Εθνικο Μετσοβιο Πολυτεχνειο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Επικοινωνιων, ηλεκτρονικης
και συστηματων πληροφορικης

Συμπερασμός δικτύων διάχυσης πληροφορίας μέσω αποδοτικών αλγορίθμων ιχνηλάτησης περιεχομένου σε πλατφόρμες κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γκλιάτης Α. Απόλλων - Θεόδωρος

Επιβλέπων : Παπαβασιλείου Συμεών
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 25η Οκτωβρίου 2023.

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

Αθήνα, Ιούλιος 2024

.....

Γκλιάτης Α. Απόλλων - Θεόδωρος

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

ΠΕΡΙΛΗΨΗ

Στα πλαίσια της Ανάλυσης Κοινωνικών Δικτύων, ιδιαίτερο ενδιαφέρον σημειώνεται στη μελέτη της διάχυσης της πληροφορίας ως αποτέλεσμα της αλληλεπίδρασης μεταξύ των χρηστών ενός δικτύου. Ο σκοπός της παρούσας διπλωματικής εργασίας είναι ο συμπερασμός της ροής διάδοσης της πληροφορίας σε μια πλατφόρμα κοινωνικής δικτύωσης (Online Social Network - OSN), λαμβάνοντας υπόψη τη δημόσια δραστηριότητα και κατ' επέκταση το δημόσιο διαμοιρασμό περιεχομένου μεταξύ των χρηστών.

Η τυπική αναπαράσταση των Online Social Networks και η μελέτη των συσχετίσεων μεταξύ των συστατικών τους μερών γίνεται μέσω της Θεωρίας Γραφημάτων. Στην παρούσα εργασία, η διάχυση της πληροφορίας εξετάζεται σε δίκτυα που κατασκευάζονται με το μοντέλο Small-World, καθώς αυτό το μοντέλο έχει δομικά χαρακτηριστικά παραπλήσια με αυτά των κοινωνικών δικτύων που συναντώνται σήμερα στο διαδίκτυο. Για την μοντελοποίηση της διάδοσης της πληροφορίας, χρησιμοποιείται το ανεξάρτητο πολλαπλασιαστικό μοντέλο (Independent Cascade), οι πιθανότητες του οποίου διαμορφώνονται από μετρικές της Ανάλυσης Σύνθετων Δικτύων. Η διάχυση της πληροφορίας μελετάται σε δίκτυα με τρεις κατηγορίες χρηστών, βάσει των ρυθμίσεων απορρήτου τους. Η πρώτη κατηγορία περιλαμβάνει τους χρήστες με δημόσιο προφίλ και δραστηριότητα. Οι χρήστες αυτοί αναφέρονται ως χρήστες - monitors και αξιοποιούνται για την παρακολούθηση της ροής της πληροφορίας στο δίκτυο. Η δεύτερη κατηγορία περιλαμβάνει τους χρήστες με πιο αυστηρές ρυθμίσεις απορρήτου, για τους οποίους όμως είναι διαθέσιμη η πληροφορία σχετικά με τη χρονική στιγμή που συμμετείχαν στη διάδοση της πληροφορίας εντός του δικτύου. Στην τρίτη κατηγορία ανήκουν οι χρήστες που δεν παρέχουν πληροφορία για τίποτα από τα παραπάνω.

Τέλος, λαμβάνοντας υπόψη τα δομικά και συμπεριφορικά χαρακτηριστικά του δικτύου, όπως αυτά διαμορφώνονται από την αλληλεπίδραση των χρηστών, καθως και την πληροφορία που λαμβάνεται από τους χρήστες - monitors, επιχειρείται ο συμπερασμός του δικτύου διάχυσης της πληροφορίας (diffusion network) μέσω ενός σχήματος πιθανοτικής οπισθοδρόμησης. Εξάγεται δηλαδή το δίκτυο των χρηστών που έλαβαν την πληροφορία, διασφαλίζοντας παράλληλα και την γνώση της πηγής από την οποία την έλαβαν. Τέλος, η μεθοδολογία της πιθανοτικής οπισθοδρόμησης αξιολογείται ως προς την ακρίβεια του συμπερασμού μέσω προσομοιώσεων σε δίκτυα ποικίλων δομικών χαρακτηριστικών.

Λέξεις κλειδιά: Ανάλυση Κοινωνικών Δικτύων, Κοινωνικά Δίκτυα στο Διαδίκτυο, Δίκτυο Διάχυσης Πληροφορίας, Αλγόριθμος Independent Cascade, Συμπερασμός Δικτύου Διάχυσης Πληροφορίας, Κοινωνικά Δίκτυα, Θεωρία Γραφημάτων

ABSTRACT

An interesting area of research in social network analysis is the inference of the information flow between the users of a network as well as the way in which the relationships between the users affects it. The purpose of this diploma thesis is the inference of the flow that a diffusion process follows inside an Online Social Network (OSN), by using as input the trading of information performed between pairs of users of the OSN with a public profile.

Graph theory, and in particular graphs, are used for the representation of online social networks. Small-World (SW) graphs are considered as the most appropriate ones for the representation of the structure and evolution of social networks. In this diploma thesis, information diffusion is modeled by the Independent Cascade model. The cascade probabilities are determined by Complex Network Analysis metrics. The process of the information diffusion is studied in networks of users which are divided into three categories, based on their privacy settings. The first category consists of users with a public profile and activity, which are referenced as users - monitors and are utilized for the monitoring of the flow of information inside the network. The second category contains users with more strict privacy settings. However, for those users it is possible to know the timestamp at which they contributed to the diffusion process. The third category consists of users for which no information is available, regarding any of the above.

Finally, based on the network topology, the association between the users and the information that is received from the users - monitors, inference of the information diffusion network -which consists of the users that obtained the information, as well as the source from which they got it - is attempted. This diploma thesis also consists of the evaluation of the inference process in terms of accuracy. This is achieved by the extensive demonstration and review of the results that occurred from the simulations on synthetic networks of different structural characteristics.

Keywords: Social Network Analysis, Online Social Networks, Information Diffusion Network, Independent Cascade Algorithm, Inference of Information Diffusion Network, Social Networks, Graph Theory

Εισαγωγή

1.1 Πλατφόρμες κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης (social media) είναι διαδραστικές τεχνολογίες οι οποίες επιτρέπουν και ενθαρρύνουν τους χρήστες τους να δημιουργούν και να μοιράζονται περιεχόμενο, πληροφορίες, ιδέες, ενδιαφέροντα και άλλες μορφές έκφρασης μέσω εικονικών ομάδων και δικτύων.
Οι χρήστες αποκτούν πρόσβαση στα μέσα κοινωνικής δικτύωσης μέσω των πλατφορμών κοινωνικής δικτύωσης οι οποίες παρέχουν τόσο web-based εκδόσεις, όσο και για smartphones ή/και tablets.

Μέσω των πλατφορμών, κάθε χρήστης μπορεί να ανταλλάξει πληροφορία γραπτά ή προφορικά, να δημιουργήσει ή να επεξεργαστεί ήδη υπάρχον περιεχόμενο. Επιπρόσθετα, τα social media χρησιμοποιούνται για να αποθηκεύονται αναμνήσεις και να έρχεται ο χρήστης σε επαφή με νέα ενδιαφέροντα.
Δε θα μπορούσε επίσης, να παραληφθεί η δυνατότητα που δίνουν τα social media σε συλλόγους, οργανισμούς και επιχειρήσεις, να διαφημίσουν τα προϊόντα και τις υπηρεσίες τους. Μάλιστα, τα τελευταία χρόνια έχει αναπτυχθεί η δυνατότητα διαφήμισης σε συγκεκριμένες ομάδες ανθρώπων, ανάλογα την ηλικία, το φύλο, την περιοχή κ.α.
Ένας ακόμη χώρος στον οποίο τα social media έχουν κάνει μεγάλη διαφορά είναι αυτός της ενημέρωσης. Σε αντίθεση με τα παραδοσιακά μέσα μαζικής ενημέρωσης (τηλεόραση, εφημερίδα, ραδιόφωνο) το γεγονός πως τα social media, ως μέρος του internet, παρέχουν άμεση πρόσβαση σε κάθε γεγονός ανάλογα με τα ενδιαφέροντα του κάθε χρήστη καθώς και το ότι υπάρχει η δυνατότητα για πολλή περισσότερη πολυφωνία, αφού ο καθένας σχολιάζει και "αντιδρά", έχει οδηγήσει τους χρήστες να στρέφονται πλέον σχεδόν αποκλειστικά σε αυτά για την ενημέρωσή τους. Αυτό, όπως είναι λογικό, έχει διπλή ανάγνωση. Δεν είναι λίγες οι φορές που έχει παρατηρηθεί δημιουργία και διακίνηση fake news μέσω των social media με αποτελέσματα επικίνδυνα έως και καταστροφικά για το κοινωνικό σύνολο.

