{2} \cdot ((\frac{n}{2})^2 + 1) + k = k'\end{aligned} \tag{26}$$

Δηλαδή θα παραβιαζόταν η μία απαίτηση. Έτσι, εφόσον για το πρόβλημα CMM στο γράφο $G' = (V', E'')$ υπάρχει ο διαμερισμός S, T με $|S| = B$ και $|E'| \leq k'$, τότε για το πρόβλημα Bisection Width στο γράφο $G = (V, E)$ υπάρχει διαμερισμός S', T' με $S' = S \setminus A, |S| = B - (\frac{n}{2})^2 + 1 = \frac{n}{2}, T' = T, |T'| = n - |S'| = \frac{n}{2}$, όπου το πλήθος των ακμών μεταξύ των συνόλων είναι το πολύ $k' - \frac{n}{2} \cdot ((\frac{n}{2})^2 + 1)$(το πλήθος των ακμών που εισάγουν ανάμεσα στα δύο σύνολα οι κόμβοι A) $= k$. Αντίστροφα, αν υπάρχει διαμερισμός S, T με $|S| = |T| = \frac{n}{2}$ που λύνει το πρόβλημα Bisection Width στο γράφο $G = (V, E)$, τότε είναι άμεση η εύρεση λύσης στο πρόβλημα CMM στο γράφο $G' = (V', E'')$. Αυτή θα ήταν τα σύνολα S', T' με $S' = S \cup A, |S| = \frac{n}{2} + (\frac{n}{2})^2 + 1 \leq B, T' = T$, όπου οι κόμβοι A θα εισήγαγαν ανάμεσα στα δύο σύνολα $((\frac{n}{2})^2 + 1) \cdot \frac{n}{2}$ ακμές, δηλαδή $|E'| \leq ((\frac{n}{2})^2 + 1) \cdot \frac{n}{2} + k = k'$. Επομένως, το πρόβλημα CMM είναι NP-Hard. □

6.3.2 Δυσκολία προβλήματος υπό το μοντέλο LT

Αρχικά, θα περιγραφούν κάποιες σημαντικές έννοιες τόσο για την απόδειξη της δυσκολίας του προβλήματος CMM όσο και για την εν συνεχεία ανάπτυξη του ευριστικού αλγορίθμου επίλυσής του.

Ορισμός 5. *Έστω E ένα πεπερασμένο σύνολο. Μια σύναρτηση $f : 2^E \to \mathbf{R}$ είναι supermodular αν και μόνο αν για $\forall S \subseteq T \subset E, \forall e \in E \setminus T$ ισχύει:*

$$f(S \cup \{e\}) - f(S) \leq f(T \cup \{e\}) - f(T) \tag{27}$$

Ορισμός 6. *Έστω E ένα πεπερασμένο σύνολο. Μια σύναρτηση $f : 2^E \to \mathbf{R}$ είναι submodular αν και μόνο αν για $\forall S \subseteq T \subset E, \forall e \in E \setminus T$ ισχύει:*

$$f(S \cup \{e\}) - f(S) \geq f(T \cup \{e\}) - f(T) \tag{28}$$

Είναι εμφανές ότι αν μια συνάρτηση f είναι submodular τότε η συνάρτηση $-f$ είναι supermodular, κι αντιστρόφως.

Ορισμός 7. *Μια σύναρτήση $f : 2^E \to \mathbf{R}$ είναι μονότονη ή μη φθίνουσα αν και μόνο αν για $\forall S \subset E, \forall e \in E \setminus S$ ισχύει:*

$$f(S \cup \{e\}) \geq f(S) \tag{29}$$

Στο πλαίσιο των παραπάνω ιδιοτήτων ισχύει το ακόλουθο ενδιαφέρον θεώρημα:

Θεώρημα 5. *Έστω μια μη αρνητική, μονότονη και submodular συνάρτηση $f : 2^E \to \mathbf{R}$. Έστω S το σύνολο μεγέθους k που προκύπτει με την επιλογή στοιχείων από το σύνολο E σε k επαναλήψεις, σε έκαστη εκ των οποίων επιλέγεται το στοιχείο του οποίου η προσθήκη στο τρέχον σύνολο S προκαλεί τη μεγαλύτερη αύξηση στην τιμή της f. Έστω S^* το σύνολο που μεγιστοποιεί την τιμή της συνάρτησης f μεταξύ όλων των δυνατών συνόλων μεγέθους k. Τότε, ισχύει:*

$$f(S) \geq (1 - \frac{1}{e}) \cdot f(S^*) \tag{30}$$

Δηλαδή, το σύνολο S εξασφαλίζει μια $(1-\frac{1}{e})$-προσέγγιση της βέλτιστης λύσης [119].

Βέβαια, αυτό προϋποθέτει να είναι δυνατός ο ακριβής υπολογισμός της συνάρτησης f. Ειδάλλως, για οποιοδήποτε $\epsilon > 0$, υπάρχει ένα $\gamma > 0$ τέτοιο ώστε αξιοποιώντας $(1+\gamma)$-προσεγγιστικές τιμές της f το σύνολο S να εξασφαλίζει μια $(1-\frac{1}{e}-\epsilon)$-προσέγγιση της βέλτιστης λύσης [5].

Όπως αναφέρθηκε στην Ενότητα 4.2.1, ο ακριβής υπολογισμός της επιρροής σ ενός κόμβου ή ενός συνόλου κόμβων είναι #P-Δύσκολος υπό το μοντέλο διάδοσης LT. Στο [5] έχει προταθεί ένας εναλλακτικός τρόπος υπολογισμός της επιρροής σ, του οποίου η χρησιμότητα θα φανεί τόσο στην απόδειξη της δυσκολίας του προβλήματος CMM όσο και στον αλγόριθμο επίλυσής του.

Live-edge Γράφοι

Η ιδέα των live-edge γράφων προέκυψε από την ανάγκη για τη μαθηματική ανάλυση της επιρροής σ σε ένα γράφο $G = (V, E, \mathbf{w})$. Έτσι, ένας τυχαίος live-edge γράφος X υπό το μοντέλο LT κατασκευάζεται ως εξής: Ανεξάρτητα, $\forall v \in V$ επιλέγεται το πολύ μια εισερχόμενη ακμή (u, v) με πιθανότητα $w(u, v)$, ενώ δεν επιλέγεται καμία με πιθανότητα $1 - \sum_{u:(u,v)\in E} w(u, v)$. Έτσι, προκύπτει ο γράφος $X = (V, E_X)$, όπου $E_X \subseteq E$ το σύνολο των επιλεγμένων ζωντανών (live) ακμών. Στον εν λόγω γράφο η διάδοση ξεκινώντας από ένα αρχικό κόμβο u είναι ντετερμινιστική και αποτελείται από όλα τα μονοπάτια που αρχίζουν από τον κόμβο u και διασχίζουν live ακμές. Έστω $r(u, X)$ το πλήθος των κόμβων που συμμετέχουν σε αυτά τα μονοπάτια, δηλαδή των κόμβων που δέχθηκαν την πληροφορία που διέδωσε ο κόμβος u. Τότε, η επιρροή ενός κόμβου u υπολογίζεται ως εξής:

$$\sigma(u, G) = \mathbb{E}_X[r(u, X)] = \sum_{X \in \mathscr{X}_G} Pr[X|G] \cdot r(u, X), \tag{31}$$

όπου $\mathscr{X}_G$ το σύνολο όλων των δυνατών live-edge γράφων X που μπορούν να προκύψουν από το γράφο G και $Pr[X|G]$ η πιθανότητα κατάληξης στο συγκεκριμένο live-edge γράφο X δεδομένου του γράφου G. Η τελευταία μπορεί να υπολογιστεί ως το γινόμενο των πιθανοτήτων των κόμβων να έχουν εισερχόμενη ακμή στο δειγματοληπτημένο γράφο X, δηλαδή:

$$Pr[X|G] = \prod_{v \in V} p(v, X, G), \tag{32}$$

όπου

$$p(v, X, G) = \begin{cases} w(u, v), & \text{αν } \exists (u, v) \in E_X \\ 1 - \sum_{u:(u,v)\in E} w(u, v), & \text{αλλιώς} \end{cases} \tag{33}$$

Θεώρημα 6. *Η συνάρτηση επιρροής $\sigma(u, (V, E \backslash S))$ είναι supermodular και μονότονη (μη αύξουσα) ως προς το σύνολο αφαιρέσιμων ακμών S υπό το μοντέλο διάδοσης LT.*

Απόδειξη. Η απόδειξη παρουσιάζεται αναλυτικά στο [6] με χρήση της θεωρίας των live-edge γράφων και την αξιοποίηση σχετικών προτάσεων (propositions). □

Θεώρημα 7. *Το πρόβλημα CMM είναι NP-hard υπό το μοντέλο LT.*

Απόδειξη. Έστω ένα στιγμιότυπο του προβλήματος CMM, κατά το οποίο το σύνολο των αρχικών ενστερνιστών της κλάσης I_T είναι κενό, δηλαδή $S_T = \emptyset$. Τότε, η αντικειμενική συνάρτηση 21 απλοποιείται ως εξής:

$$f_{LT}(E') = \sum_{b \in S_F} \sigma(b, (V, E \setminus E')) \tag{34}$$

Με αυτόν τον τρόπο ακριβώς ορίζεται το πρόβλημα στο [6], όπου με χρήση μιας ισοδύναμης συνάρτησης $h(E') = \sum_{b \in S_F} \sigma(b, (V, E)) - \sum_{b \in S_F} \sigma(b, (V, E \setminus E'))$, η οποία εμφανώς είναι μονότονη (μη φθίνουσα), submodular και μη αρνητική, ζητείται η μεγιστοποίησή αυτής (δηλ. η μεγιστοποίηση της μείωσης της διάδοσης της παραπληροφόρησης), ούτως ώστε να αξιοποιηθεί το Θεώρημα 5. Τότε, η λύση του στιγμιοτύπου του προβλήματος CMM αποτελεί τη λύση του προβλήματος μεγιστοποίησης μιας τέτοιας συνάρτησης υπό περιορισμό στην πληθικότητα της λύσης, το οποίο είναι NP-Hard [120], αφού γνωστά NP-Hard προβλήματα, όπως το πρόβλημα max-k-cover [121] και το πρόβλημα knapsack [122], μπορούν να εκφραστούν κατά αυτό τον τρόπο. Συνεπώς, εφόσον ένα στιγμιότυπο του προβλήματος είναι NP-Hard, τότε και το γενικό πρόβλημα CMM είναι NP-Hard. □

6.3.3 Δυσκολία προβλήματος υπό το μοντέλο DLT

Θεώρημα 8. *Το πρόβλημα CMM είναι NP-hard υπό το μοντέλο DLT.*

Απόδειξη. Η απόδειξη είναι ίδια με αυτή του Θεωρήματος 4 με την εξής διαφορά: Στο γράφο $G' = (V', E'', \mathbf{w})$ για κάθε ακμή $(u, v) \in E''$ ισχύει $w(u, v) > 0$, και για κάθε κόμβο $u \in V'$ ισχύει ότι $e(u) = 0$, με αποτέλεσμα σύμφωνα με τις εξισώσεις 17, 18 οι τιμές των κατωφλίων να διαμορφώνονται ως $\theta_T(u) = \theta_F(u) = 0$. Δηλαδή πάλι η διάδοση πληροφοριών τόσο της κλάσης I_T όσο και της κλάσης I_F είναι ντετερμινιστική, αφού η ενεργοποίηση οποιουδήποτε εισερχόμενου γείτονα ενός κόμβου $u \in V'$ επιφέρει σύμφωνα με την εξίσωση 10 την ενεργοποίηση του τελευταίου. □

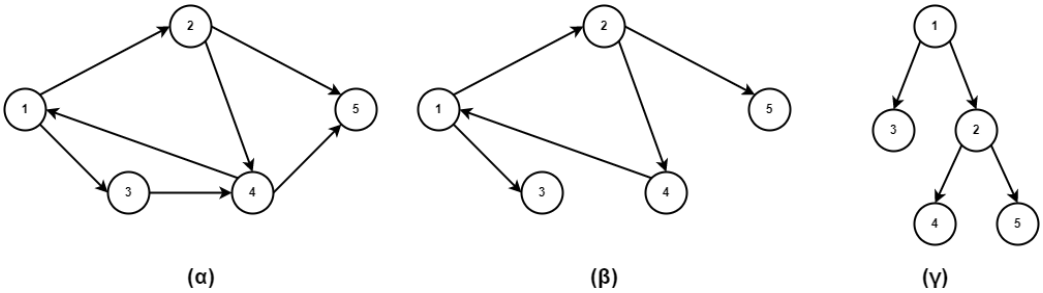
6.4 Προτεινόμενος αλγόριθμος

Στην παρούσα διπλωματική εργασία προτείνεται ένας επαναληπτικός άπληστος αλγόριθμος για την αντιμετώπιση του προβλήματος CMM υπό τα μοντέλα LT και DLT. Η βασική ιδέα του αλγορίθμου είναι η αφαίρεση ακμών σε k ή λιγότερες επαναλήψεις, σε καθεμία εκ των οποίων επιλέγεται η ακμή που προκαλεί τη μέγιστη μείωση της αντικειμενικής συνάρτησης 21 ή 23. Πιο συγκεκριμένα θα παρουσιαστούν δύο εκδοχές του αλγορίθμου ανάλογα με το μοντέλο διάδοσης:

6.4.1 Αλγόριθμος υπό το μοντέλο LT

Στην περίπτωση του μοντέλου LT, θα αξιοποιηθούν ιδέες και δομές δεδομένων που προτάθηκαν στα [5, 6]. Όπως προκύπτει από τον τρόπο κατασκευής τους, στους

live-edge γράφους κάθε κόμβος έχει το πολύ μια εισερχόμενη ακμή. Επομένως, από ένα κόμβο προς έναν άλλο υφίσταται το πολύ ένα μονοπάτι. Αυτό αποδεικνύεται εύκολα ως εξής: Έστω δύο κόμβοι $u, v \in V$, για τους οποίους υπάρχουν δύο διαφορετικά μονοπάτια, p_1, p_2 από τον u στον v. Αφού τα εν λόγω μονοπάτια είναι διαφορετικά, τότε σε ένα εξ αυτών, έστω στο p_2 συμμετέχει τουλάχιστον ένας κόμβος $w \in V$ που δε συμμετέχει στο άλλο, p_1, δηλαδή τα μονοπάτια ταυτίζονται μέχρι τον κόμβο $t \in V$, τον εισερχόμενο γείτονα του w, όπου διαχωρίζονται. Προκειμένου και τα δύο μονοπάτια να καταλήξουν στον τελικό κόμβο v, θα πρέπει είτε να ταυτιστούν από ένα κόμβο $s \in V$ μέχρι τον κόμβο v είτε να συναντηθούν στον τελικό κόμβο v. Με αυτόν τον τρόπο, θα απαιτείτο ο κόμβος s ή ο κόμβος v αντίστοιχα να έχει δύο εισερχόμενους γείτονες, το οποίο είναι αδύνατο εκ κατασκευής του γράφου X_G.
Λόγω της ύπαρξης των μοναδικών μονοπατιών, μπορεί κανείς να κατασκευάσει ένα δένδρο επιρροής T_X^u για κάθε κόμβο $u \in V$, εκτελώντας τον αλγόριθμο διάσχισης BFS της Ενότητας 2.4, στον οποίο επιλέγεται ως πρώτος κόμβος στο Βήμα 1 ο u. Το εν λόγω δένδρο είναι κατευθυνόμενο δένδρο με ρίζα τον κόμβο u, του οποίου οι ακμές είναι αυτές που εξετάζονται με τη σειρά στο Βήμα 3 κατά την εκτέλεση του αλγορίθμου BFS και δεν καταλήγουν σε προσπελαυσμένους κόμβους (με αυτό τον τρόπο εξασφαλίζεται η ακυκλικότητα του δένδρου). Στο Σχήμα 6 αποτυπώνεται η παραπάνω διαδικασία με στόχο την εύρεση του δένδρου επιρροής του κόμβου 1.



Σχήμα 6: Παραγωγή δένδρου επιρροής: (α) Αρχικός γράφος $G = (V, E, w)$, (β) Ένας τυχαίος live-edge γράφος X_G (γ) Δένδρο επιρροής του κόμβου 1 μέσω της εκτέλεσης BFS.

Με αυτόν τον τρόπο, το πλήθος των κόμβων που θα αποδεχθούν την πληροφορία της διάδοσης που ξεκινά από τον κόμβο $u \in V$ σε ένα live-edge γράφο X ορίζεται ισοδύναμα ως:

$$r(u, X) = r(u, T_X^u) \tag{35}$$

Επιπλέον, επειδή όπως αναφέρθηκε στην Ενότητα 4.2.1 ο υπολογισμός της επιρροής $\sigma(\cdot, G)$ είναι #P-Δύσκολος υπό το μοντέλο LT, για λόγους κλιμακωσιμότητας και εξοικονόμησης χρόνου η επιρροή ενός κόμβου θα υπολογίζεται προσεγγιστικά ως ο μέσος όρος των επιρροών του κόμβου σε ένα σύνολο τυχαίων live-edge γράφων,

δηλαδή η εξίσωση 31 μεταβάλλεται ως κάτωθι:

$$\sigma(u, G) \approx \bar{\sigma}(u, G) = \frac{1}{|\mathscr{X}_S|} \cdot \sum_{X_i \in \mathscr{X}_S} r(u, T^u_{X_i}), \tag{36}$$

όπου $\mathscr{X}_S = \{X_i : 1 \leq i \leq x_S\} \subseteq \mathscr{X}$ το σύνολο των δειγματοληπτημένων live-edge γράφων του γράφου G. Είναι εμφανές ότι η εφαρμογή ενός επαναληπτικού άπληστου αλγορίθμου, σε κάθε επανάληψη του οποίου επιλέγεται προς αφαίρεση η ακμή που επιφέρει τη μέγιστη μείωση ή αλλιώς οριακή απώλεια της αντικειμενικής συνάρτησης, επιβάλλει τον υπολογισμό της τελευταίας σε κάθε επανάληψη. Αυτό φαίνεται καθαρά μέσω της μαθηματικής έκφρασης του εν λόγω κριτηρίου, θεωρώντας E_t το σύνολο των ακμών που έχουν αφαιρεθεί μέχρι την τρέχουσα επανάληψη:

$$e = (u, v) = \arg\max_{e \in E \setminus E_t} \{\Delta(e|E_t)\} = \arg\max_{e \in E \setminus E_t} \{f_{LT}(E_t) - f_{LT}(E_t \cup \{e\}\} \tag{37}$$

Αν κανείς επιχειρήσει να εφαρμόσει το παραπάνω κριτήριο αφελώς, θα πρέπει σε κάθε επανάληψη $t \leq k$ να υπολογίζει για κάθε ακμή $e \in E \setminus E_t$ και κάθε κόμβο $s \in S_T \cup S_F$ την επιρροή $\bar{\sigma}(s, G)$ της εξίσωσης 36. Δηλαδή, θα πρέπει να υπολογίσει το νέο δένδρο επιρροής $T^s_{X_i}$ από το live-edge γράφο X_i, από τον οποίο έχουν αφαιρεθεί οι ακμές $E_t \cup \{e\}$, εφαρμόζοντας τον αλγόριθμο BFS πολυπλοκότητας $O(|V| + |E|)$. Αυτό θα επιφέρει χρονικό κόστος κάθε επανάληψης τουλάχιστον $O(x_S \cdot |E| \cdot (|V| + |E|))$, το οποίο κρίνεται απαγορευτικό σε μεγάλους γράφους.

Είναι εμφανές ότι η διαδικασία που εισάγει μεγάλο υπολογιστικό κόστος είναι η διάσχιση BFS των live-edge γράφων για την κατασκευή του δένδρου επιρροής και τον υπολογισμό της επιρροής ενός κόμβου. Ωστόσο, είναι δυνατή η αξιοποίηση των ιδιοτήτων των δένδρων προκειμένου να υπολογίζεται αποτελεσματικά η οριακή απώλεια στην εξίσωση 37. Σύμφωνα με την ιδιότητα 2 των δένδρων στην Ενότητα 2.5, κάθε ακμή ενός δένδρου είναι γέφυρα και η αφαίρεσή της επιφέρει την απώλεια της συνεντικότητας αυτού. Συνεπώς, σε ένα δένδρο επιρροής ενός κόμβου $s \in V$, $T^s_{X_i}$ η αφαίρεση μιας ακμής $e = (u, v) \in E \setminus E_t$ θα κοστίσει στην επιρροή του κόμβου s την επιρροή του κόμβου v και τον ίδιο τον κόμβο v, δηλαδή:

$$r(s, T^s_{X_i} \setminus E_t) - r(s, T^s_{X_i} \setminus (E_t \cup \{e\})) = r(v, T^s_{X_i} \setminus E_t) + 1 \tag{38}$$

Επομένως, αρκεί να είναι γνωστές οι επιρροές όλων των κόμβων $v \in V$ στο δένδρο επιρροής $T^s_{X_i}$. Αυτό μπορεί να επιτευχθεί με τη χρήση του αλγορίθμου BFS, ο οποίος πλέον έχει πολυπλοκότητα $O(|V|)$ λόγω της ιδιότητας 3 ενός δένδρου με n κόμβους να έχει $n - 1$ ακμές. Παρακάτω παρουσιάζεται σε μορφή ψευδοκώδικα ο τρόπος υπολογισμού της μετρικής $r(u, T^s_{X_i})$ για κάθε κόμβο $u \in V$, ο οποίος βασίζεται στο γεγονός ότι οι απόγονοι ενός κόμβου u αποτελούνται από τους εξερχόμενους γείτονες του κόμβου u και τους απογόνους τους:

```
Input: influence tree T^s_{X_i}
Variables: queue Q, stack H, set of traversed nodes visited
Output: number of descendants of vertex u  r(u, T^s_{X_i}), ∀u ∈ V
```

```
4
5  // initialization
6  for each vertex u ∈ V:
7      r(u, T^s_{X_i}) = 0
8
9  // BFS traversal
10 Q.enqueue(s)
11 visited = {s}
12 while Q is not empty:
13     u = Q.dequeue()
14     for each vertex v ∈ N_out(u) in T^s_{X_i}:
15         if v ∉ visited:
16             visited.add(v)
17             Q.enqueue(v)
18             H.push((u, v))
19
20 // compute influences
21 while H is not empty:
22     (u, v) = H.pop()
23     r(u, T^s_{X_i}) += r(v, T^s_{X_i}) + 1
```

Ομοίως, είναι αναγκαίος ο επαναϋπολογισμός των επιρροών όλων των κόμβων $u \in V$ στο δένδρο επιρροής $T^s_{X_i}$ μετά την αφαίρεση μιας ακμής $e = (u, v)$. Ένας κόμβος του δένδρου δύναται να επηρεαστεί από την αφαίρεση της ακμής με έναν εκ των δύο ακόλουθων τρόπων:

- Αν είναι ο κόμβος v ή απόγονος του κόμβου v, η επιρροή του μηδενίζεται αφού πλέον δεν υπάρχει μονοπάτι μεταξύ του κόμβου s και αυτού, ώστε να επηρεασθεί και να διαδώσει την πληροφορία.

- Αν είναι ο κόμβος u ή πρόγονος του κόμβου u, η επιρροή του μειώνεται κατά την επιρροή του κόμβου v και τη μονάδα (ο κόμβος v).

Έχοντας αυτά τα εργαλεία παρατίθεται κάτωθι ο συνολικός επαναληπτικός άπληστος αλγόριθμος για το πρόβλημα CMM.

```
1  Input: original graph G(V, E, w), set of seeds of true
       information S_T, set of seeds of false information S_F,
       maximum number of removed edges k
2  Variables: queue Q, stack H, live-edge graphs X_i, influence
       trees T^s_{X_i}, ∀s ∈ S_T ∪ S_F, number of influenced nodes by node u
       r(u, T^s_{X_i}), ∀u ∈ V, set of traversed nodes visited, marginal
       loss for true information's propagation Δ_T(e), ∀e ∈ E,
       marginal loss for false information's propagation
       Δ_F(e), ∀e ∈ E
3  Output: set of removed edges E'
4
5  // initialization
6  Sample a set of live-edge graphs 𝒳_S = {X_i : 1 ≤ i ≤ x_S} from G.
```

```
7  Create influence trees of true information's propagation
       trees_T = {T^s_{X_i} : X_i ∈ X_S, s ∈ S_T}.
8  Create influence trees of false information's propagation
       trees_F = {T^s_{X_i} : X_i ∈ X_S, s ∈ S_F}.
9
10 E' = ∅
11 for each edge e ∈ E:
12     Δ_T(e) = 0
13     Δ_F(e) = 0
14 for each vertex u ∈ V:
15     for each influence tree T^s_{X_i}:
16         r(u, T^s_{X_i}) = 0
17
18 // initial influences
19 for each influence tree T^s_{X_i} ∈ trees_T:
20     Q.enqueue(s)
21     visited = {s}
22     while Q is not empty:
23         u = Q.dequeue()
24         for each vertex v ∈ N_out(u) in T^s_{X_i}:
25             if v ∉ visited:
26                 visited.add(v)
27                 Q.enqueue(v)
28                 H.push((u, v))
29
30     while H is not empty:
31         (u, v) = H.pop()
32         r(u, T^s_{X_i}) += r(v, T^s_{X_i}) + 1
33         Δ_T((u, v)) += r(v, T^s_{X_i}) + 1
34
35 Repeat lines 19-33 for trees_F and Δ_F instead of trees_T and Δ_T
       respectively, to compute false information's propagation.
36
37 // greedy approach
38 while k is not 0:
39         Find edge e = (u, v) = arg max_{e∈E\E'}{Δ_F(e) − Δ_T(e)}.
40         if Δ_F(e) − Δ_T(e) ≤ 0:
41             Stop. // f_LT cannot be decreased more
42         else:
43             E' = E' ∪ {e}
44             k− = 1
45             for each influence tree T^s_{X_i} ∈ trees_T:
46                 curr = u
47                 while curr is not s:
48                     r(curr, T^s_{X_i}) −= r(v, T^s_{X_i}) + 1
49                     father = father of curr in T^s_{X_i}
50                     Δ_T((father, curr)) −= r(v, T^s_{X_i}) + 1
51                     curr = father
```

```
52
53                  r(v, T^s_{X_i}) = 0
54                  Q.enqueue(v)
55                  visited = {v}
56                  while Q is not empty:
57                      t = Q.dequeue()
58                      for each vertex w ∈ N_out(t) in T^s_{X_i}:
59                          if w ∉ visited:
60                              visited.add(w)
61                              Q.enqueue(w)
62                              Δ_T((t, w)) −= r(w, T^s_{X_i}) + 1
63                              r(w, T^s_{X_i}) = 0
64
65              Repeat lines 45-63 for trees_F and Δ_F instead of
66              trees_T and Δ_T respectively,to recompute false
67              information's propagation.
```