Μπορεί εύκολα να συμπεραθεί πως ο περιορισμός της διακίνησης ψευδούς πληροφορίας (fake news) θα γινόταν ευκολότερος αν υπήρχε η δυνατότητα εξακρίβωσης της προέλευσης της πληροφορίας. Επιπρόσθετα, γίνεται αντιληπτό πως η μελέτη της διάχυσης της πληροφορίας σε ένα δίκτυο στα μέσα κοινωνικής δικτύωσης θα είχε ενδιαφέρον και για άλλους σκοπούς, όπως η διαφήμιση, η οποία αναφέρθηκε προηγουμένως.

Αυτά αποτελούν μερικά μόνο παραδείγματα σχετιζόμενα με ένα ευρύτερο πλαίσιο στο οποίο φαίνεται η σημασία της δυνατότητας να προβλεφθεί η διάχυση της πληροφορίας σε μια πλατφόρμα κοινωνικής δικτύωσης.

1.2 Αντικείμενο Διπλωματικής και Συνεισφορά

Για τη μελέτη των πλατφορμών κοινωνικής δικτύωσης, είναι χρήσιμο να ληφθεί υπόψιν πως αυτές ανήκουν στην ευρύτερη κατηγορία των σύνθετων δικτύων. Ως Σύνθετα Δίκτυα (complex networks) ορίζονται τα δίκτυα στα οποία παρουσιάζονται συμπεριφορές που δε μπορούν να προβλεφθούν a priori με βάση τα γνωστά χαρακτηριστικά των συστατικών μερών τους. Η μελέτη των σύνθετων δικτύων αποτελεί έναν νέο τομέα επιστημονικής έρευνας και αφορμάται από την εμπειρική μελέτη πραγματικών δικτύων όπως δίκτυα υπολογιστών, τεχνολογίας, αλλά και κοινωνικών δικτύων [1]. Συνεπώς, για την μελέτη των πλατφορμών κοινωνικής δικτύωσης χρησιμοποιείται η Ανάλυση Κοινωνικών Δικτύων, η οποία αποτελεί τμήμα της Ανάλυσης Σύνθετων Δικτύων. Στο ευρύτερο επιστημονικό πεδίο της ανάλυσης κοινωνικών δικτύων ανήκει και η εξαγωγή του δικτύου διάχυσης πληροφορίας σε ένα online κοινωνικό δίκτυο.

Η παρούσα εργασία έχει ως στόχο, λαμβάνοντας υπόψη τα δομικά και χρονικά χαρακτηριστικά (structural and temporal features) ενός κοινωνικού δικτύου, να εξάγει το δίκτυο διάχυσης πληροφορίας (diffusion network), το οποίο αναπαριστά σε ποιο τμήμα του δικτύου που μελετάται, κινήθηκε η πληροφορία. Συγκεκριμένα, αξιοποιώντας γράφους για την αναπαράσταση των σχετικών δικτύων, επιχειρείται αρχικά η καταγραφή της διάχυσης πληροφορίας μεταξύ των χρηστών του δικτύου χρησιμοποιώντας γνωστό μοντέλο διάχυσης πληροφορίας και στη συνέχεια ο σχεδιασμός ενός αλγορίθμου συμπερασμού του δικτύου διάχυσης μέσω της αποδοτικής καταγραφής της διάχυσης αυτής.

1.3 Οργάνωση κειμένου

Η συγκεκριμένη εργασία αποτελείται από 6 κεφάλαια.

Το πρώτο κεφάλαιο αποτελεί την παρουσίαση των πλατφορμών κοινωνικής δικτύωσης και των λόγων για τους οποίους υπάρχει η ανάγκη ανάλυσής τους και διερεύνησης των χαρακτηριστικών τους. Επίσης, γίνεται αναφορά στο επιστημονικό πεδίο στο οποίο ανήκει η ανάλυσή τους.

Στο δεύτερο κεφάλαιο παρουσιάζονται συνοπτικά βασικά θέματα σχετικά με το θεωρητικό υπόβαθρο της ανάλυσης κοινωνικών δικτύων καθώς και στοιχεία της θεωρίας Γραφημάτων τα οποία αξιοποιούνται στη συγκεκριμένη εργασία.

Στο τρίτο κεφάλαιο παρουσιάζεται το πρόβλημα του συμπερασμού του δικτύου διάχυσης πληροφορίας, όπως συναντάται στη βιβλιογραφία, καθώς και η σχέση αυτού με τα δομικά χαρακτηριστικά του δικτύου. Επιπρόσθετα, παρουσιάζονται και περιορισμοί σχετικά με τη λήψη δεδομένων για τη διάχυση της πληροφορίας στο δίκτυο.

Στο τέταρτο κεφάλαιο παρουσιάζεται το πρόβλημα του συμπερασμού του δικτύου διάχυσης πληροφορίας, όπως αντιμετωπίστηκε σε αυτήν την εργασία. Συγκεκριμένα, παρουσιάζονται αναλυτικά οι αλγόριθμοι διάχυσης πληροφορίας και συμπερασμού του δικτύου διάχυσης, καθώς και η μέθοδος που χρησιμοποιήθηκε για την κατηγοριοποίηση των χρηστών και το συμπερασμό του δικτύου διάχυσης πληροφορίας.

Στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα μέσω της αξιολόγησης των οποίων προκύπτουν και τα συμπεράσματα σχετικά με την αποδοτικότητα των αλγορίθμων που χρησιμοποιήθηκαν.

Το έκτο κεφάλαιο αποτελεί τον επίλογο της εργασίας στον οποίο παρουσιάζεται η σύνοψη των αποτελεσμάτων καθώς και μελλοντικές επεκτάσεις της.

2

Θεωρητικό υπόβαθρο

2.1 Βασικά στοιχεία θεωρίας γραφημάτων

Για την αναπαράσταση κοινωνικών δικτύων αξιοποιείται η Θεωρία Γραφημάτων, βασικές έννοιες της οποίας παρουσιάζονται ακολούθως.

2.1.1 Βασικές Έννοιες

Ένας γράφος ή ένα γράφημα G(V,E) είναι ένα σύνολο από κορυφές (κόμβους) που ενώνονται μεταξύ τους με γραμμές (ακμές), οι οποίες αναπαριστούν τη σχέση που έχουν οι κόμβοι μεταξύ τους[2].

Ένα γράφημα ονομάζεται κατευθυνόμενο (directed) αν οι ακμές του έχουν προσανατολισμό, δηλαδή αν η μία κορυφή δηλώνει την αρχή και η άλλη το τέλος της ακμής. Σε περίπτωση που δεν τηρείται αυτή η συνθήκη, το γράφημα είναι μη-κατευθυνόμενο (undirected)[3].

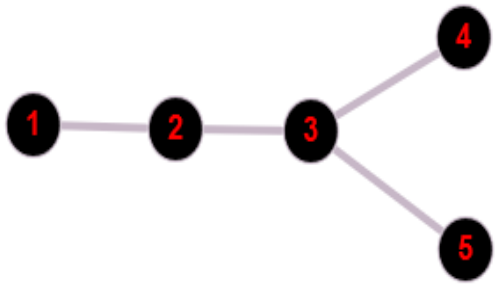
Δύο κορυφές είναι γειτονικές αν υπάρχει ακμή που ξεκινάει από τη μία και καταλήγει στην άλλη. Βαθμός μιας κορυφής σε ένα μη κατευθυνόμενο γράφο ονομάζεται ο αριθμός των γειτονικών της κορυφών.

Πολλές φορές στη θεωρία γραφημάτων οι ακμές χρησιμοποιούνται για την αναπαράσταση φυσικών ποσοτήτων όπως, για παράδειγμα, η απόσταση. Αν χρειάζεται για την ανάλυση του προβλήματος αυτή η φυσική ποσότητα, τότε ανατίθεται στην ακμή ένας αριθμός που την προσδιορίζει. Αυτός ο αριθμός ονομάζεται βάρος της ακμής. Ένας γράφος του οποίου οι ακμές έχουν βάρη ονομάζεται γράφος με βάρη (weighted graph) [4].

Ένας γράφος και οι συνδέσεις ανάμεσα στις κορυφές του μπορούν να αναπαρασταθούν αλγεβρικά μέσω ενός πίνακα γειτνίασης (adjacency matrix).

Ο πίνακας γειτνίασης Α είναι ένας τετραγωνικός πίνακας διαστάσεων nxn, όπου n το πλήθος των κορυφών του γράφου και κάθε στοιχείο του Aij δηλώνει την ύπαρξη ή όχι της ακμής ανάμεσα στις κορυφές i,j. Συγκεκριμένα, για τους μη-κατευθυνόμενους, χωρίς βάρη, γράφους, ο πίνακας γειτνίασης έχει "1" στις θέσεις που αντιστοιχούν σε ακμές που ανήκουν στο γράφο και "0" στις υπόλοιπες [5].