Ουσιαστικά τα βήματα που ακολουθούνται σύμφωνα με τον παραπάνω ψευδοκώδικα αποτυπώνονται ως εξής:

1. Υπολόγισε ένα σύνολο τυχαίων live-edge γράφων και τα αντίστοιχα δένδρα διάδοσης για κάθε κόμβο στο σύνολο των αρχικών ενστερνιστών της αληθούς πληροφορίας και για κάθε κόμβο στο σύνολο των αρχικών ενστερνιστών της ψευδούς πληροφορίας (Γραμμές 6-8)

2. Υπολόγισε το πλήθος των απογόνων όλων των κόμβων και την οριακή απώλεια που προκαλεί η αφαίρεση κάθε ακμής σε κάθε δένδρο διάδοσης της αληθούς πληροφορίας (Γραμμές 19-33). Εδώ ως οριακή απώλεια ορίζεται αυτή που αναφέρεται μόνο στη διάδοση της αληθούς πληροφορίας:

$$\Delta_T(e|E_t) = \sum_{a \in S_T} \sigma(a, (V, E \setminus E_t)) - \sum_{a \in S_T} \sigma(a, (V, E \setminus (E_t \cup \{e\}))) \quad (39)$$

3. Υπολόγισε το πλήθος των απογόνων όλων των κόμβων και την οριακή απώλεια που προκαλεί η αφαίρεση κάθε ακμής σε κάθε δένδρο διάδοσης της ψευδούς πληροφορίας (Γραμμη 35). Εδώ ως οριακή απώλεια ορίζεται αυτή που αναφέρεται μόνο στη διάδοση της ψευδούς πληροφορίας:

$$\Delta_F(e|E_t) = \sum_{b \in S_F} \sigma(b, (V, E \setminus E_t)) - \sum_{b \in S_F} \sigma(b, (V, E \setminus (E_t \cup \{e\}))) \quad (40)$$

4. Όσο δεν έχει εξαντληθεί το πλήθος των διαθέσιμων αφαιρέσιμων ακμών επανάλαβε τα ακόλουθα:

   (α′) Βρες την ακμή που προκαλεί τη μέγιστη μείωση $\Delta(e|E_t)$ της αντικειμενι-

κής συνάρτησης (Γραμμή 39):

$$
\begin{aligned}
&\Delta(e|E_t) \\
&= f_{LT}(E_t) - f_{LT}(E_t \cup \{e\}) \\
&= \sum_{a \in S_T} \sigma(a, (V, E)) - \sum_{a \in S_T} \sigma(a, (V, E \setminus E_t)) + \sum_{b \in S_F} \sigma(b, (V, E \setminus E_t)) \\
&- \sum_{a \in S_T} \sigma(a, (V, E)) + \sum_{a \in S_T} \sigma(a, (V, E \setminus (E_t \cup \{e\}))) \\
&- \sum_{b \in S_F} \sigma(b, (V, E \setminus (E_t \cup \{e\}))) \\
&= \sum_{b \in S_F} \sigma(b, (V, E \setminus E_t)) - \sum_{b \in S_F} \sigma(b, (V, E \setminus (E_t \cup \{e\}))) \\
&- [\sum_{a \in S_T} \sigma(a, (V, E \setminus E_t)) - \sum_{a \in S_T} \sigma(a, (V, E \setminus (E_t \cup \{e\})))] \\
&= \Delta_F(e|E_t) - \Delta_T(e|E_t)
\end{aligned}
\tag{41}
$$

(β′) Αν η μέγιστη μείωση δεν είναι θετικού προσήμου, δηλαδή πρόκειται για αύξηση της αντικειμενικής συνάρτησης, σταμάτα (Γραμμές 40-41). Αλλιώς, πρόσθεσε την ακμή στο σύνολο των αφαιρεμένων ακμών και συνέχισε στο επόμενο βήμα (Γραμμές 43-44).

(γ′) Επαναϋπολόγισε το πλήθος των απογόνων όλων των κόμβων (που συμμετέχουν στη διάδοση της πληροφορίας) και την οριακή απώλεια που προκαλεί η αφαίρεση κάθε ακμής σε κάθε δένδρο διάδοσης της αληθούς πληροφορίας (Γραμμές 45-63).

(δ′) Επαναϋπολόγισε το πλήθος των απογόνων όλων των κόμβων (που συμμετέχουν στη διάδοση της πληροφορίας) και την οριακή απώλεια που προκαλεί η αφαίρεση κάθε ακμής σε κάθε δένδρο διάδοσης της ψευδούς πληροφορίας (Γραμμές 65-67).

Η χρονική πολυπλοκότητα του επαναληπτικού μέρους αλγορίθμου πλέον είναι $O(k(|E|+ x_S \cdot (|S_T|+|S_F|) \cdot |V|))$, όπου ο παράγοντας $|E|$ οφείλεται στο βήμα (α') (εύρεση μέγιστης τιμής) και ο παράγοντας $x_S \cdot (|S_T| + |S_F|) \cdot |V|$ προκαλείται από τα βήματα (γ') και (δ') (διάσχιση των δένδρων επιρροής). Φυσικά, στις περισσότερες περιπτώσεις τα σύνολα S_T και S_F είναι μικρής πληθικότητας σε σχέση με το σύνολο όλων των κόμβων, οπότε τότε μπορεί να θεωρηθεί ότι ο αλγόριθμος είναι γραμμικός κατά ένα παράγοντα ως προς το μέγεθος του δικτύου.

6.4.2 Αλγόριθμος υπό το μοντέλο DLT

Σε αντίθεση με το πιθανοτικό μοντέλο LT, στο μοντέλο DLT ο υπολογισμός της επιρροής ενός συνόλου A, $\sigma(A, G)$ σε ένα γράφο $G = (V, E)$ γίνεται με ακρίβεια

σε γραμμικό χρόνο [117], με αλγόριθμο ο οποίος ελέγχει σε κάθε διακριτό χρονικό βήμα t αν οι για πρώτη φορά ενεργοποιημένοι κόμβοι επιφέρουν την ενεργοποίηση ανενεργών κόμβων μέχρις ότου να μην ενεργοποιηθεί κανένας νέος κόμβος σε κάποιο βήμα, και παρουσιάζεται με τη μορφή ψευδοκώδικα παρακάτω:

```
Input: graph G(V,E,w), threshold θ(u),∀u∈V, set of seeds of
    information S
Variables: set of activated vertices activated, set of
    activated vertices in the current iteration
    newly_activated, sum of influences in a vertex u
    sum_influence[u]
Output: set of activated vertices activated

// initialization
for each vertex u∈V:
    sum_influence[u] = 0

activated = S
newly_activated = S
// propagation
while newly_activated is not empty:
    temp_activated = ∅
    for each vertex u∈newly_activated:
        for each v∈N_out(u):
            sum_influence[v] += w(u,v)
            if v∉activated and sum_influence[v]≥theta[v]:
                temp_activated.add(v)
                activated.add(v)
    newly_activated = temp_activated
```

Όπως και προηγουμένως, ο αλγόριθμος που θα εφαρμοστεί για το πρόβλημα CMM είναι άπληστος επαναληπτικός, σε κάθε επανάληψη του οποίου επιλέγεται προς αφαίρεση η ακμή που επιφέρει τη μέγιστη μείωση ή αλλιώς οριακή απώλεια της αντικειμενικής συνάρτησης. Ωστόσο, σε αντίθεση με την περίπτωση του μοντέλου LT, επιλέγεται η ακμή με βάση την ελάχιστη τιμή της αντικειμενικής συνάρτησης μετά την αφαίρεσή της κι όχι με βάση τη διαφορά των τιμών της αντικειμενικής συνάρτησης πριν και μετά την αφαίρεσή της. Δηλαδή πλέον το κριτήριο εκφράζεται μαθηματικά ως:

$$e = (u, v) = \arg\min_{e \in E \setminus E_t} \{f_{DLT}(E_t \cup \{e\})\} \tag{42}$$

Έτσι, ο αλγόριθμος διαμορφώνεται όπως φαίνεται παρακάτω:

```
Input: original graph G(V,E,w), expertise e(u),∀u∈V, set of
    seeds of true information S_T, set of seeds of false
    information S_F, maximum number of removed edges k
Variables: threshold for accepting true information θ_T(u),∀u∈V
    , threshold for accepting false information θ_F(u),∀u∈V,
    subgraph of propagation of true information G_true, subgraph
     of propagation of false information G_false
```

```
3  Output: set of removed edges E'
4
5  // initialization
6  for each vertex u ∈ V:
7       θ_T(u) = min(1 − e(u), e(u))
8       θ_F(u) = e(u)
9
10 E' = ∅
11 Compute activated_true, influence_true: activated set and influence of
       true information's seed set S_T in graph G(V, E, w) with
       thresholds θ_T(u), ∀u ∈ V.
12 Compute activated_false, influence_false: activated set and influence
       of false information's seed set S_F in graph G(V, E, w) with
       thresholds θ_F(u), ∀u ∈ V.
13
14 // initial true information's propagation
15 t_init = influence_true
16
17 Build induced subgraph of G, G_true = (V_T, E_T, w) with
       V_T = activated_true.
18 Build induced subgraph of G, G_false = (V_F, E_F, w) with
       V_F = activated_false.
19
20 f_curr = influence_false
21 while k is not 0:
22      Find edge e = arg min_{e∈E_F} {t_init − influence_true(e) + influence_false(e)}
23      if t_init − influence_true(e) + influence_false(e) ≥ f_curr:
24           Stop. // f_DLT cannot be decreased more
25      else:
26           E' = E' ∪ {e}
27           k− = 1
28           f_curr = t_init − influence_true(e) + influence_false(e)
29           Remove edge e from G_true, G_false.
30           Build induced subgraph of G_true, G_true = (V_T, E_T, w) with
       V_T = activated_true(e).
31           Build induced subgraph of G_false, G_false = (V_F, E_F, w) with
       V_F = activated_false(e).
```

Ουσιαστικά τα βήματα που ακολουθούνται σύμφωνα με τον παραπάνω ψευδοκώδικα αποτυπώνονται ως εξής:

1. Υπολόγισε το κατώφλι αποδοχής της αληθούς και της ψευδούς πληροφορίας για κάθε κόμβο (Γραμμές 6-8).

2. Υπολόγισε το σύνολο των κόμβων που αποδέχονται την αληθή πληροφορία και το σύνολο των κόμβων που αποδέχονται την ψευδή πληροφορία (Γραμμές 11-12).

3. Διατήρησε την αρχική τιμή της διάδοσης της αληθούς πληροφορίας (Γραμμή 15).

4. Κατασκεύασε τον επαγόμενο γράφο διάδοσης της αληθούς πληροφορίας και τον επαγόμενο γράφο διάδοσης της ψευδούς πληροφορίας με βάση τους αντίστοιχους κόμβους του βήματος 2 (Γραμμές 17-18). Με αυτόν τον τρόπο περιορίζεται το μέγεθος του γράφου και ως εκ τούτου, το πλήθος των ακμών που εξετάζονται.

5. Όσο δεν έχει εξαντληθεί το πλήθος των διαθέσιμων αφαιρέσιμων ακμών επανάλαβε τα ακόλουθα:

   (α′) Βρες την ακμή που οδηγεί στην ελάχιστη τιμή της αντικειμενικής συνάρτησης (Γραμμή 22). Σημειώνεται ότι οι μεταβλητές $influence\_true(e)$ και $influence\_false(e)$ είναι το πλήθος των κόμβων που αποδέχονται τη διαδιδόμενη πληροφορία στον επαγόμενο γράφο διάδοσης της αληθούς και στον επαγόμενο γράφο διάδοσης της ψευδούς πληροφορίας αντίστοιχα, από τους οποίους έχει αφαιρεθεί προσωρινά η ακμή e. Επιπλέον, εξετάζονται μόνο οι ακμές που βρίσκονται στο γράφο διάδοσης της ψευδούς πληροφορίας, διότι αν δεν υπάρχουν σε αυτόν, τότε είτε δεν υπάρχουν ούτε στο γράφο διάδοσης της αληθούς πληροφορίας, οπότε η αφαίρεσή τους δεν επηρεάζει καθόλου τη διάδοση των πληροφοριών, είτε υπάρχουν μόνο στο γράφο διάδοσης της αληθούς πληροφορίας, οπότε η αφαίρεσή τους στην καλύτερη περίπτωση δεν επηρεάζει τη διάδοση της αληθούς πληροφορίας, ενώ στη χειρότερη την περιορίζει.

   (β′) Αν η ελάχιστη τιμή δεν είναι μικρότερη από την τρέχουσα τιμή της αντικειμενικής συνάρτησης, σταμάτα (Γραμμές 23-24). Αλλιώς, πρόσθεσε την ακμή στο σύνολο των αφαιρεμένων ακμών και συνέχισε στο επόμενο βήμα (Γραμμές 26-27).

   (γ′) Θέσε την τρέχουσα τιμή της αντικειμενικής συνάρτησης ίση με την ελάχιστη που βρέθηκε (Γραμμή 28).

   (δ′) Αφαίρεσε από τους επαγόμενους γράφους διάδοσης την ακμή (Γραμμή 29).

   (ε′) Κατασκεύασε τον επαγόμενο γράφο διάδοσης της αληθούς πληροφορίας και τον επαγόμενο γράφο διάδοσης της ψευδούς πληροφορίας με βάση τους αντίστοιχους κόμβους του βήματος (α') (Γραμμές 30-31).

Όσον αφορά την απόδοση του αλγορίθμου, ο υπολογιστικός φόρτος προκαλείται κυρίως κατά τον υπολογισμό της διάδοσης μιας πληροφορίας με χρονική πολυπλοκότητα $O(|E|)$ [117]. Έτσι, ο παραπάνω αλγόριθμος έχει συνολική πολυπλοκότητα $O(k{\cdot}|E|^2)$, λόγω του υπολογισμού της διάδοσης των πληροφοριών μετά την αφαίρεση κάθε ακμής σε κάθε επανάληψη. Ωστόσο, συνήθως η πληροφορία δε διαδίδεται σε ολόκληρο το δίκτυο (λόγω των πιθανοτήτων επιρροής των ακμών, και των κατωφλίων), με αποτέλεσμα το πλήθος των προς εξέταση ακμών στους επαγόμενους γράφους να είναι αρκετά μικρότερο από αυτό του αρχικού γράφου.

7 Πειράματα

Στο συγκεκριμένο κεφάλαιο θα παρουσιαστούν και θα σχολιαστούν τα αποτελέσματα της εφαρμογής του προτεινόμενου αλγορίθμου σε πραγματικά κοινωνικά δίκτυα. Προς τούτο χρησιμοποιήθηκε η βιβλιοθήκη NetworkX [123], ένα πολύ εύχρηστο πακέτο της γλώσσας προγραμματισμού Python για τη δημιουργία, επεξεργασία και μελέτη της δομής, της δυναμικής και των λειτουργιών ενός δικτύου. Περαιτέρω, ο κώδικας γράφτηκε στην έκδοση Python 3.9.5 εντός του ολοκληρωμένου περιβάλλοντος ανάπτυξης (IDE) PyCharm [124] και παρέχεται στο GitHub repository [125]. Τα πειράματα εκτελέστηκαν σε Προσωπικό Υπολογιστή με επεξεργαστή AMD Ryzen 7 3700U with Radeon Vega Mobile Gfx, 2.30 GHz, μνήμη RAM 12.0 GB (9.95 GB usable) και λειτουργικό σύστημα Windows 11 Home. Παρακάτω θα παρατεθούν τα μελετούμενα πραγματικά δίκτυα και στη συνέχεια θα γίνει σύγκριση της απόδοσης της προτεινόμενης μεθόδου για την αντιμετώπιση του προβλήματος CMM με άλλες συγκριτικές μεθόδους υπό τα μοντέλα διάδοσης LT και DLT.

7.1 Μελετούμενα Δίκτυα

Καθώς το υπό μελέτη πρόβλημα συναντάται στα κοινωνικά δίκτυα, κρίνεται σκόπιμη η εξέταση της αποτελεσματικότητας του αλγορίθμου σε πραγματικά κοινωνικά δίκτυα, τα οποία βέβαια μπορούν να χαρακτηριστούν ως Δίκτυα Ελεύθερης Κλίμακας στο χώρο των σύνθετων δικτύων.

7.1.1 Πραγματικά Δίκτυα

Από τη συλλογή συνόλων δεδομένων μεγάλων δικτύων του Πανεπιστημίου Stanford [126] επιλέχθηκαν οι τοπολογίες 3 κοινωνικών δικτύων, τα οποία παρουσιάζονται κάτωθι:

- Δίκτυο email-Eu-core [127]
  Το εν λόγω δίκτυο προέρχεται από την καταγραφή της κίνησης της ηλεκτρονικής αλληλογραφίας μεταξύ των μελών ενός μεγάλου Ευρωπαικού Ερευνητικού Ινστιτούτου κατά τη χρονική περίοδο Οκτώβριος 2003 - Μάϊος 2005. Έτσι, κάθε κόμβος αντιπροσωπεύει ένας ακαδημαϊκό μέλος και κάθε κατευθυνόμενη ακμή μεταξύ δύο κόμβων αντιπροσωπεύει την αποστολή τουλάχιστον ενός μηνύματος από το ένα μέλος στο άλλο.

- Δίκτυο Social circles: Facebook [128]
  Αυτό το δίκτυο αποτελείται από τους κύκλους, δηλαδή τις λίστες φίλων, των χρηστών της κοινωνικής πλατφόρμας Facebook, οι οποίοι συλλέχθησαν μέσω της εθελοντικής χρήσης μιας ορισμένης εφαρμογής της πλατφόρμας. Κατά συνέπεια, κάθε κόμβος αντιστοιχεί σε ένα χρήστη και κάθε μη κατευθυνόμενη ακμή μεταξύ δύο κόμβων αντιστοιχεί στην ύπαρξη αμφίδρομης φιλίας μεταξύ των χρηστών.

- Δίκτυο Wikipedia vote [129, 130]
  Το συγκεκριμένο δίκτυο προκύπτει από την καταγραφή των ψηφοφοριών μέχρι τον Ιανουάριο 2008 για την προαγωγή χρηστών σε διαχειριστές της πλατφόρμας Wikipedia [131], η οποία αποτελεί μια δωρεάν διαδικτυακή εγκυκλοπαίδεια συντασσόμενη συνεργατικά από εθελοντές. Συνεπώς, κάθε κόμβος αντιπροσωπεύει ένα χρήστη (διαχειριστή ή συγγραφέα) και κάθε κατευθυνόμενη ακμή μεταξύ δύο κόμβων αντιπροσωπεύει τη θετική ψήφο του ενός χρήστη υπέρ του άλλου.

Στον Πίνακα 4 παρουσιάζονται τα κυριότερα τοπολογικά χαρακτηριστικά των προαναφερθέντων δικτύων. Αξίζει να αναφερθεί ότι στην περίπτωση του δικτύου email-Eu-core αφαιρέθηκαν οι ανακυκλώσεις (ακμές με τον ίδιο κόμβο ως αρχή και τέλος) και οι απομονωμένοι κόμβοι (κόμβοι χωρίς καμία προσπίπτουσα ακμή), καθώς δε φέρουν σημασία στην παρούσα μελέτη. Για τον ίδιο λόγο, στην περίπτωση του δικτύου Wikipedia Vote αφαιρέθηκαν οι πολλαπλές ακμές (ακμές με τον ίδιο κόμβο ως αρχή και τον ίδιο κόμβο ως τέλος).

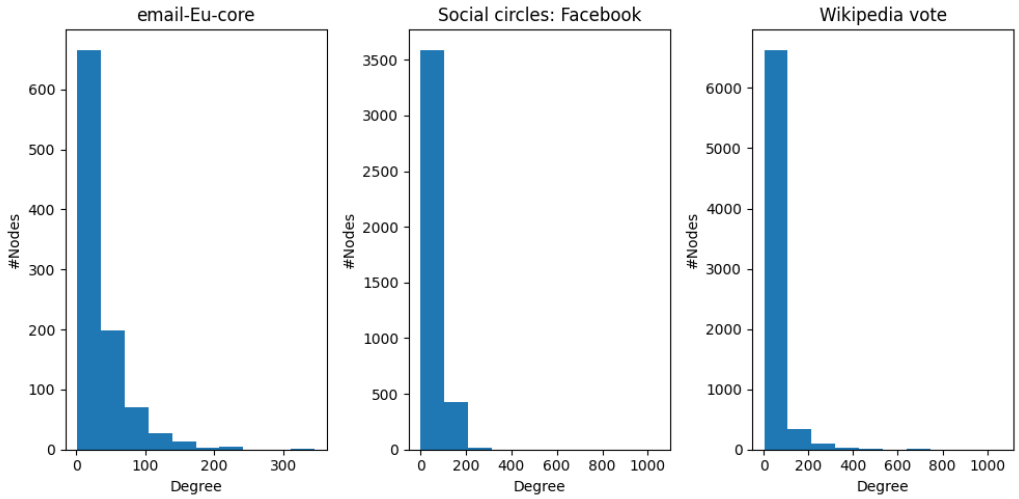
Πίνακας 4: Τοπολογικά χαρακτηριστικά των πραγματικών κοινωνικών δικτύων email-Eu-core, Social circles: Facebook και Wikipedia vote

Δίκτυο	**Γράφος**	**Κόμβοι**	**Ακμές**	**Μέσο μήκος μονοπατιού**	**Μέσος συντελεστής ομαδοποίησης**
email-Eu-core	Κατευθυνόμενος	986	16064	2.5869	0.4071
Social circles: Facebook	Μη κατευθυνόμενος	4039	88234	3.6925	0.6055
Wikipedia vote	Κατευθυνόμενος	7115	100762	3.247	0.1409

Στο Σχήμα 7 απεικονίζεται η κατανομή βαθμού κόμβου των 3 δικτύων. Είναι προφανές ότι ακολουθούν την κατανομή νόμου-δύναμης (power-law distribution), κατά την οποία λίγοι κόμβοι-επίκεντρα συγκεντρώνουν υψηλό αριθμό γειτόνων, ενώ οι περισσότεροι κόμβοι διαθέτουν μικρό βαθμό. Επιπλέον, όπως φαίνεται στον Πίνακα 4, και στα 3 δίκτυα το μέσο μήκος μονοπατιού είναι μικρό, ενώ ταυτόχρονα αν εξαιρέσει κανείς το δίκτυο Wikipedia vote, ο μέσος συντελεστής ομαδοποίησης είναι σχετικά υψηλός. Αυτά τα χαρακτηριστικά επιτρέπουν την κατηγοριοποίηση των εν λόγω κοινωνικών δικτύων στα αντίστοιχα σύνθετα Δίκτυα Ελεύθερης Κλίμακας.

7.1.2 Απόδοση πιθανοτήτων

Προκειμένου να είναι δυνατή η προσομοίωση των μοντέλων διάδοσης LT και DLT είναι αναγκαία η απόδοση βάρους σε κάθε ακμή $(u, v) \in E$ του δικτύου $G = (V, E)$, το οποίο αντιστοιχεί στην ισχύ της επιρροής που ασκεί ο κόμβος u στον κόμβο v, και πρέπει να συμμορφώνεται στην εξίσωση 10. Προς τούτο, ακολουθείται για κάθε κόμβο $v \in V$ η παρακάτω διαδικασία:



Σχήμα 7: Κατανομή βαθμού κόμβου για τα δίκτυα email-Eu-core, Social circles: Facebook και Wikipedia vote

1. Για κάθε εισερχόμενη ακμή του κόμβου v, $(u, v) \in E$, επιλέγεται τυχαία σύμφωνα με την ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ η πιθανότητα $w_{unconstrained}(u, v)$.

2. Επιλέγεται τυχαία σύμφωνα με την ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ η πιθανότητα αδυναμίας ενεργοποίησης του κόμβου, ακόμα κι αν όλοι οι εισερχόμενοι γείτονές του μολυνθούν, η οποία συμβολίζεται ως $w_{no\_activation}(v)$.

3. Προκειμένου να ικανοποιείται η εξίσωση 10, γίνεται κανονικοποίηση των παραπάνω πιθανοτήτων των εισερχόμενων ακμών του κόμβου v ως εξής:

$$w(u, v) = \frac{w_{unconstrained}(u, v)}{\sum_{u:(u,v)\in E} w_{unconstrained}(u, v) + w_{no\_activation}(v)} \tag{43}$$

Επιπλέον, στην περίπτωση του μοντέλου DLT, για κάθε κόμβο $u \in V$ επιλέγεται η εξειδίκευση $e(u)$ τυχαία σύμφωνα με την ομοιόμορφη κατανομή στο διάστημα $[0, 1]$.

7.2 Συγκριτικές μέθοδοι

Με στόχο την αξιολόγηση της αποδοτικότητας του προτεινόμενου αλγορίθμου, εξετάζονται οι παρακάτω ευριστικές μέθοδοι που αξιοποιούν κυρίως τη δομή του δικτύου κι όχι τη δυναμική συμπεριφορά των μοντέλων διάδοσης [6]:

- Random: Αφαίρεση k τυχαίων ακμών. Αυτό επιτυγχάνεται ακολουθώντας τον αλγόριθμο Fisher-Yates [132, 133] για την παραγωγή μιας τυχαίας μετάθεσης του συνόλου των ακμών του δικτύου E_{lim}, οι οποίες έχουν ως αρχή ένα κόμβο

που ανήκει στο σύνολο των αρχικών ενστερνιστών της ψευδούς πληροφορίας, δηλαδή $E_{lim} = \{(u, v) : (u, v) \in E, u \in S_F\}$. Αυτό αιτιολογείται από το γεγονός ότι είναι πιο αποτελεσματική η τυχαία αφαίρεση ακμών κοντά στο σημείο εκκίνησης της διάδοσης της ψευδούς πληροφορίας, προτού διαδοθεί σε μεγαλύτερο μέρος του δικτύου.