Ακολούθως, παρουσιάζεται ο μη-κατευθυνόμενος γράφος του προηγούμενου παραδείγματος και ο πίνακας γειτνίασής του.



	1	2	3	4	5
1	0	1	0	0	0
2	1	0	1	0	0
3	0	1	0	1	1
4	0	0	1	0	0
5	0	0	1	0	0

Κλίκα (clique) ενός μη-κατευθυνόμενου γραφήματος είναι ένας υπογράφος του οποίου οι κορυφές είναι όλες μεταξύ τους γειτονικές[6].

2.1.2 Περίπατοι και Μονοπάτια

Ένας περίπατος μήκους a στο γράφημα είναι μία ακολουθία
$\Pi = (u_1,v_1,u_2,v_2,...v_{i-1},u_i)$ από κορυφές (u) και ακμές (v) που αρχίζει και τελειώνει με κορυφή ώστε η ακμή v_j να προσπίπτει στις κορυφές u_j και u_{j+1}, για $1 \leq j < i$[7].

Μονοπάτι είναι ένας περίπατος στον οποίο κάθε κορυφή εμφανίζεται το πολύ μία φορά. Αν η αρχική και η τελική κορυφή ταυτίζονται, τότε λέμε ότι έχουμε ένα κύκλο[8].
Ελάχιστο ή συντομότερο (shortest path) ανάμεσα σε δύο κορυφές καλείται ένα μονοπάτι όταν είναι αυτό με το μικρότερο μήκος[9].

Αν για κάθε ζεύγος κορυφών ενός γράφου υπάρχει μονοπατι που τους συνδέει, τότε ο γράφος χαρακτηρίζεται συνδεδεμένος (connected). Αν δεν ισχύει αυτό, τότε ο γράφος χωρίζεται σε συνεκτικές συνιστώσες (connected components), οι οποίες αποτελούνται από συνδεδεμένους υπογράφους του G [10].

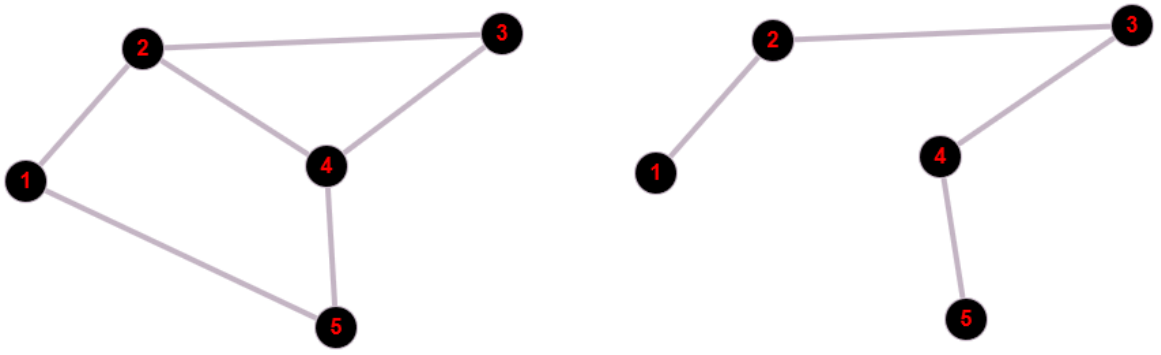
Στους συνδεδεμένους γράφους, αν η αφαίρεση μιας ακμής διαιρεί το γράφο σε τουλάχιστον δύο συνεκτικές συνιστώσες, αυτή ονομάζεται γέφυρα (bridge) [11].

2.1.3 Δέντρα

Ως Δέντρο (tree) ορίζεται ένας γράφος έστω G = (V,E) όπου V οι κορυφές και E οι ακμές, για τον οποίο ισχύει τουλάχιστον ένα από τα παρακάτω.

1. Ο G είναι συνεκτικός χωρίς κύκλους.
2. Ο G είναι συνεκτικός και ισχύει ότι |E| = |V| - 1
3. Ο G δεν έχει κύκλους και |E| = |V|-1
4. Ο G είναι συνεκτικός και κάθε ακμή του είναι γέφυρα.
5. Για κάθε ζεύγος κορυφών του G υπάρχει μοναδικό μονοπάτι που τις συνδέει.
6. Ο G δεν έχει κύκλους, και για κάθε καινούρια ακμή e, ο γράφος G έχει ακριβώς ένα κύκλο[12].

Ως δέντρο επικάλυψης (spanning tree) ενός γράφου G = (V, E) ορίζουμε ένα υπογράφο του G ο οποίος περιλαμβάνει όλες τις κορυφές του G και είναι δέντρο, δηλαδή ικανοποιεί τουλάχιστον ένα από τα παραπάνω κριτήρια [13].



2.1.4 Κυρίαρχα και ανεξάρτητα σύνολα

Στη θεωρία γραφημάτων, ως independent set ορίζεται το σύνολο κάποιων κορυφών ενός δικτύου από τις οποίες καμία δε γειτονεύει με άλλη.

Το μεγιστικό ανεξάρτητο σύνολο (maximal independent set) ενός δικτύου είναι το independent set το οποίο δεν αποτελεί υποσύνολο άλλου independent set[14].

Ένα άλλο αντικείμενο μελέτης της θεωρίας γραφημάτων είναι τα κυρίαρχα σύνολα (dominating sets). Dominating set ενός γράφου G είναι ένα σύνολο D, για το οποίο ισχύει ότι κάθε κορυφή του G που δεν είναι στο D είναι γειτονική μιας τουλάχιστον κορυφής η οποία ανήκει στο D[15].

Το πρόβλημα εύρεσης ενός minimum dominating set αφορά την εύρεση του μικρότερου από όλα τα πιθανά dominating sets ενός γράφου.

Αξίζει να αναφερθεί πως ένα ανεξάρτητο σύνολο είναι και κυρίαρχο σύνολο αν και μόνο αν είναι μεγιστικό ανεξάρτητο σύνολο.

2.2 Ανάλυση κοινωνικών δικτύων

Για την ανάλυση κοινωνικών δικτύων αξιοποιούνται οι ορισμοί της θεωρίας γραφημάτων που αναλύθηκαν ανωτέρω, καθώς και μετρικές οι οποίες παρέχουν μία ποσοτική περιγραφή της δομής ενός δικτύου τόσο σε επίπεδο κόμβου, όσο και συνολικά σε επίπεδο δικτύου.

2.2.1 Κατανομή βαθμού κόμβου

Βαθμός (degree) ενός κόμβου ενός μη-κατευθυνόμενου γραφήματος χωρίς βάρη είναι ο αριθμός των γειτόνων του.

Κατανομή του βαθμού κόμβων (degree distribution) P(m) ενός δικτύου είναι ο αριθμός των κόμβων του δικτύου με βαθμό m. Κατά συνέπεια, αν ένα δίκτυο αποτελείται από n κόμβους και n_m από αυτούς έχουν βαθμό m, τότε το δίκτυο έχει degree distribution $P(m) = n_m/n$.

Σε επίπεδο δικτύου, η κατανομή βαθμών κόμβων μπορεί να προκύπτει είτε ντετερμινιστικά, είτε πιθανοτικά. Συγκεκριμένα, σε δίκτυα στα οποία οι σχέσεις γειτνίασης δε μεταβάλλονται - όπως για παράδειγμα ένας χάρτης συγκοινωνιών - κατανομή βαθμών κόμβων αποτελεί το σύνολο των βαθμών των κόμβων του δικτύου. Αντίθετα, σε δίκτυα που οι σχέσεις ανάμεσα στους κόμβους - γείτονες μεταβάλλονται χρονικά ή με άλλο τρόπο, τότε η κατανομή βαθμών κόμβων του δικτύου προκύπτει στοχαστικά. Σε αυτήν την περίπτωση, η κατανομή βαθμών κόμβων P(m) είναι η πιθανότητα ένας κόμβος να έχει m αριθμό γειτόνων [16].

2.2.2 Μέσο μήκος μονοπατιού

Χρήσιμη μετρική σχετιζόμενη με το μήκος των μονοπατιών ενός γραφήματος είναι το μέσο μήκος μονοπατιού (average path length) το οποίο είναι ο μέσος αριθμός βημάτων που περιλαμβάνονται στα shortest paths ανάμεσα σε κάθε πιθανό ζευγάρι κόμβων του γραφήματος.
Επομένως, για τον υπολογισμό του μέσου μήκους μονοπατιού σε ένα δίκτυο απαιτείται ο προσδιορισμός όλων των πιθανών ζευγών κόμβων του δικτύου και στη συνέχεια ο

υπολογισμός του μέσου όρου των μηκών των μικρότερων μονοπατιών ανάμεσα σε όλα αυτά τα πιθανά ζεύγη.