- Weighted: Αφαίρεση των k ακμών με την υψηλότερη πιθανότητα διάδοσης της πληροφορίας, $w(u, v)$. Και σε αυτή την περίπτωση λαμβάνονται υπόψη μόνο οι ακμές οι οποίες έχουν ως αρχή ένα κόμβο που ανήκει στο σύνολο των αρχικών ενστερνιστών της ψευδούς πληροφορίας. Με αυτό τον τρόπο, λαμβάνεται μερικώς υπόψη ο τρόπος διάδοσης της πληροφορίας, αφού στόχο αποτελούν οι ακμές που συνδράμουν κατά το μεγαλύτερο ποσοστό στα αρχικά στάδια της διάδοσης. Βέβαια σύμφωνα με αυτόν τον ευριστικό αλγόριθμο, αλλά και τον προηγούμενο, αγνοείται η εξάπλωση της αληθούς πληροφορίας. Για αυτό, αντιμετωπίζονται κυρίως ως μέθοδοι αναφοράς για τις άλλες μεθόδους.

- DistanceDiff (μόνο υπό το μοντέλο LT): Αφαίρεση των k ακμών με τη μικρότερη διαφορά της απόστασης από το σύνολο των αρχικών ενστερνιστών μόνο της αληθούς πληροφορίας από την απόσταση από το σύνολο των αρχικών ενστερνιστών μόνο της ψευδούς πληροφορίας. Αυτή η διαφορά εκφράζεται μαθηματικά για κάθε ακμή $(u, v) \in E$ ως:

$$diff(u, v) = \min_{sf \in S_F \setminus S_T} \{d(sf, u)\} - \min_{st \in S_T \setminus S_F} \{d(st, u)\} \tag{44}$$

  Πρέπει να σημειωθεί ότι προκειμένου να συμπεριληφθεί έστω μερικώς ο τρόπος διάδοσης της πληροφορίας, στον υπολογισμό των συντομότερων μονοπατιών λαμβάνεται υπόψη η πιθανότητα διάδοσης σε κάθε ακμή $(u, v) \in E$ θεωρώντας ως βάρος $w'(u, v) = \frac{1}{w(u,v)}$, έτσι ώστε να δίνεται προτεραιότητα στα μονοπάτια που αποτελούνται από ακμές με μεγαλύτερη ισχύ στη διάχυση της πληροφορίας στο δίκτυο. Επιπλέον, σε περίπτωση ισοπαλίας με βάση την παραπάνω μετρική, επιλέγεται η ακμή με τη μεγαλύτερη πιθανότητα διάδοσης. Με αυτόν τον ευριστικό αλγόριθμο, αξιοποιείται κυρίως η δομή του δικτύου και επιδιώκεται ο περιορισμός της διάδοσης πληροφοριών όσο πιο κοντά στο σημείο εκκίνησης της διάδοσης της ψευδούς είδησης κι όσο πιο μακριά από το αντίστοιχο σημείο της αληθούς είδησης.

- EdgeBetweennessDiff (μόνο υπό το μοντέλο DLT): Αφαίρεση των k ακμών με τη μεγαλύτερη διαφορά της κεντρικότητας ενδιαμεσικότητας στον επαγόμενο γράφο διάδοσης της αληθούς πληροφορίας από την κεντρικότητα ενδιαμεσικότητας στον επαγόμενο γράφο διάδοσης της ψευδούς πληροφορίας. Αυτή η διαφορά εκφράζεται μαθηματικά για κάθε ακμή $e \in E$ ως:

$$diff(e) = C_{B,false}(e) - C_{B,true}(e) \tag{45}$$

  Είναι προφανές από τον ορισμό της μετρικής ότι λαμβάνονται υπόψη μόνο οι ακμές που ανήκουν σε τουλάχιστον έναν εκ τους δύο επαγόμενους γράφους,

αφού οι υπόλοιπες δε συμμετέχουν στη διάδοση των πληροφοριών και έτσι η αφαίρεσή τους δεν έχει κάποιο αντίκτυπο. Όπως και προηγουμένως, στον υπολογισμό των συντομότερων μονοπατιών λαμβάνεται υπόψη η πιθανότητα διάδοσης σε κάθε ακμή $(u, v) \in E$ θεωρώντας ως βάρος $w'(u, v) = \frac{1}{w(u,v)}$, έτσι ώστε μια ακμή με μεγάλη πιθανότητα να συμμετέχει σε περισσότερα συντομότερα μονοπάτια και ως τούτου να εμφανίζει μεγαλύτερη κεντρικότητα ενδιαμεσικότητας. Υπενθυμίζεται ότι η κεντρικότητα ενδιαμεσικότητας αποτελεί ένα καλό κριτήριο ισχύος μιας ακμής όταν πρόκειται για τη μελέτη διάδοσης πληροφοριών. Έτσι, στόχος αποτελεί η αφαίρεση ακμών οι οποίες είναι σημαντικές στη διάχυση της ψευδούς πληροφορίας αλλά ασήμαντες στη διάχυση της αληθούς πληροφορίας. Επιπλέον, προκειμένου να επιτευχθεί εξοικονόμηση χρόνου και να ληφθεί υπόψη η γνώση των αρχικών ενστερνιστών των δύο κλάσεων πληροφορίας, στον υπολογισμό της κεντρικότητας ενδιαμεσικότητας λαμβάνονται υπόψη μόνο τα μονοπάτια με αρχή έναν κόμβο που ανήκει στο σύνολο των αρχικών ενστερνιστών της είδησης στην οποία ανφέρεται ο επαγόμενος γράφος και τέλος οποιοδήποτε κόμβο εκτός αυτών.

Αξίζει να σημειωθεί ότι σε κάθε περίπτωση οι ακμές αφαιρούνται κατά ομάδες των 10 ή 25 ή 50 ανάλογα με την πληθικότητα του συνόλου των ακμών που λαμβάνονται υπόψη με στόχο την ισορροπία μεταξύ της εξοικονόμησης χρόνου εκτέλεσης και του κατάλληλου βαθμού λεπτομέρειας στην εξέλιξη της τιμής της αντικειμενικής συνάρτησης. Επίσης, σε περίπτωση που επιτευχθεί η εξάλειψη της διάδοσης της ψευδούς πληροφορίας με την αφαίρεση ακμών πλήθους μικρότερου του k, ο αλγόριθμος σταματάει διότι η περαιτέρω αφαίρεση ακμών θα επιφέρει μόνο τη μείωση της διάδοσης της αληθούς πληροφορίας (και την αύξηση της αντικειμενικής συνάρτησης).

7.3 Παρουσίαση και σχολιασμός της απόκρισης των μεθόδων

Για την εκτέλεση των πειραμάτων ορίζονται τα κάτωθι σε κάθε περίπτωση δικτύου $G = (V, E)$ υπό το μοντέλο διάδοσης LT ή DLT, ώστε να λειτουργήσουν ως είσοδος στην εφαρμογή του προτεινόμενου αλγορίθμου και των παραπάνω συγκριτικών μεθόδων:

- S_T, S_F: Επιλέγονται τα σύνολα των αρχικών ενστερνιστών των κλάσεων I_T, I_F αντίστοιχα, τυχαία σύμφωνα με την ομοιόμορφη κατανομή με πληθικότητα $|S_T| = |S_F| = \lceil 1\% \cdot |V| \rceil$ [85].

- k: Ορίζεται ο μέγιστος αριθμός αφαιρέσιμων ακμών ως:

$$k = \lceil 3\% \cdot |E| \rceil \tag{46}$$

  Θεωρούμε αυτό τον περιορισμό με στόχο μεν την αποτελεσματική επίδραση του προτεινόμενου αλγορίθμου δε την όσο το δυνατό μικρότερη διατάραξη του δικτύου και της εμπερίας των χρηστών εντός αυτού [55, 56].

Στην περίπτωση του μοντέλου διάδοσης LT, ορίζεται επιπλέον το πλήθος των δειγματοληπτημένων γράφων, $x_S = 5000$, όπως προτείνεται στο [6]. Στη συνέχεια, θα παρουσιαστούν και θα σχολιαστούν τα αποτελέσματα των πειραμάτων υπό τα μοντέλα διάδοσης LT και DLT.

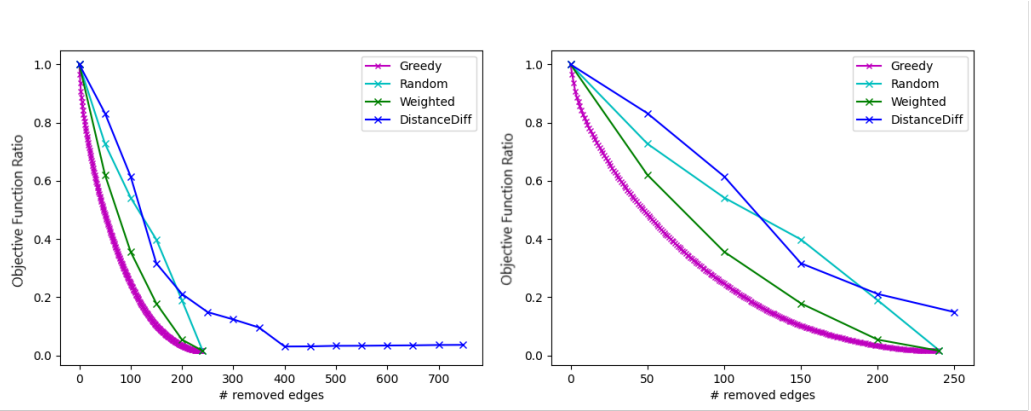
7.3.1 Αποτελέσματα υπό το μοντέλο διάδοσης LT

Στα Σχήματα 8, 9 και 10 φαίνεται η απόδοση των διαφόρων μεθόδων υπό το μοντέλο LT για τα δίκτυα email-Eu-core, Social circles: Facebook και Wikipedia vote αντίστοιχα. Αυτή η απόδοση ποσοτικοποιείται χρησιμοποιώντας το πηλίκο της τιμής της αντικειμενικής συνάρτησης μετά την αφαίρεση k ακμών προς την αρχική τιμή αυτής:

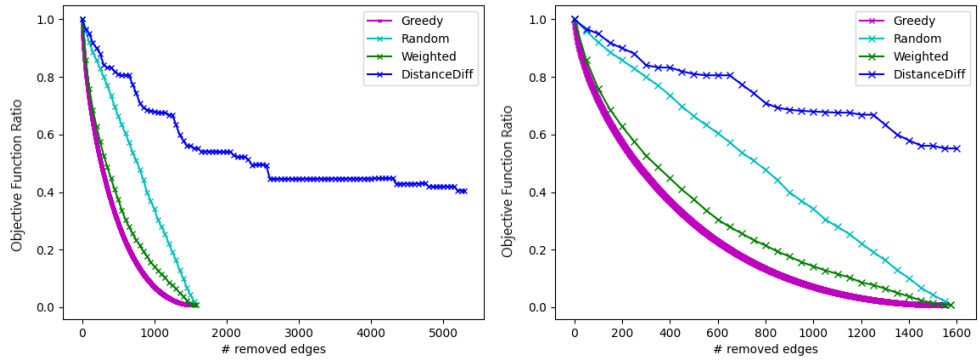
$$ratio(k) = \frac{f_{LT}(E')}{f_{LT}(\emptyset)}, \text{ με } |E'| = k \tag{47}$$

Σημειώνεται ότι ο προτεινόμενος αλγόριθμος αναφέρεται στις γραφικές παραστάσεις ως "Greedy". Επιπλέον, λόγω της συνθήκης να σταματούν όλοι οι αλγόριθμοι όταν η διάδοση της ψευδούς πληροφορίας έχει εξαλειφθεί, παρατηρείται το φαινόμενο όλες οι μέθοδοι εκτός της "DistanceDiff" να σταματούν με την αφαίρεση μικρότερου πλήθους ακμών από τον ακέραιο k. Για αυτό το λόγο, κάθε σχήμα περιλαμβάνει δύο γραφικές παραστάσεις: Η πρώτη απεικονίζει τα πλήρη αποτελέσματα, ενώ η δεύτερη εστιάζει στο διάστημα του οριζόντιου άξονα μέχρι την τιμή του πλήθους των ακμών που αφαιρέθηκαν κατά την εφαρμογή των μεθόδων εκτός της "DistanceDiff", με στόχο τη λεπτομερέστερη παρατήρηση της επίδρασής τους στην αντιμετώπιση του μελετούμενου προβλήματος.
Είναι εμφανής η υπεροχή του προτεινόμενου άπληστου αλγορίθμου, αφού επιτυγχάνει

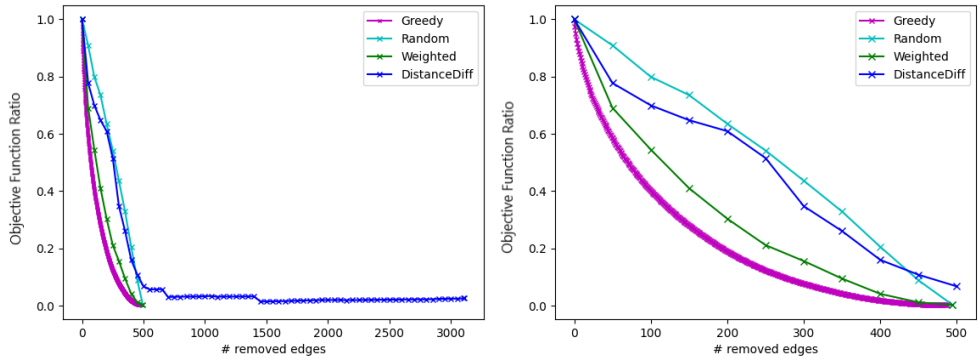
τη ραγδαία μείωση της αντικειμενικής συνάρτησης με την αφαίρεση ακόμα και μικρού πλήθους ακμών. Ακολουθεί με σχετικά καλή απόδοση και παρόμοια συμπεριφορά, αλλά με μικρότερο ρυθμό μείωσης της τιμής της συνάρτησης η μέθοδος "Weighted", όπως και αναμενόταν χάρη στην επιλογή ακμών με μεγαλύτερο βάρος, άρα και μεγαλύτερη επίδραση στη διάδοση της πληροφορίας. Εν γένει, παρατηρεί κανείς ότι η μέθοδος "Random" επιφέρει σχεδόν γραμμική μείωση στην τιμή της συνάρτησης, αφού γίνεται με τυχαίο τρόπο η επιλογή των ακμών. Επιπλέον, στα δίκτυα email-Eu-core και Wikipedia vote είναι συγκρίσιμη η απόδοσή της με αυτή της μεθόδου "DistanceDiff". Ωστόσο, η τελευταία φαίνεται να μην κατορθώνει την εξάλειψη της παραπληροφόρησης με τόσες ακμές όσο οι υπόλοιπες μέθοδοι, αλλά να αυξομειώνει με πολύ μικρό ρυθμό την τιμή της αντικειμενικής συνάρτησης εξαντλώντας το διαθέσιμο πλήθος αφαιρέσιμων ακμών. Ειδικά στην περίπτωση του δικτύου Social circles: Facebook, η μέθοδος κρίνεται εξαιρετικά αναποτελεσματική. Έτσι, μπορεί κανείς να εξάγει το εύλογο συμπέρασμα ότι ένας τέτοιος αλγόριθμος που αξιοποιεί κατά κύριο λόγο τα τοπολογικά χαρακτηριστικά του δικτύου δεν είναι ο καταλληλότερος για την επίλυση του προβλήματος CMM.