2.2.3 Συντελεστής ομαδοποίησης

Επίσης, η μετρική του συντελεστή ομαδοποίησης (clustering coefficient) αφορά την τάση των κόμβων ενός δικτύου να συνδέονται ώστε να δημιουργούν ομάδες κόμβων μέσα στο δίκτυο. Στα πλαίσια της συγκεκριμένης εργασίας έχουν ενδιαφέρον δύο μορφές του συντελεστή ομαδοποίησης, ο τοπικός(local) και ο μέσος συντελεστής ομαδοποίησης ενός δικτύου.
Ο τοπικός συντελεστής ομαδοποίησης ενός κόμβου ενός δικτύου εκτιμά το κατά πόσο οι συνδέσεις ανάμεσα στους γείτονες του οδηγούν ώστε η γειτονιά του να αποτελεί μία κλίκα (clique).
Δεδομένου ότι για κάθε κόμβο με k γείτονες ο μέγιστος αριθμός ακμών που θα μπορούσαν να υπάρχουν στη γειτονιά αυτή είναι k(k-1)/2, ο local clustering coefficient υπολογίζεται ως,

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

Ο μέσος όρος των συντελεστών ομαδοποίησης όλων των κορυφών ενός γράφου, αποτελεί το μέσο συντελεστή ομαδοποίησης ενός δικτύου[16].

2.2.4 Μετρικές κεντρικότητας κόμβων

Τρεις είναι οι κύριες μετρικές οι οποίες χρησιμοποιούνται για την εκτίμηση της κεντρικότητας ενός κόμβου.

Κεντρικότητα βαθμού κόμβου

Η κεντρικότητα βαθμού κόμβου (degree centrality), η οποία μετράται ως συνάρτηση του βαθμού κόμβου και αναφέρεται και ως κεντρικότητα σημείου στην ουσία δηλώνει την δυνατότητα ενός κόμβου να ελέγχει τη ροή της πληροφορίας στο δίκτυο μέσω της δημοτικότητάς του. Ενδεικτικά, ένας κόμβος με μεγαλύτερο βαθμό αναμένεται να ελέγχει μεγαλύτερο μέρος της ροής της πληροφορίας σε ένα δίκτυο, σε σχέση με έναν λιγότερο δημοφιλή κόμβο.

Όπως αναφέρθηκε, προκύπτει πως η κεντρικότητα βαθμού κόμβου μπορεί να υπολογιστεί ως μία συνάρτηση της τιμής του βαθμού που έχει ένας κόμβος. Η πιο απλή από αυτές τις πιθανές συναρτήσεις είναι η γραμμική συνάρτηση που εκμεταλλεύεται τα χαρακτηριστικά του πίνακα γειτνίασης. Δεδομένου του πίνακα γειτνίασης ενός δικτύου, εύλογα προκύπτει πως ο βαθμός ενός κόμβου είναι το άθροισμα των τιμών που υπάρχουν στη γραμμή/στήλη για τον εν λόγω κόμβο. Συγκεκριμένα η κεντρικότητα βαθμού ενός κόμβου k καθώς και η

κεντρικότητά του σε σχέση με την κεντρικότητα βαθμού κόμβου των υπόλοιπων κόμβων του δικτύου υπολογίζεται ως εξής:

$$C_D(k) = \sum_{i=1}^{n} a_{ik}$$

$$C'_D = \frac{\sum_{i=1}^{n} a_{ik}}{n-1}$$

Ωστόσο, για ένα κόμβο ο οποίος είναι απομονωμένος από τον κύριο όγκο του δικτύου, αλλά έχει πολλούς γείτονες θα προκύψει υψηλή κεντρικότητα με βάση αυτόν τον τρόπο υπολογισμού, κάτι που δεν ανταποκρίνεται στην πραγματικότητα. Το πρόβλημα που γίνεται αντιληπτό σχετικά με την αξιοπιστία της μετρικής του βαθμού κόμβου, είναι πως η αποτελεσματικότητά της δεν είναι εξασφαλισμένη καθώς εξαρτάται από τα δομικά χαρακτηριστικά του δικτύου. Απαιτείται μια πιο ακριβής μετρική κεντρικότητας, η οποία σχετίζεται με τον έλεγχο της πληροφορίας, και λαμβάνει υπόψιν τη σχέση που έχει ένας κόμβος με τη διαδικασία της διάδοσης της πληροφορίας. Στη θεωρία γραφημάτων, ο πιο συνηθισμένος τρόπος μέτρησης των αποστάσεων είναι η μέτρηση βημάτων - hops. Επομένως, χρειάζεται μία μετρική κεντρικότητας η οποία θα θεωρεί κεντρικούς τους κόμβους που βρίσκονται σε μία θέση κεντρική ως προς το υπόλοιπο δίκτυο, σε αντίθεση με τους κόμβους που παρουσιάζονται απομονωμένοι από το υπόλοιπο δίκτυο [16].

Κεντρικότητα εγγύτητας

Ο ορισμός της κεντρικότητας εγγύτητας, βασίζεται στην απόσταση κάθε κόμβου του δικτύου από όλους τους υπόλοιπους. Δεδομένου του τρόπου μέτρησης των αποστάσεων που αναφέρθηκε παραπάνω, αλλά και των δομικών χαρακτηριστικών του δικτύου, η απόσταση ανάμεσα σε δύο κόμβους είναι το μήκος του μικρότερου μονοπατιού (shortest path) από έναν κόμβο σε έναν άλλον. Συνεπώς, η κεντρικότητα εγγύτητας ενός κόμβου, είναι ο αντίστροφος του αθροίσματος των shortest paths από αυτόν τον κόμβο προς κάθε άλλο κόμβο του δικτύου. Θεωρώντας ως d(i,k) την απόσταση ανάμεσα στους κόμβους i,k η κεντρικότα εγγύτητας του κόμβου k αλγεβρικά υπολογίζεται ως εξής:

$$C_P(k) = \left(\sum_{i=1}^{n} d(i,k) \right)^{-1}$$

Όπως αναφέρθηκε, η κεντρικότητα εγγύτητας εξαρτάται από το πόσο κοντά βρίσκεται ένας κόμβος στους υπόλοιπους κόμβους του δικτύου. Έτσι, ένας κόμβος που είναι σχετικά κοντά με τους περισσότερους κόμβους του δικτύου θα έχει υψηλή κεντρικότητα εγγύτητας, σε αντίθεση με έναν απομονωμένο κόμβο - με πολλούς ή λίγους γείτονες - ο οποίος θα έχει μικρή κεντρικότητα εγγύτητας. Ενδιαφέρον παρουσιάζει η σχέση ανάμεσα σε κεντρικότητα εγγύτητας και κόστος επικοινωνίας, δεδομένου πως τα δύο αυτά σχετίζονται με την απόσταση ανάμεσα σε πομπό και δέκτη.

Γίνεται λοιπόν αντιληπτό πως για τον υπολογισμό της κεντρικότητας εγγύτητας παίζει καθοριστικό ρόλο η δομή του δικτύου και ο αριθμός των κόμβων οι οποίοι συντελούν το δίκτυο [16].

Κεντρικότητα ενδιαμεσικότητας

Η κεντρικότητα ενδιαμεσικότητας ασχολείται με την εύρεση του αριθμού των φορών που ένας κόμβος αποτελεί μέρος ενός shortest path ανάμεσα σε δύο άλλους κόμβους. Ουσιαστικά θεωρείται πως ένας κόμβος που βρίσκεται με μεγάλη συχνότητα σε shortest paths ανάμεσα σε άλλους, έχει μεγαλύτερη δυνατότητα ελέγχου της ροής πληροφορίας μέσα στο δίκτυο και άρα θα πρέπει να θεωρείται πιο κεντρικός από τους άλλους. Επομένως, σε αυτήν την περίπτωση είναι η δυναμική ενός κόμβου να ελέγξει τη ροή της πληροφορίας αυτή η οποία καθορίζει την κεντρικότητά του.

Ο ορισμός της μετρικής ενδιαμεσικότητας είναι απλός όταν υπάρχει μόνο ένα shortest path ανάμεσα σε κάθε ζευγάρι κόμβων. Ωστόσο, όταν υπάρχουν περισσότερα του ενός shortest paths ανάμεσα σε ένα ζευγάρι κόμβων (εναλλακτικά μονοπάτια, όλα όμως με ίδιο μήκος) τότε ένας κόμβος που βρίσκεται σε κάποια αλλά όχι σε όλα αυτά τα shortest paths έχει περιορισμένη δυνατότητα ελέγχου της ροής της πληροφορίας.