Σχήμα 8: Γραφική παράσταση της μετρικής $ratio(k)$ ως προς το πλήθος των αφαιρεμένων ακμών k με εφαρμογή των μεθόδων Greedy, Random, Weighted και DistanceDiff υπό το μοντέλο LT για το δίκτυο email-Eu-core.



Σχήμα 9: Γραφική παράσταση της μετρικής $ratio(k)$ ως προς το πλήθος των αφαιρεμένων ακμών k με εφαρμογή των μεθόδων Greedy, Random, Weighted και DistanceDiff υπό το μοντέλο LT για το δίκτυο Social circles: Facebook.



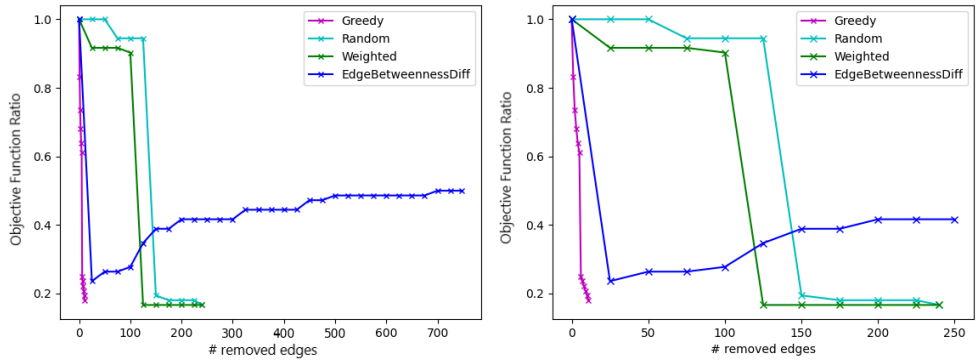
Σχήμα 10: Γραφική παράσταση της μετρικής $ratio(k)$ ως προς το πλήθος των αφαιρεμένων ακμών k με εφαρμογή των μεθόδων Greedy, Random, Weighted και DistanceDiff υπό το μοντέλο LT για το δίκτυο Wikipedia vote.

7.3.2 Αποτελέσματα υπό το μοντέλο διάδοσης DLT

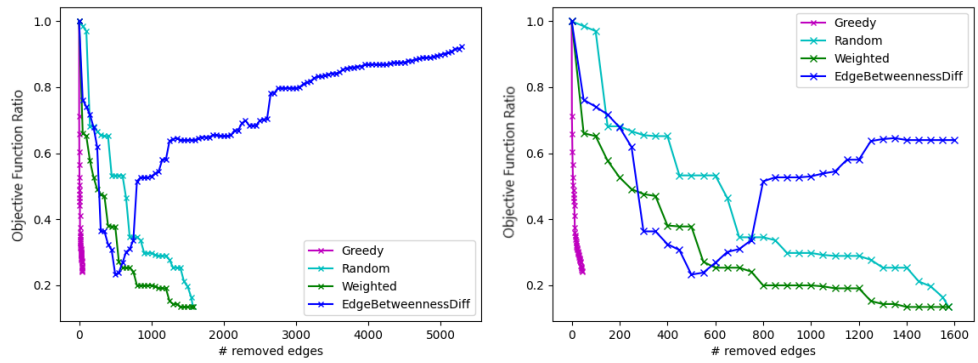
Στα Σχήματα 11, 12 και 13 φαίνεται η απόδοση των διαφόρων μεθόδων υπό το μοντέλο DLT για τα δίκτυα email-Eu-core, Social circles: Facebook και Wikipedia vote αντίστοιχα. Αυτή η απόδοση ποσοτικοποιείται, όπως και προηγουμένως, χρησιμοποιώντας το πηλίκο της τιμής της αντικειμενικής συνάρτησης μετά την αφαίρεση k ακμών προς την αρχική τιμή αυτής:

$$ratio(k) = \frac{f_{DLT}(E')}{f_{DLT}(\emptyset)}, \text{ με } |E'| = k \tag{48}$$

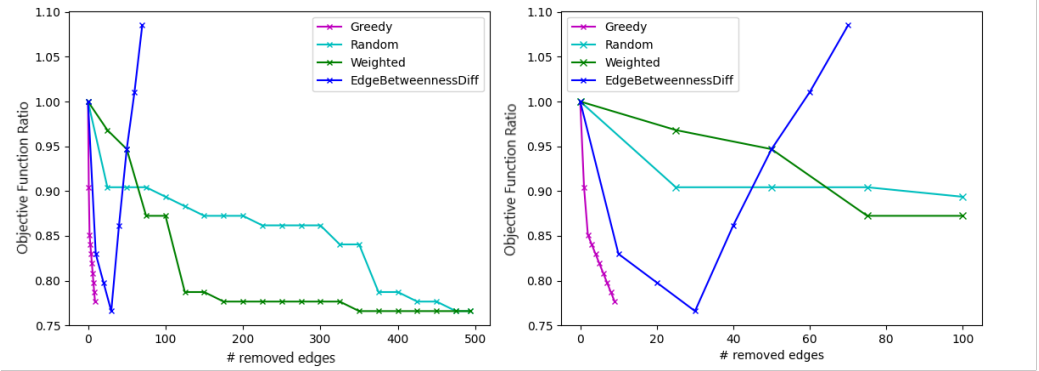
Όπως και προηγουμένως κάθε σχήμα περιέχει 2 γραφικές παραστάσεις, με τη δεύτερη να αποτελεί εστιασμένο μέρος της πρώτης με στόχο την αναλυτικότερη μελέτη της απόδοσης των μεθόδων που αφαιρούν μικρότερο πλήθος ακμών από τις υπόλοιπες. Είναι και πάλι εμφανής, αλλά και πιο έντονη η υπεροχή του προτεινόμενου άπληστου αλγορίθμου, αφού επιτυγχάνει συγκρίσιμη τελική τιμή της αντικειμενικής συνάρτησης με την επιτευγμένη μικρότερη αυτής με αισθητά μικρότερο πλήθος ακμών από τις υπόλοιπες μεθόδους. Σε αντίθεση με το μοντέλο LT, υπό το μοντέλο DLT δεν είναι ξεκάθαρη η κατάταξη της αποτελεσματικότητας των μεθόδων πλην της προτεινόμενης λόγω της συμπεριφοράς του αλγορίθμου "EdgeBetweennessDiff". Αυτός παρατηρείται να επιτυγχάνει καλά αποτελέσματα με την αφαίρεση λίγων ακμών και να είναι αποδοτικότερος από τις μεθόδους "Weighted" και "Random" αν λάβει κανείς υπόψη την ελάχιστη τιμή που επιτυγχάνει έχοντας αφαιρέσει λιγότερες ακμές από αυτές. Ωστόσο, το κριτήριο παύσης του αλγορίθμου όταν έχει εξαλειφθεί η παραπληροφόρηση, φαίνεται να μην ικανοποιείται ποτέ, με αποτέλεσμα να συνεχίζεται η αφαίρεση ακμών σε βάρος της διάδοσης της αληθούς πληροφορίας και ως εκ τούτου να αυξάνεται εκ νέου η τιμή της συνάρτησης. Είναι ιδιαίτερα ενδιαφέρουσα η περίπτωση του δικτύου Wikipedia vote, στο οποίο ο αλγόριθμος "EdgeBetweennessDiff" αφαιρεί λιγότερες ακμές από τις άλλες δύο συγκριτικές μεθόδους, λόγω του μικρού μεγέθους των επαγόμενων γράφων και ως εκ τούτου του μικρότερου πλήθους αφαιρέσιμων ακμών. Αν και επιτυγχάνει την ελάχιστη τιμή μεταξύ όλων των αλγορίθμων διαγράφοντας λίγες ακμές, συνεχίζει με την αφαίρεση επιπλέον ακμών μέχρι το μέγιστο δυνατό, με αποτέλεσμα εν τέλει να σημειώνει τελική τιμή της αντικειμενικής συνάρτησης μεγαλύτερη ακόμα και από την αρχική. Δηλαδή αντί να αντιμετωπίσει έστω μερικώς το πρόβλημα CMM, το επιδείνωσε μειώνοντας κατά μεγάλο βαθμό τη διάδοση της αληθούς πληροφορίας. Συνεπώς, όπως είναι ορισμένη η εν λόγω μέθοδος, λαμβάνοντας υπόψη τα τελικά αποτελέσματα, κρίνεται αναποτελεσματική.



Σχήμα 11: Γραφική παράσταση της μετρικής $ratio(k)$ ως προς το πλήθος των αφαιρεμένων ακμών k με εφαρμογή των μεθόδων Greedy, Random, Weighted και EdgeBetweennessDiff υπό το μοντέλο DLT για το δίκτυο email-Eu-core.



Σχήμα 12: Γραφική παράσταση της μετρικής $ratio(k)$ ως προς το πλήθος των αφαιρεμένων ακμών k με εφαρμογή των μεθόδων Greedy, Random, Weighted και EdgeBetweennessDiff υπό το μοντέλο DLT για το δίκτυο Social circles: Facebook.



Σχήμα 13: Γραφική παράσταση της μετρικής $ratio(k)$ ως προς το πλήθος των αφαιρεμένων ακμών k με εφαρμογή των μεθόδων Greedy, Random, Weighted και EdgeBetweennessDiff υπό το μοντέλο DLT για το δίκτυο Wikipedia vote.

8 Επίλογος

Στη συγκεκριμένη ενότητα παρουσιάζεται μια σύνοψη των αποτελεσμάτων της παρούσας Διπλωματικής Εργασίας και των συμπερασμάτων από την εφαρμογή του προτεινόμενου άπληστου επαναληπτικού αλγορίθμου και των συγκριτικών μεθόδων για την αντιμετώπιση του προβλήματος CMM. Τέλος, προτείνονται κάποιες ενδεχόμενες μελλοντικές επεκτάσεις της ερευνητικής εργασίας.