Η μερική ενδιαμεσικότητα (partial betweenness) ενός κόμβου n σε σχέση με ένα ζευγάρι κόμβων (a,b) όπου οι τρεις αυτοί κόμβοι δεν ταυτίζονται, ορίζεται ως εξής:

- αν δεν υπάρχει μονοπάτι ανάμεσα στον a και στον b, τότε η μερική ενδιαμεσικότητα $p_{ab}(n) = 0$

- αν υπάρχει μονοπάτι ανάμεσα στον a και στον b, τότε η μερική ενδιαμεσικότητα $p_{ab}(n) = k_{ab} / k_{ab}(n)$, όπου k_{ab} είναι ο αριθμός των μονοπατιών που συνδέουν a και b και $k_{ab}(n)$ είναι ο αριθμός των μονοπατιών που συνδέουν a και b, τα οποία περιλαμβάνουν και τον κόμβο n. Αν το μονοπάτι που περιλαμβάνει τον n είναι το μοναδικό, τότε $p_{ab}(n) = 1$.

Ο υπολογισμός της συνολικής κεντρικότητας ενδιαμεσικότητας ενός κόμβου n προκύπτει αθροίζοντας όλες τις μερικές ενδιαμεσικότητες για όλα τα ζεύγη κορυφών στο γράφο:

$$C_B(k) = \sum_{i \neq j \neq k, i<j}^{n} \sum^{n} b_{ij}(k)$$

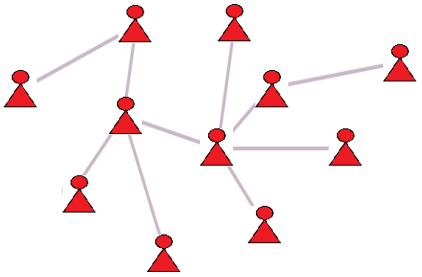
Το βασικό πρόβλημα του υπολογισμού της κεντρικότητας ενδιαμεσικότητας είναι πως χρειάζεται πρώτα η εύρεση όλων των shortest paths ενός γράφου - δικτύου, κάτι που είναι πολύ απαιτητικό ειδικά για δίκτυα μεγάλης κλίμακας [16].

2.3 Τύποι Σύνθετων Δικτύων και Μοντέλα Αναπαράστασης

Ως δίκτυο ορίζεται οποιοδήποτε σύνολο αντικειμένων εκ των οποίων κάποια συνδέονται με κάποιο τρόπο μεταξύ τους σε ζευγάρια, ενώ ως κοινωνικό δίκτυο (SN) ορίζεται το σύνολο των κοινωνικών σχέσεων ανάμεσα σε ανθρώπους.
Στο εξής, το Online Κοινωνικό Δίκτυο ή αλλιώς Online Social Network θα αναγράφεται ως OSN. Συμπεραίνεται εύκολα πως το OSN είναι ένα λογικό και όχι ένα φυσικό δίκτυο, από την άποψη πως οι συνδέσεις ανάμεσα στους λογαριασμούς που οι άνθρωποι διατηρούν σε αυτά δεν έχουν κάποια φυσική υπόσταση.
Με την έννοια του "φίλου" σε ένα OSN εννοείται η πρόσβαση ενός χρήστη στις δημοσιεύσεις ενός άλλου, καθώς και η δυνατότητα να ανταλλάσσουν προσωπικά μηνύματα.

Στη συγκεκριμένη εργασία, για την αναπαράσταση των OSNs χρησιμοποιούνται γράφοι, των οποίων οι κόμβοι αναπαριστούν τους χρήστες του OSN ενώ οι ακμές αναπαριστούν την πιθανή τους σύνδεση με την έννοια του φίλου στο OSN.

Επομένως, κόμβοι με μεγάλο βαθμό αναπαριστούν χρήστες του δικτύου με μεγάλη λίστα φίλων. Το αντίθετο ισχύει για κόμβους με μικρό βαθμό.



Υπάρχουν συγκεκριμένοι τύποι μαθηματικών γραφημάτων, οι οποίοι εξυπηρετούν στην αναπαράσταση OSNs. Οι δύο στους οποίους θεωρείται σημαντικό να γίνει εκτενής αναφορά είναι οι εξής: [16]

<u>Δίκτυα μικρού κόσμου (small world networks)</u>

Σε ένα δίκτυο μικρού κόσμου (SW Network) οι περισσότεροι κόμβοι δε γειτονεύουν μεταξύ τους. Ωστόσο, οι γείτονες κάθε κόμβου είναι πολύ πιθανό να είναι γείτονες μεταξύ τους, προκειμένου κάθε κόμβος να είναι προσβάσιμος με μικρό αριθμό βημάτων. Σε περίπτωση που ένα SW Network αποτελείται από N κόμβους, τότε η απόσταση μεταξύ δύο κόμβων του είναι ανάλογη της ποσότητας logN.

Αναλύοντας τα δίκτυα μικρού κόσμου με βάση κάποιες εκ των προαναφερθεισών μετρικών, μπορούν να εξαχθούν σημαντικά συμπεράσματα.
Λόγω του υψηλού αριθμού συνδέσεων ανάμεσα στους γείτονες κάθε κόμβου, γίνεται αντιληπτό πως υπάρχει η τάση να δημιουργούνται κλίκες (cliques) καθώς παρατηρούνται υψηλές τιμές του συντελεστή ομαδοποίησης (clustering coefficient). Επίσης, το μικρό μέσο

μήκος μονοπατιού προκύπτει ως άμεση συνέπεια από τα κύρια χαρακτηριστικά των SW Networks[17].

Η εφαρμογή των small world networks έχει οδηγήσει στην καθιέρωση πρόσθετων μετρικών οι οποίες εξετάζουν το κατά πόσο ένα δίκτυο προσεγγίζει τη δομή ενός small world network:

- Μέσω της μετρικής του small-coefficient (σ) συγκρίνεται η τάση για ομαδοποίηση και το μήκος μονοπατιού ενός δικτύου με ένα παρόμοιο τυχαίο δίκτυο το οποίο έχει τον ίδιο μέσο βαθμό κόμβου. Συγκεκριμένα,

  $\sigma = (C/C\tau) / (L/L\tau)$.

  Αν $\sigma >> 1$ ($C >> C\tau$ και $L = L\tau$), τότε το δίκτυο είναι small world.

Η μετρική αυτή δεν είναι, ωστόσο, αρκετή, καθώς επηρεάζεται από το μέγεθος του δικτύου.

- Εναλλακτικός τρόπος εκτίμησης του πόσο παρεμφερές είναι ένα δίκτυο με ένα small world network είναι η σύγκριση του clustering του δικτύου με ένα παρεμφερές δίκτυο-πλέγμα, ενώ παράλληλα γίνεται σύγκριση του μέσου μήκους μονοπατιού με την αντίστοιχη μετρική για ένα τυχαίο δίκτυο. Σε αυτήν την περίπτωση,

  $\omega = (L\tau/L) - (C/C\pi)$,

  όπου L,C το μήκος μονοπατιού και ο συντελεστής ομαδοποίησης αντίστοιχα για το δίκτυο που ελέγχουμε, ενώ $C\pi$ είναι ο συντελεστής ομαδοποίησης του δικτύου-πλέγματος και $L\tau$ το μέσο μήκος μονοπατιού του τυχαίου δικτύου.

- Τελευταίος τρόπος, είναι ο small world index (SWI) ο οποίος με δεδομένα τα δίκτυα της προηγούμενης μεθόδου υπολογίζεται ως εξής:

  $SWI = [(L-L\pi) / (L\tau-L\pi)] * [(C-C\tau) / (C\pi-C\tau)]$.

Τέλος, κρίνεται χρήσιμο να παρουσιαστεί ένα μοντέλο κατασκευής small-world δικτύων τα οποία χρησιμοποιούνται και για αυτήν την εργασία.

➔ *Μοντέλο Watts-Strogatz (Watts-Strogatz model)*

Το μοντέλο Watts-Strogatz αποτελεί μία διαδικασία κατασκευής small world καθώς και άλλων τύπων τυχαίων γράφων ξεκινώντας με μια κανονική δομή πλέγματος. Συγκεκριμένα, με βάση αυτή τη διαδικασία λαμβάνεται ένα κανονικό δίκτυο - πλέγμα και εισάγονται διαδοχικά ακμές οι οποίες συνδέουν κόμβους οι οποίοι με βάση την αρχική δομή του δικτύου θεωρούνται απομακρυσμένοι από άποψη μέτρησης βημάτων - hops. Αυτές τις τυχαίες ακμές θα καλούνται στο εξής “shortcuts“ (συντομεύσεις/παρακάμψεις) .
Η αρχική δομή πλέγματος εξασφαλίζει υψηλή τιμή του συντελεστή ομαδοποίησης, ενώ η εισαγωγή ενός κατάλληλου αριθμού από shortcuts μειώνει περεταίρω το μέσο

μήκος μονοπατιού στο δίκτυο ώστε ο νέος γράφος να μπορεί να χαρακτηριστεί ως small world[16].

Δίκτυα ελεύθερης κλίμακας (scale free networks)

Σε ένα δίκτυο ελεύθερης κλίμακας (SF Network), η πλειοψηφία των κόμβων έχει μικρό βαθμό και οι κόμβοι με υψηλή κεντρικότητα είναι λίγοι.
Η κατανομή βαθμού κόμβου γίνεται με βάση το νόμο της δύναμης (power law), σύμφωνα με τον οποίο η μεταβολή μιας ποσότητας οδηγεί σε μια αναλογική μεταβολή μιας δεύτερης ποσότητας με την οποία σχετίζεται.

Ο τρόπος με τον οποίο δομείται το δίκτυο, έχει σαν αποτέλεσμα να υπάρχουν λίγοι κόμβοι με μεγάλο βαθμό - και άρα επιρροή - οι οποίοι αποτελούν ταυτόχρονα και πλεονέκτημα και μειονέκτημα για το δίκτυο. Συγκεκριμένα, διασφαλίζουν πως αν αφαιρεθούν κάποιοι κόμβοι από το δίκτυο δε θα εμποδιστεί η διάδοση της πληροφορίας. Ωστόσο, αν αφαιρεθούν αρκετοί από τους κόμβους με μεγάλο βαθμό, τότε το δίκτυο θα αποτελείται από διάφορα, μικρότερα υποδίκτυα, μη συνδεδεμένα μεταξύ τους.

Όσον αφορά τον συντελεστή ομαδοποίησης (clustering coefficient), παρατηρείται πως όσο μεγαλύτερος είναι ο βαθμός του κόμβου, τόσο μικρότερος θα είναι ο συντελεστής. Συνεπώς, προκύπτει πως οι κόμβοι με χαμηλό βαθμό συνδέονται με τους γείτονές τους σχηματίζοντας μικρούς υπογράφους, και στη συνέχεια λίγοι εξ αυτών συνδέονται με έναν κόμβο με μεγάλη κεντρικότητα [16].

Τέλος, κρίνεται χρήσιμο να παρουσιαστεί ένα μοντέλο κατασκευής scale-free δικτύων:

➔ *Μοντέλο Barabasi-Albert (Barabasi-Albert model)*

Για την κατασκευή ενός γραφήματος με τον μοντέλο Barabasi - Albert, θεωρείται πως ο χρόνος είναι κατανεμημένος σε επιμέρους χρονικές μονάδες. Το αρχικό δίκτυο αποτελείται από m0 κόμβους. Κάθε νεοεισαχθής κόμβος στο δίκτυο συνδέεται με m κατάλληλα επιλεγμένους κόμβους, ήδη παρόντες στο δίκτυο.
Σε κάθε χρονική μονάδα t πραγματοποιούνται τα εξής:

- ➢ Ένας νέος κόμβος εισάγεται στο δίκτυο και συνδέεται σε m υπάρχοντες κόμβους
- ➢ Κάθε ένας από τους m υπάρχοντες κόμβους επιλέγεται με βάση τον κανόνα της διασύνδεσης κατά προτίμηση (preferential attachment rule), για παράδειγμα ο κόμβος i με βαθμό $k_i(t)$ επιλέγεται με πιθανότητα

$$\Pi(k_i) = \frac{k_i(t)}{\sum_{\forall j} k_j(t)},$$

όπου το άθροισμα στον παρονομαστή εκτείνεται σε όλους τους κόμβους του δικτύου

- Ο αριθμός των κόμβων στο δίκτυο είναι ίσως με $N_t = m_0 + t$

3

Διάχυση πληροφορίας σε πλατφόρμες κοινωνικής δικτύωσης

3.1 Διάχυση πληροφορίας σε ένα Online Social Network

Ως διάχυση πληροφορίας σε ένα OSN ορίζουμε τη διαδικασία μέσω της οποίας η πληροφορία επικοινωνείται μέσα από συγκεκριμένους διαύλους ανάμεσα στους χρήστες του OSN[19].
Οι κύριοι παράγοντες οι οποίοι επηρεάζουν τη διάχυση της πληροφορίας σε ένα OSN είναι

- η δομή και τα χαρακτηριστικά του δικτύου (κατανομή βαθμού κόμβου, κεντρικότητα κ.α.)
- η ισχύς των δεσμών (συχνότητα αλληλεπιδράσεων, ισχύς επιρροής)
- τα πρόσκαιρα δυναμικά, δηλαδή πώς ο ρυθμός διάδοσης εξελίσσεται με το χρόνο[20,21]

Η ερευνητική δραστηριότητα στη διάχυση πληροφορίας ασχολείται με τρεις κύριες κατηγορίες προβλημάτων. Την ανίχνευση δημοφιλών θεμάτων, το συμπερασμό ή/και την πρόβλεψη του δικτύου διάχυσης πληροφορίας και τέλος την αναγνώριση των πιο σημαντικών κόμβων στη διαδικασία της διάδοσης της πληροφορίας, αντικείμενο στενά συνδεδεμένο με το πρόβλημα του υπολογισμού και της μεγιστοποίησης της επιρροής [22].

3.1.1 Ανίχνευση και πρόβλεψη δημοφιλών θεμάτων

Για την ανίχνευση δημοφιλών θεμάτων στα OSNs, οι περισσότερες μέθοδοι στη βιβλιογραφία βασίζονται στην παρατήρηση της εμφάνισης συγκεκριμένων όρων σε ασυνήθιστη συχνότητα [23]. Στο [24] χρησιμοποιείται μία μηχανή καταστάσεων (state machine) προκειμένου να μοντελοποιηθούν οι χρόνοι άφιξης των εγγράφων μιας ροής ώστε να αναγνωριστούν οι ριπές, δηλαδή τα θέματα τα οποία αναπτύσσονται σε συχνότητα για μία χρονική περίοδο και μετά ατονούν, υποθέτοντας πάντα ότι αυτά τα έγγραφα ανήκουν στην ίδια θεματική ενότητα. Στο [25] αναπτύσσεται μία μέθοδος για την παρακολούθηση μονάδων πληροφορίας με τη μορφή των μικρών διακριτών εκφράσεων, οι οποίες αποκαλούνται “memes”. Η συγκεκριμένη μέθοδος ομαδοποιεί γραπτές παραλλαγές των συγκεκριμένων “memes” και εφαρμόζοντάς τη σε πραγματικά δεδομένα παρατηρείται πως ο κύκλος ενημέρωσης στις πλατφόρμες κοινωνικής δικτύωσης ακολουθεί ένα επίμονο χρονικό μοτίβο αυξανόμενης και μειούμενης δημοφιλίας γύρω από το σημείο της μέγιστης τιμής αυτής. Μία προσέγγιση η οποία επιχειρεί να προβλέψει τα θέματα μεγάλης δημοτικότητας στο άμεσο μέλλον αναπτύσσεται στο [26]. Ένας χρηματιστηριακός δείκτης πρόθεσης γνωστός και ως MACD (Moving Average Convergence Divergence) προσαρμόζεται στο να αναγνωρίζει ριπές από θεματικές ενότητες ορισμένες από λέξεις-κλειδιά στις πλατφόρμες κοινωνικής δικτύωσης. Το αντικείμενο του βελτιωμένου δείκτη MACD είναι να μετατρέψει δύο trend-following δείκτες, ένα μικρό και ένα μεγάλο χρονικό διάστημα κινούμενου μέσου από λέξεις-κλειδιά, σε έναν δυναμικό ταλαντωτή. Η δυναμική της κάθε τάσης υπολογίζεται αφαιρώντας το μεγάλο από το μικρό χρονικό διάστημα κινούμενου μέσου. Όταν η αξία της δυναμικής της τάσης αλλάζει από αρνητική

σε θετική, τότε η θεματική ενότητα γίνεται δημοφιλής. Αντίθετα, η δημοτικότητά της μειώνεται όταν η τιμή της δυναμικής της γίνεται αρνητική.