8.1 Σύνοψη και συμπεράσματα

Στο πλαίσιο της Εργασίας προτάθηκε ως παραλλαγή του πολυμελετημένου προβλήματος της ελαχιστοποίησης της παραπληροφόρησης εντός ενός κοινωνικού δικτύου, το πρόβλημα Cautious Misinformation Minimization (CMM), κατά το οποίο θεωρώντας την ανεξάρτητη διάδοση ειδήσεων που ανήκουν είτε στην κλάση ορθών πληροφοριών είτε στην κλάση ψευδών πληροφοριών επιδιώκεται μέσω της αφαίρεσης ακμών περιορισμένου πλήθους η ελαχιστοποίηση της διάδοσης της ψευδούς είδησης με την ταυτόχρονη ελαχιστοποίηση της μείωσης της διάδοσης της ορθής είδησης. Για την προσομοίωση της διάδοσης των πληροφοριών παρουσιάστηκαν 3 μοντέλα, IC, LT και DLT, εκ των οποίων τα μοντέλα IC και DLT λαμβάνουν υπόψη και την εξειδίκευση του χρήστη ως προς τη θεματική κατηγορία στην οποία ανήκουν οι διαδιδόμενες ειδήσεις. Στη συνέχεια, αποδείχθηκε ότι υπό τα παραπάνω μοντέλα το πρόβλημα CMM είναι NP-Hard και προτάθηκε ένας άπληστος επαναληπτικός αλγόριθμος για την επίλυσή του υπό τα μοντέλα LT και DLT. Το βασικό κριτήριο του εν λόγω αλγορίθμου για την αφαίρεση ακμής αποτελεί η μέγιστη μείωση της τιμής της αντικειμενικής συνάρτησης ή η ελάχιστη προκύπτουσα τιμή της αντικειμενικής συνάρτησης αντίστοιχα. Υπό το μοντέλο LT, αξιοποιείται ο πιθανοτικός χαρακτήρας του, που αποτυπώνεται στην επιλογή τυχαίων τιμών για τα κατώφλια των κόμβων, ούτως ώστε να κατασκευαστούν δειγματοληπτημένοι live-edge γράφοι και τα αντίστοιχα δένδρα επιρροής για έκαστο κόμβο εκ των αρχικών υποστηρικτών μιας είδησης, και να είναι αποδοτικότερος ο υπολογισμός της μεταβολής της αντικειμενικής συνάρτησης μετά την αφαίρεση μιας οποιασδήποτε ακμής. Όμοια, υπό το μοντέλο DLT προσομοιώνεται ντετερμινιστικά και υπολογίζεται με ακρίβεια η διάδοση των ειδήσεων στο αρχικό δίκτυο. Τέλος, εκτελώντας πειράματα πάνω σε 3 πραγματικά κοινωνικά δίκτυα, αξιολογήθηκε η απόδοση των αλγορίθμων συγκρινόμενη με την αντίστοιχη άλλων 3 ευριστικών μεθόδων, εκ των οποίων οι 2 πρώτες συναντούνται συχνά στη βιβλιογραφία ως επίπεδο αναφοράς και η τρίτη επιχειρεί να αξιοποιήσει κυρίως την τοπολογία του δικτύου. Σε κάθε περίπτωση, παρατηρήθηκε ότι ο προτεινόμενος άπληστος αλγόριθμος επιτυγχάνει γρήγορη μείωση της αντικειμενικής συνάρτησης με την αφαίρεση μικρού πλήθους ακμών, υπερέχοντας των άλλων συγκριτικών μεθόδων.

8.2 Μελλοντικές κατευθύνσεις

Η παρούσα Διπλωματική Εργασία διαθέτει ενδιαφέρουσες επεκτάσεις οι οποίες θα μπορούσαν να μελετηθούν μελλοντικά. Μερικές προτεινόμενες εξ αυτών είναι οι

ακόλουθες:

- Εκμάθηση των παραμέτρων των μοντέλων διάδοσης, αντί της ανάθεσης τυχαίας τιμής στο βάρος μιας ακμής (ισχύς επιρροής ενός κόμβου στον άλλο) και στην παράμετρο εξειδίκευσης ενός χρήστη. Αυτό θα μπορούσε να επιτευχθεί αξιοποιώντας εργαλεία από το χώρο της Μηχανικής Μάθησης, ούτως ώστε καταγράφοντας και παρατηρώντας παρελθοντικές διαδόσεις ειδήσεων εντός του κοινωνικού δικτύου να είναι δυνατή η πρόβλεψη των τιμών των παραμέτρων. Με αυτόν τον τρόπο, θα ήταν πιο ρεαλιστικά και ουσιώδη τα αποτελέσματα των πειραμάτων.

- Μελέτη του προβλήματος CMM υπό το μοντέλο IC, υιοθετώντας ακόμη και την ίδια ιδέα του άπληστου επαναληπτικού αλγορίθμου. Η διεξαγωγή πειραμάτων υπό αυτό το μοντέλο κρίθηκε υπολογιστικά απαγορευτική για τις δυνατότητες του διαθέσιμου μηχανήματος. Επομένως, διαθέτοντας μεγαλύτερη υπολογιστική ισχύ ενδέχεται να είναι δυνατή η εκτέλεση πειραμάτων υπό το μοντέλο IC και η εξαγωγή αντίστοιχων χρήσιμων συμπερασμάτων.

- Εύρεση κατάλληλου κριτηρίου παύσης της μεθόδου "EdgeBetweennessDiff' υπό το μοντέλο DLT, η οποία αποδείχθηκε αποτελεσματική με την αφαίρεση μικρού πλήθους ακμών. Είναι εύλογο να ερευνήσει κανείς τα πιθανά κριτήρια με βάση τα οποία θα μπορούσε να κριθεί ότι η αφαίρεση περαιτέρω ακμών δε θα επιφέρει κάποια εκ νέου μείωση της αντικειμενικής συνάρτησης. Κατά αυτό τον τρόπο, ίσως αναδειχθεί μια καλή μέθοδος αντιμετώπισης του προβλήματος με την αξιοποίηση κυρίως τοπολογικών χαρακτηριστικών και δευτερευόντως της διάδοσης των πληροφοριών.

- Μελέτη δυναμικών δικτύων, αφού τα πειράματα διεξήχθησαν βάσει στατικών δεδομένων. Στην πραγματικότητα, τα κοινωνικά δίκτυα αλλάζουν συνεχώς στη διάρκεια του χρόνου. Επομένως, φαίνεται ενδιαφέρουσα η μελέτη του προβλήματος CMM και η εύρεση ενός προσαρμοστικού αλγορίθμου που να λαμβάνει υπόψη τη διάδοση πάνω στο εξελισσόμενο δίκτυο. Αυτό θα απέφερε αρκετά ρεαλιστικά αποτελέσματα, αλλά πιθανότατα απαιτεί μεγάλο υπολογιστικό φόρτο.

Αναφορές

[1] https://datareportal.com/social-media-users

[2] https://www.marketwatch.com/story/this-day-in-history-hacked-ap-tweet-about-white-house-explosions-triggers-panic-2018-04-23

[3] Shu, K., Bernard, H.R., Liu, H., 2019. Studying Fake News via Network Analysis: Detection and Mitigation. Agarwal, N., Dokoohaki, N., Tokdemir, S. (eds) Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining. Lecture Notes in Social Networks. Springer, Cham.

[4] Zareie, A., Sakellariou, R., 2021. Minimizing the spread of misinformation in online social networks: A survey. Journal of Network and Computer Applications, Volume 186, 103094, ISSN 1084-8045.

[5] Kempe, D., Kleinberg, J., Tardos, E., 2003. Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03). Association for Computing Machinery, New York, NY, USA, 137–146.

[6] Khalil, E.B., Dilkina, B., Song, L., 2014. Scalable diffusion-aware optimization of network topology. Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1226–1235.

[7] Bondy, A., Murty, U.S.R., 2008. Graph Theory. Springer.

[8] Dijkstra, E. W., 1959. A note on two problems in connexion with graphs. Numerische Mathematik. 1: 269–271.

[9] Mehlhorn, K., Sanders, P., 2008. Chapter 10. Shortest Paths. Algorithms and Data Structures: The Basic Toolbox. Springer.

[10] Bellman, R., 1958. On a routing problem. Quarterly of Applied Mathematics. 16: 87–90.

[11] Ford, L. R. Jr., 1956. Network Flow Theory. Paper P-923. Santa Monica, California: RAND Corporation.

[12] Floyd, R. W., 1962. Algorithm 97: Shortest Path. Communications of the ACM. 5 (6): 345.

[13] Roy, B., 1959. Transitivité et connexité. C. R. Acad. Sci. Paris. 249: 216–218.

[14] Warshall, S., 1962. A theorem on Boolean matrices. Journal of the ACM. 9 (1): 11–12.

[15] Pan, C., Han, Y., Lu, J., 2020. Design and Optimization of Lattice Structures: A Review. Appl. Sci. 10, 6374.

[16] Barabási, A.-L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

[17] Watts, D., Strogatz, S., 1998. Collective dynamics of "small-world" networks. Nature.

[18] Biggs, N. L., 1993. Algebraic Graph Theory (2nd ed.), Cambridge Mathematical Library.

[19] Penrose, M., 2003. Random geometric graphs. Oxford: Oxford University Press.

[20] Bollobás, B., 2001. Random Graphs (2nd ed.). Cambridge University Press.

[21] Easley, D., Kleinberg, J., 2010. Networks, Crowds and Markets: Reasoning about a Highly Connected World. Cambridge University Press.

[22] Borgatti, S. P., 2005. Centrality and network flow. Social Networks, Volume 27, Issue 1, Pages 55-71, ISSN 0378-8733.

[23] Freeman, L., 1979. Centrality in networks: I. Conceptual clarification. Social Networks 1:215-239.

[24] Freeman L., 1977. A set of measures of centrality based on betweenness. Sociometry 40: 35–41.

[25] Newman, M.E., 2010. Networks: An Introduction. Oxford University Press, USA, pp. 169.

[26] Girvan, M., Newman, M.E., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. 99, 7821–7826.

[27] Brandes, U. 2001. A Faster Algorithm for Betweenness Centrality. The Journal of Mathematical Sociology 25 (2): 163–177.

[28] Brandes, U. 2008. On variants of shortest-path betweenness centrality and their generic computation. Social Networks, Volume 30, Issue 2, Pages 136-145, ISSN 0378-8733.

[29] Goldenberg, J., Libai, B., Muller, E., 2001. Using Complex Systems Analysis to Advance Marketing Theory Development. Academy of Marketing Science Review.

[30] Goldenberg, J., Libai, B., Muller, E., 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Wordof-Mouth. Marketing Letters, 12(3): pp. 211–223.

[31] Granovetter, M., 1978. Threshold Models of Collective Behavior. American Journal of Sociology, 83(6): pp. 1420–1443.

[32] Schelling T., 1978. Micromotives and Macrobehavior. Norton.

[33] Chou, Chung-Kuang, Chen, M., 2015. Multiple factors-aware diffusion in social networks. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham.

[34] Lu, Z., Zhang, W., Wu, W., 2012. The complexity of influence maximization problem in the deterministic linear threshold model. J Comb Optim 24, 374–378.

[35] Litou, I., Kalogeraki, V., Katakis, I., Gunopulos, D., 2017. Efficient and timely misinformation blocking under varying cost constraints, Online Social Networks and Media, Volume 2, Pages 19-31, ISSN 2468-6964.

[36] Canzani, E., Lechner, U., 2015. Insights from Modeling Epidemics of Infectious Diseases – A Literature Review.

[37] Scheibehenne, B., Rainer, G., Peter, M. T., 2010. Can there ever be too many options? A meta-analytic review of choice overload. Journal of Consumer Research 37 (3): 409–425.

[38] Chen, W., Wang, C., Wang, Y., 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10). Association for Computing Machinery, New York, NY, USA, 1029–1038.

[39] Chen, W., Yuan, Y., Zhang, L., 2010. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. IEEE International Conference on Data Mining, pp. 88-97.

[40] https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/

[41] Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. International Conference on Machine Learning. pp. 1188–1196.

[42] Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543.

[43] Abbasi, M.A., Liu, H., 2013. Measuring user credibility in social media. SBP. pp.441–448. Springer.

[44] Shu, K., Wang, S., Liu, H., 2017. Exploiting tri-relationship for fake news detection.arXiv preprint arXiv:1712.07709.

[45] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19(1), 22–36.

[46] Ruchansky, N., Seo, S., Liu, Y., 2017. Csi: A hybrid deep model for fake news. arXiv preprint arXiv:1703.06959.

[47] Wu, L., Liu, H., 2018. Tracing fake-news footprints: Characterizing social media messages by how they propagate.

[48] Perozzi, B., Al-Rfou, R., Skiena, S., 2014. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 701–710. ACM.

[49] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q., 2015. Line: Large-scale information network embedding. Proceedings of the 24th International Conference on World Wide Web. pp. 1067–1077. International World Wide Web Conferences Steering Committee.

[50] Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., Yang, S., 2017. Community preserving network embedding. AAAI. pp. 203–209.

[51] Newman, M.E., 2006. Finding community structure in networks using the eigenvectors of matrices. Physical review E 74(3), 036104.

[52] Shu, K., Wang, S., Liu, H., 2018. Understanding user profiles on social media for fake news detection. 1st IEEE International Workshop on "Fake Multi-Media"(FakeMM'18). IEEE.

[53] Jin, Z., Cao, J., Zhang, Y., Luo, J., 2016. News verification by exploiting conflicting social viewpoints in microblogs. AAAI. pp. 2972–2978.

[54] Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S., de Alfaro, L., 2017. Some like it hoax: Automated fake news detection in social networks. arXiv preprint. arXiv:1704.07506.

[55] Hosni, A.I.E., Li, K., Ahmad, S., 2019. DARIM: Dynamic approach for rumor influence minimization in online social networks. In: International Conference on Neural Information Processing. Springer, pp. 619–630.

[56] Wang, B., Chen, G., Fu, L., Song, L., Wang, X., 2017. Drimux: Dynamic rumor influence minimization with user experience in social networks. IEEE Trans. Knowl. Data Eng.29 (10), 2168–2181.

[57] Lü, L., Chen, D., Ren, X.-L., Zhang, Q.-M., Zhang, Y.-C., Zhou, T., 2016. Vital nodes identification in complex networks. Phys. Rep. 650, 1–63.

[58] Chen, W., Wang, C., Wang, Y., 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1029–1038.

[59] Borgs, C., Brautbar, M., Chayes, J., Lucier, B., 2014. Maximizing social influence in nearly optimal time. Proceedings of the 25th Annual Symposium on Discrete Algorithms. SIAM, pp. 946–957.

[60] Tang, Y., Xiao, X., Shi, Y., 2014. Influence maximization: Near-optimal time complexity meets practical efficiency. Proceedings of the 2014 International Conference on Management of Data. ACM, pp. 75–86.

[61] Ruchansky, N., Seo, S., Liu, Y., 2017. Csi: A hybrid deep model for fake news. arXiv preprint arXiv:1703.06959.