3.1.2 Συμπερασμός και πρόβλεψη δικτύων επιρροής

Στη δεύτερη κατηγορία μεθόδων ανήκουν αυτές της εύρεσης, μέσω πρόβλεψης ή συμπερασμού, ποιος χρήστης επηρέασε/μόλυνε ποιον, δηλαδή τα μονοπάτια μέσα από τα οποία διαδίδεται η πληροφορία. Τα μοντέλα διάδοσης πληροφορίας μπορούν να κατηγοριοποιηθούν σε επεξηγηματικά και πρόβλεψης.
Ο στόχος των επεξηγηματικών μοντέλων είναι ο συμπερασμός του δικτύου το οποίο χρησιμοποιήθηκε για τη διάδοση της πληροφορίας, δοθείσων των χρονικών στιγμών κατά τις οποίες οι χρήστες έλαβαν ένα μέρος της πληροφορίας. Οι συσχετισμοί ανάμεσα στις χρονικές στιγμές "μόλυνσης" των χρηστών μελετώνται στο [27], ώστε να συμπεραθεί το υπάρχον δίκτυο στο οποίο διαδίδονται οι "μολύνσεις". Υποθέτοντας μία στατική τοπολογία δικτύου και πως οι ενεργοί κόμβοι μεταδίδουν τη "μόλυνση" σε κάθε γείτονά τους ανεξάρτητα με κάποια συγκεκριμένη πιθανότητα, η διάδοση της πληροφορίας ανάμεσα σε δύο κόμβους μειώνεται με την αύξηση της διαφοράς των χρόνων ενεργοποίησής τους. Για την εύρεση της ροής διάδοσης που προσομοιάζει στο μέγιστο τα πραγματικά δεδομένα, χρησιμοποιείται ένας αλγόριθμος βασισμένος στη βελτιστοποίηση της υπομετρικής συνάρτησης (submodular function optimization). Μία προέκταση του συγκεκριμένου αλγορίθμου, η οποία λαμβάνει ως δεδομένα δυναμικα, χρονικά μεταβαλλόμενα δίκτυα παρουσιάζεται στο [28], όπου η διαδικασία της διάχυσης πληροφορίας μοντελοποιείται ως ένα χωρικά ανεξάρτητο δίκτυο από συνεχείς, υπό όρους ανεξάρτητες χρονικές διαδικασίες οι οποίες πραγματοποιούνται σε διαφορετικούς ρυθμούς. Η πιθανότητα ένας κόμβος να "μολύνει" έναν άλλον κόμβο σε μία συγκεκριμένη χρονικη στιγμή μοντελοποιείται μέσω μιας συνάρτησης πυκνότητας πιθανότητας, η οποία εξαρτάται από τις ανα ζεύγη χρονικές στιγμές ενεργοποίησης και τους ρυθμούς διάδοσης. Η συγκεκριμένη μέθοδος αναγνωρίζει τη δομή του δικτύου και τους ρυθμούς διάδοσης πληροφορίας ανάμεσα σε ζεύγη κόμβων μέσω της διατύπωσης ενός προβλήματος μέγιστης πιθανοφάνειας το οποίο επιλύεται με την στοχαστική κλίση κλίσης (stochastic gradient descent). Στο [29], το πρόβλημα του συμπερασμού της διάδοσης πληροφορίας αντιμετωπίζεται ως ένα πρόβλημα συστάσεων. Δεδομένης της διάδοσης πληροφορίας, που μοντελοποιείται ως μία ακολουθία από πλειάδες(user id, χρονική στιγμή μόλυνσης) αλλά και από ιστορικά δεδομένα των χρηστών, εξάγονται τα συμπεριφορικά χαρακτηριστικά των χρηστών τα οποία είναι, η σχέση τους με τις μελετώμενες θεματικές ενότητες αλλά και γραπτά χαρακτηριστικά της μεταξύ τους επικοινωνίας. Θεωρώντας τη διαδιδόμενη πληροφορία ως ένα αντικείμενο το οποίο μπορεί να προταθεί, τα χαρακτηριστικά που περιγράφησαν ανωτέρω αξιοποιούνται από ένα αναδρομικό νευρωνικό δίκτυο το οποίο επιλύει το πρόβλημα των προτάσεων και συμπεραίνει τη σχέση διάδοσης ανάμεσα στους χρήστες.
Με βάση το γεγονός ότι τα δεδομένα σχετικά με τη διασπορά που παρατηρείται σε ένα OSN είναι περιορισμένα λόγω των περιορισμών σχετικά με την ιδιωτικότητα των χρηστών, ένα επεξηγηματικό μοντέλο διάχυσης πληροφορίας παρουσιάζεται στην παράγραφο 3.2. Το πρόβλημα της μερικώς παρατηρήσιμης διάδοσης πληροφορίας αντιμετωπίζεται με την αξιοποίηση των δομικών χαρακτηριστικών ενός OSN προκειμένου να επιλεγεί το μικρότερο δυνατό σύνολο από κόμβους που χρειάζεται να χρησιμοποιηθούν για παρακολούθηση της ροής της πληροφορίας στο δίκτυο (π.χ. χρήστες με δημόσιο προφίλ που είναι δημοφιλείς στο δίκτυο), προκειμένου να συμπεραθεί η διαφορετική πληροφορία ως προς το περιεχόμενο και τη δυναμική διάδοσης που διαδίδονται την ίδια χρονική στιγμή, αλλά

ανεξάρτητα, σε ένα δίκτυο. Για την ιχνηλάτηση των μονοπατιών διάδοσης που προκύπτουν, χρησιμοποιούνται οι τεχνικές του χρωματισμού ακμών (edge coloring).

Σε αντίθεση με τα επεξηγηματικά μοντέλα, τα μοντέλα πρόβλεψης έχουν ως στόχο την πρόβλεψη των αποτελεσμάτων μιας συγκεκριμένης διαδικασίας διάχυσης η οποία βασίζεται στα χρονικά ή χωρικά χαρακτηριστικά του δικτύου [23]. Ευρέως γνωστά μοντέλα πρόβλεψης διάχυσης πληροφορίας σε γράφους είναι το πιθανοτικό μοντέλο Independent Cascade (IC) το οποίο είναι προσανατολισμένο γύρω από τον αποστολέα και χρησιμοποιείται και στη συγκεκριμένη εργασία και το Linear Threshold (LT) το οποίο είναι προσανατολισμένο γύρω από τον αποδέκτη. Στο μοντέλο IC ο κόμβος u τη χρονική στιγμή t γίνεται ενεργός (δηλαδή λαμβάνει την πληροφορία) και έχει μία μοναδική ευκαιρία να ενεργοποιήσει κάθε ανενεργό γείτονά του v τη χρονική στιγμή t+1 με πιθανότητα p_{uv}. Η διαδικασία συνεχίζεται μέχρι να μη μπορούν να ενεργοποιηθούν νέοι κόμβοι. Το LT μοντέλο είναι το πιο γνωστό από τα μοντέλα κατωφλιού τα οποία χρησιμοποιούνται στη μελέτη των δικτύων διάχυσης πληροφορίας. Στο συγκεκριμένο μοντέλο, μία τιμή κατωφλιού (ή ένα σύνολο από τιμές) ορίζουν το εύρος των τιμών στο οποίο η συμπεριφορα που προβλέφθηκε από το μοντέλο αλλάζει σημαντικά. Σε κάθε ακμή (u,v) του δικτύου ανατίθεται ένα βάρος w_{uv}, και κάθε κόμβος v έχει ένα κατώφλι t_v. Ο κόμβος u ενεργοποιείται αν το κλάσμα των ενεργών του γειτόνων ξεπερνά το t_v [19].
Ένα άλλο επεξηγηματικό μοντέλο είναι αυτό των τυχαίων περιπάτων (Random Walks). Δοθέντος ενός γράφου και ενός κόμβου-αρχικού σημείου από τον οποίο ξεκινά η διάδοση της πληροφορίας, ένας από τους γείτονές του επιλέγεται τυχαία για να λειτουργήσει ως νέο αρχικό σημείο για μία τυχαία επιλογή γείτονά του. Η ακολουθία των σημείων που δημιουργούνται επαναλαμβάνοντας τη συγκεκριμένη τυχαία διαδικασία αναφέρεται ως τυχαίος περίπατος σε γράφο [30] . Οι τυχαίοι περίπατοι είναι από τους πιο βασικούς τύπους στοχαστικών διαδικασιών και χρησιμοποιούνται ευρέως για τη μοντελοποίηση της διάχυσης της πληροφορίας και των αλληλεπιδράσεων ανάμεσα στις οντότητες σε δίκτυα διαφόρων δομικών χαρακτηριστικών. Μπορούν να κατηγοριοποιηθούν σε τρεις κύριους τύπους:

- Τυχαίοι περίπατοι διακριτού χρόνου
- Κομβοκεντρικοί τυχαίοι περίπατοι συνεχούς χρόνου
- Ακμοκεντρικοί τυχαίοι περίπατοι συνεχούς χρόνου [31]

Μοντέλα διάδοσης πληροφορίας τα οποία δε βασίζονται στην ύπαρξη γράφου, σε αντίθεση με τα προαναφερθέντα, βασίζονται κυρίως στις επιδημιολογικές διάδοσης. Στη μετάδοση επιδημιών, οι χρήστες μολύνονται με έναν ιό ενώ άλλοι είναι ευάλωτοι σε αυτόν. Ο ιος μπορεί να μεταδίδεται από μολυσμένους σε ευάλωτους χρήστες, με τρόπο παρόμοιο με αυτόν που η πληροφορία μεταδίδεται από αποστολείς σε δέκτες. Επομένως, οι χρήστες κατηγοριοποιούνται σε συγκεκριμένες τάξεις και μελετάται η απόκλιση στον αριθμό των χρηστών κάθε τάξης, λόγω των μεταβάσεων από μία κατάσταση σε άλλη. Τα επιδημιολογικά μοντέλα εκφράζονται μέσω διαφορικών εξισώσεων και τα πιο αντιπροσωπευτικά ντετερμινιστικά είναι τα SI (Susceptible-Infected), SIS (Susceptible-Infected-Susceptible) και SIR (Susceptible-Infected-Removed) [31][32] . Στα SI, SIR [36], ευάλωτοι χρήστες (S) μεταβαίνουν στην κατάσταση infected (I) με μία σταθερή πιθανότητα. Στο μοντέλο SIS επιπρόσθετα θεωρείται πως οι χρήστες στην κατάσταση I μπορεί να μεταβούν στην κατάσταση S με μία δεδομένη πιθανότητα. Αντίθετα, στο SIR, οι χρήστες από την I μεταβαίνουν στην R.