[62] Holme, P., Kim, B.J., Yoon, C.N., Han, S.K., 2002. Attack vulnerability of complex networks. Phys. Rev. E 65 (5), 056109.

[63] Schneider, C.M., Mihaljev, T., Havlin, S., Herrmann, H.J., 2011. Suppressing epidemics with a limited amount of immunization units. Phys. Rev. E 84 (6), 061911.

[64] Dey, P., Roy, S., 2017. Centrality based information blocking and influence minimization in online social network. 2017 International Conference on Advanced Networks and Telecommunications Systems. IEEE, pp. 1–6.

[65] Wang, S., Zhao, X., Chen, Y., Li, Z., Zhang, K., Xia, J., 2013. Negative influence minimizing by blocking nodes in social networks. Proceedings of the 17th Conference on Late-Breaking Developments in the Field of Artificial Intelligence. AAAI Press, pp. 134–136.

[66] Tanınmış, K., Aras, N., Altınel, İ.K., Güney, E., 2020. Minimizing the misinformation spread in social networks. IISE Trans. 52 (8), 850–863.

[67] Yao, Q., Shi, R., Zhou, C., Wang, P., Guo, L., 2015. Topic-aware social influence minimization. Proceedings of the 24th International Conference on World Wide Web. ACM, pp. 139–140.

[68] Pham, C.V., Thai, M.T., Duong, H.V., Bui, B.Q., Hoang, H.X., 2018. Maximizing misinformation restriction within time and budget constraints. J. Combin. Optim. 35 (4), 1202–1240.

[69] Wu, Q., Wang, T., Cai, Y., Tian, H., Chen, Y., 2017. Rumor restraining based on propagation prediction with limited observations in large-scale social networks. Proceedings of the Australasian Computer Science Week Multiconference. ACM, pp. 1–8.

[70] Pham, C.V., Phu, Q.V., Hoang, H.X., 2018. Targeted misinformation blocking on online social networks. In: Asian Conference on Intelligent Information and Database Systems. Springer, pp. 107–116.

[71] Pham, C.V., Phu, Q.V., Hoang, H.X., Pei, J., Thai, M.T., 2019. Minimum budget for misinformation blocking in online social networks. J. Combin. Optim. 38 (4), 1101–1127.

[72] Zheng, J., Pan, L., 2018. Least cost rumor community blocking optimization in social networks. 3rd International Conference on Security of Smart Cities, Industrial Control System and Communications. IEEE, pp. 1–5.

[73] Yang, D., Liao, X., Shen, H., Cheng, X., Chen, G., 2018. Dynamic node immunization for restraint of harmful information diffusion in social networks. Physica A 503, 640–649.

[74] Shi, Q., Wang, C., Ye, D., Chen, J., Feng, Y., Chen, C., 2019. Adaptive influence blocking: Minimizing the negative spread by observation-based policies. 35th International Conference on Data Engineering. IEEE, pp. 1502–1513.

[75] Song, C., Hsu, W., Lee, M.L., 2015. Node immunization over infectious period. Proceedings of the 24th International Conference on Information and Knowledge Management. ACM, pp. 831–840.

[76] Pham, C.V., Dinh, H.M., Nguyen, H.D., Dang, H.T., Hoang, H.X., 2017. Limiting the spread of epidemics within time constraint on online social networks. Proceedings of the 8th International Symposium on Information and Communication Technology. ACM, pp. 262–269.

[77] Wijayanto, A.W., Murata, T., 2019. Effective and scalable methods for graph protection strategies against epidemics on dynamic networks. Appl. Netw. Sci. 4 (18), 1–32.

[78] Kimura, M., Saito, K., Motoda, H., 2008. Solving the contamination minimization problem on networks for the linear threshold model. Pacific Rim International Conference on Artificial Intelligence. Springer, pp. 977–984.

[79] Kimura, M., Saito, K., Motoda, H., 2009. Blocking links to minimize contamination spread in a social network. ACM Trans. Knowl. Discov. Data 3 (2), 1–23.

[80] Kimura, M., Saito, K., Nakano, R., 2007. Extracting influential nodes for information diffusion on a social network. Proceedings of the 22nd National Conference on Artificial Intelligence, vol. 2. AAAI Press, pp. 1371–1376.

[81] Kimura, M., Saito, K., Motoda, H., 2008. Minimizing the spread of contamination by blocking links in a network. Proceedings of the 23rd National Conference on Artificial Intelligence. 2, AAAI Press, pp. 1175–1180.

[82] Khalil, E., Dilkina, B., Song, L., 2013. CuttingEdge: influence minimization in networks. Proceedings of the Workshop on Frontiers of Network Analysis: Methods, Models, and Applications at NIPS.

[83] Tong, H., Prakash, B.A., Eliassi-Rad, T., Faloutsos, M., Faloutsos, C., 2012. Gelling, and melting, large graphs by edge manipulation. Proceedings of the 21st International Conference on Information and Knowledge Management. ACM, pp.245–254.

[84] Yao, Q., Zhou, C., Xiang, L., Cao, Y., Guo, L., 2014. Minimizing the negative influence by blocking links in social networks. International Conference on Trustworthy Computing and Services. Springer, pp. 65–73.

[85] Yan, R., Li, Y., Wu, W., Li, D., Wang, Y., 2019. Rumor blocking through online link deletion on social networks. ACM Trans. Knowl. Discov. Data 13 (2), 1–26.

[86] Kuhlman, C.J., Tuli, G., Swarup, S., Marathe, M.V., Ravi, S., 2013. Blocking simple and complex contagion by edge removal. 13th International Conference on Data Mining. IEEE, pp. 399–408.

[87] Wang, X., Deng, K., Li, J., Yu, J.X., Jensen, C.S., Yang, X., 2018. Targeted influence minimization in social networks. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 689–700.

[88] Papadimitriou, C.H., Steiglitz, K., 1982. Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall.

[89] Song, Y., Dinh, T.N., 2014. Optimal containment of misinformation in social media: A scenario-based approach. International Conference on Combinatorial Optimization and Applications. Springer, pp. 547–556.

[90] He, X., Song, G., Chen, W., Jiang, Q., 2012. Influence blocking maximization in social networks under the competitive linear threshold model. Proceedings of the International Conference on Data Mining. SIAM, pp. 463–474.

[91] Zhang, H., Zhang, H., Li, X., Thai, M.T., 2015. Limiting the spread of misinformation while effectively raising awareness in social networks. International Conference on Computational Social Networks. Springer, pp. 35–47.

[92] Liu, W., Yue, K., Wu, H., Li, J., Liu, D., Tang, D., 2016. Containment of competitive influence spread in social networks. Knowl.-Based Syst. 109, 266–275.

[93] Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank Citation Ranking : Bringing Order to the Web. WWW 1999.

[94] Yang, L., Li, Z., Giua, A., 2019. Rumor containment by spreading correct information in social networks. American Control Conference. IEEE, pp. 5608–5613.

[95] Yang, L., Li, Z., Giua, A., 2020. Containment of rumor spread in complex social networks. Inform. Sci. 506, 113–130.

[96] Budak, C., Agrawal, D., El Abbadi, A., 2011. Limiting the spread of misinformation in social networks. Proceedings of the 20th International Conference on World Wide Web. ACM, pp. 665–674.

[97] Wu, P., Pan, L., 2017. Scalable influence blocking maximization in social networks under competitive independent cascade models. Comput. Netw. 123, 38–50.

[98] Arazkhani, N., Meybodi, M.R., Rezvanian, A., 2019. Influence blocking maximization in social network using centrality measures. 5th Conference on Knowledge Based Engineering and Innovation. IEEE, pp. 492–497.

[99] Lv, J., Yang, B., Yang, Z., Zhang, W., 2019. A community-based algorithm for influence blocking maximization in social networks. Cluster Comput. 22 (3), 5587–5602.

[100] Tong, G., Wu, W., Guo, L., Li, D., Liu, C., Liu, B., Du, D.-Z., 2020. An efficient randomized algorithm for rumor blocking in online social networks. IEEE Trans. Netw. Sci. Eng. 7 (2), 845–854.

[101] Tong, G.A., Du, D.-Z., 2019. Beyond uniform reverse sampling: A hybrid sampling technique for misinformation prevention. Conference on Computer Communications. IEEE, pp. 1711–1719.

[102] Li, S., Zhu, Y., Li, D., Kim, D., Huang, H., 2013. Rumor restriction in online social networks. 32nd International Performance Computing and Communications Conference. IEEE, pp. 1–10.

[103] Tong, G., Wu, W., Du, D.-Z., 2018. Distributed rumor blocking with multiple positive cascades. IEEE Trans. Comput. Soc. Syst. 5 (2), 468–480.

[104] Song, C., Hsu, W., Lee, M.L., 2017. Temporal influence blocking: minimizing the effect of misinformation in social networks. 33rd International Conference on Data Engineering. IEEE, pp. 847–858.

[105] Fan, L., Wu, W., Zhai, X., Xing, K., Lee, W., Du, D.-Z., 2014. Maximizing rumor containment in social networks with constrained time. Soc. Netw. Anal. Min. 4 (1), 214–224.

[106] Hosni, A.I.E., Li, K., Ahmad, S., 2020. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. Inform. Sci. 512, 1458–1480.

[107] Hosni, A.I.E., Li, K., Ahmed, S., 2018. Hisbmodel: a rumor diffusion model based on human individual and social behaviors in online social networks. International Conference on Neural Information Processing. Springer, pp. 14–27.

[108] Chen, T., Liu, W., Fang, Q., Guo, J., Du, D.-Z., 2019. Minimizing misinformation profit in social networks. IEEE Trans. Comput. Soc. Syst. 6 (6), 1206–1218.

[109] Tong, G.A., Wu, W., Du, D.-Z., 2018. On misinformation containment in online social networks. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Curran Associates Inc., pp. 339–349.

[110] Zhu, W., Yang, W., Xuan, S., Man, D., Wang, W., Du, X., 2018. Location-aware influence blocking maximization in social networks. IEEE Access 6, 61462–61477.

[111] Zhu, W., Yang, W., Xuan, S., Man, D., Wang, W., Du, X., Guizani, M., 2019. Location based seeds selection for influence blocking maximization in social networks. IEEE Access 7, 27272–27287.

[112] Wu, Y., Huang, H., Zhao, J., Wang, C., Wang, T., 2018. Using mobile nodes to control rumors in big data based on a new rumor propagation model in vehicular social networks. IEEE Access 6, 62612–62621.

[113] Nguyen, N.P., Yan, G., Thai, M.T., Eidenbenz, S., 2012. Containment of misinformation spread in online social networks. Proceedings of the 4th Annual Web Science Conference. ACM, pp. 213–222.

[114] Nguyen, N.P., Yan, G., Thai, M.T., 2013. Analysis of misinformation containment in online social networks. Comput. Netw. 57 (10), 2133–2146.

[115] Fan, L., Lu, Z., Wu, W., Thuraisingham, B., Ma, H., Bi, Y., 2013. Least cost rumor blocking in social networks. 2013 IEEE 33rd International Conference on Distributed Computing Systems. IEEE, pp. 540–549.

[116] Hosni, A.I.E., Li, K., Ding, C., Ahmed, S., 2018. Least cost rumor influence minimization in multiplex social networks. International Conference on Neural Information Processing. Springer, pp. 93–105.

[117] Z. Lu, W. Zhang, W. Wu, B. Fu and D. Du, 2011. Approximation and Inapproximation for the Influence Maximization Problem in Social Networks under Deterministic Linear Threshold Model. 31st International Conference on Distributed Computing Systems Workshops, 2011, pp. 160-165.

[118] Karp, R.M., 1972. Reducibility Among Combinatorial Problems. Complexity of Computer Computations. pp. 85–104. Plenum Press, New York.

[119] Nemhauser, G., Wolsey, L., Fisher, M., 1978. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 14, 265–294.

[120] Schrijver, A., 2003. Combinatorial Optimization: Polyhedra and Efficiency. Volume 24 of Algorithms and Combinatorics. Springer.

[121] Feige, U., 1998. A Threshold of ln n for Approximating Set Cover. J. of the ACM 45(5): 634–652.

[122] Kellerer, H., Pferschy, U., Pisinger, D., 2004. Introduction to NP-Completeness of Knapsack Problems. Knapsack Problems, 483–493.

[123] https://networkx.org/

[124] https://www.jetbrains.com/pycharm/

[125] https://github.com/IlianaXn/thesis_min_misinformation

[126] Leskovec, J.. https://snap.stanford.edu/data/index.html

[127] Leskovec, J., Kleinberg, J., Faloutsos, C., 2007. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1).

[128] McAuley, J., Leskovec, J., 2012. Learning to Discover Social Circles in Ego Networks. NIPS 2012.

[129] Leskovec, J., Huttenlocher, D., Kleinberg, J., 2010. Signed Networks in Social Media. CHI 2010.

[130] Leskovec, J., Huttenlocher, D., Kleinberg, D., 2010. Predicting Positive and Negative Links in Online Social Networks. WWW 2010.

[131] https://en.wikipedia.org/wiki/Main_Page

[132] Fisher, R., Yates, F., 1948. Statistical tables for biological, agricultural and medical research (3rd ed.). London: Oliver & Boyd. pp. 26–27.

[133] Durstenfeld, R., 1964. Algorithm 235: Random permutation. Communications of the ACM. 7 (7): 420.