3.1.3 Αναγνώριση χρηστών με επιρροή

Η κοινωνική επιρροή ορίζεται στο [33] ως εξής:
Δοθέντων δύο χρηστών u,v σε ένα κοινωνικό δίκτυο, ο u επηρεάζει ή ασκεί δύναμη στον v όταν άμεσα ή έμμεσα, μπορεί να αλλάξει την άποψη του v.

Από την άποψη της διάχυσης πληροφορίας, η επιρροή μπορεί να οριστεί ως η ισχύς που έχει κάθε χρήστης στη διαδικασία της διάδοσης πληροφορίας. Η εύρεση των χρηστών με τη μεγαλύτερη επιρροή στους γείτονές τους είναι πολύ σημαντική καθώς εξασφαλίζει αποτελεσματικότητα στη μετάδοση της πληροφορίας, αλλά και στις συστάσεις. Οι πιο σημαντικές μετρικές για την αξιολόγηση της επιρροής κάθε χρήστη βασίζεται στη δομή του δικτύου, στις αλληλεπιδράσεις των χρηστών και στα χαρακτηριστικά αυτών. Οι μετρικές κεντρικότητας συμπεριλαμβανομένου του βαθμού κόμβου (degree), των κεντρικοτήτων εγγύτητας (closeness), ενδιαμεσικότητας (betweenness), eigenvector, Katz [34], θεωρούνται αξιόπιστες μετρικές αξιολόγησης με σκοπό το χαρακτηρισμό των χρηστών σε αναλογία με τη δομή του δικτύου. Για παράδειγμα, οι χρήστες με επιρροή σε ένα σύστημα συστάσεων για ένα OSN μπορούν να θεωρηθούν αυτοί που παρουσιάζουν μεγάλη ομοιότητα με ένα σημαντικό μέρος των χρηστών, έχοντας ταυτόχρονα μεγάλο αριθμό ακμών που ξεκινούν από αυτούς [35].

Το πρόβλημα της επιλογής των k χρηστών με τη μεγαλύτερη επιρροή σε ένα δίκτυο είναι γνωστό ως μεγιστοποίηση επιρροής και είχε παρουσιαστεί αρχικά στο [36] ως μια αλγοριθμική τεχνική για viral marketing. Στη συνέχεια μοντελοποιήθηκε στο [37] ως ένα πρόβλημα διακριτής βελτιστοποίησης το οποίο αποδείχθηκε NP-Hard. Με τη χρησιμοποίηση ενός πλαισίου ανάλυσης το οποίο βασίζεται σε συναρτήσεις υποδιαμορφωμένου συνόλου, παρουσιάστηκε μία άπληστη προσεγγιστική στρατηγική. Εκτός αυτής της στρατηγικής πολλοί ευριστικοί και υβριδικοί αλγόριθμοι έχουν παρουσιαστεί για την επίλυση του προβλήματος μέγιστης επιρροής σε στατικά και δυναμικά δίκτυα [38,39].

3.1.4 Μοντέλα διάδοσης πληροφορίας

Υπάρχει πληθώρα αλγορίθμων - μοντέλων διάδοσης πληροφορίας μέσω των οποίων μπορεί να επιτευχθεί η διάχυση της πληροφορίας σε σημαντικό βαθμό εντός ενός δικτύου. Στα πλαίσια της συγκεκριμένης εργασίας χρησιμοποιήθηκε το μοντέλο Independent Cascade. Κρίνεται επομένως, σκόπιμο να παρουσιαστεί ο αλγόριθμος αυτού του μοντέλου διάδοσης καθώς και αυτός του μοντέλου SIR, αφού το Independent Cascade αποτελεί μία γενίκευση του SIR.

<u>Μοντέλο SIR</u>

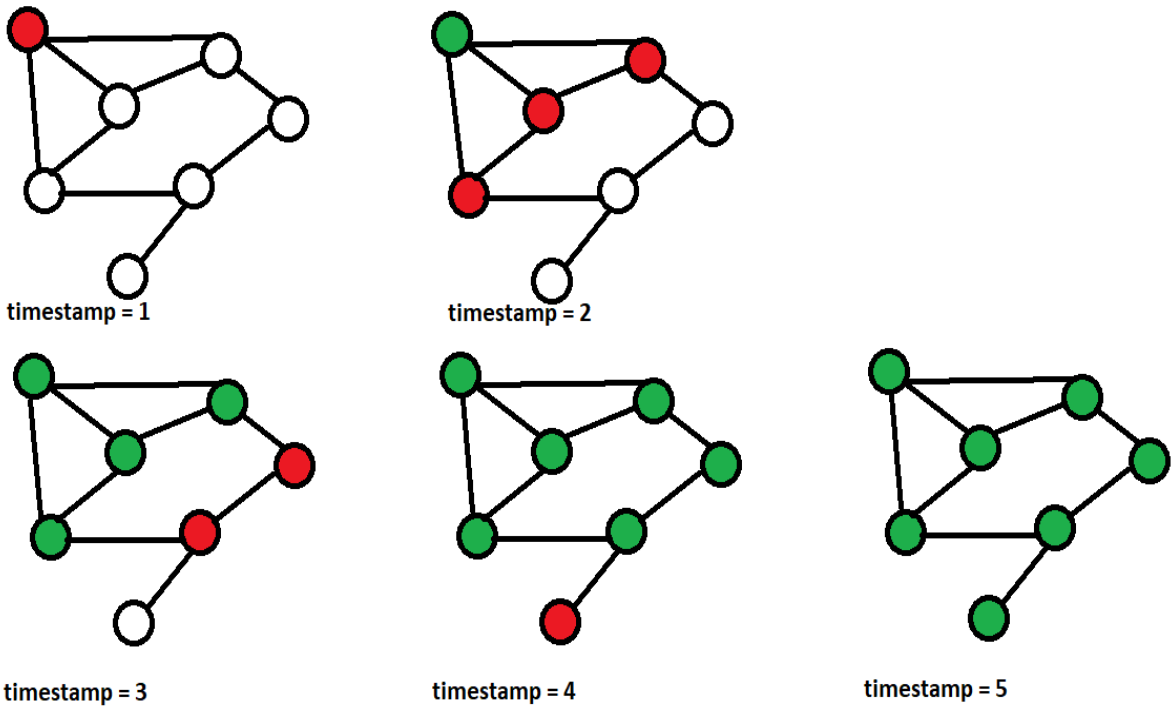
Τα αρχικά του συγκεκριμένου μοντέλου διάδοσης πληροφορίας προέρχονται από τις τρεις καταστάσεις στις οποίες μπορεί να οδηγήσει τους χρήστες ενός δικτύου, οι οποίες είναι οι Susceptible (Ευπαθής), Infected (Μολυσμένος), Recovered (Θεραπευμένος). Όλοι οι χρήστες του δικτύου κάθε χρονική στιγμή βρίσκονται σε μία εκ των τριών αυτών καταστάσεων. Πιο συγκεκριμένα, ως S χαρακτηρίζονται οι χρήστες που μπορούν να μολυνθούν, ως I οι χρήστες που είναι μολυσμένοι και ως R αυτοί που δεν έχουν πλέον τη δυνατότητα να μολυνθούν και να μολύνουν. Κάθε χρονική στιγμή, οι κόμβοι που

μολύνθηκαν την προηγούμενη χρονική στιγμή μπορούν να μολύνουν τους γείτονές τους που βρίσκονται στην κατάσταση S με πιθανότητα β. Μετά από αυτή τη χρονική στιγμή, οι κόμβοι που βρίσκονταν σε κατάσταση μολυσμένου (I) μεταβαίνουν σε κατάσταση θεραπευμένου (R) και δε μπορούν πλέον να μολύνουν ή να μολυνθούν εκ νέου.

Για κάθε κόμβο με βαθμό k και πιθανότητα βαθμού κόμβου P(k), η πιθανότητα μόλυνσης β υπολογίζεται ως

$$\left(\sum_k \frac{P(k)\cdot k\cdot(k-1)}{\langle k\rangle}\right)^{-1} = \beta$$

Ακολουθεί μία ενδεικτική απεικόνιση διάδοσης πληροφορίας με μοντέλο SIR. Σε κάθε χρονική στιγμή (timestamp), με κόκκινο απεικονίζονται οι μολυσμένοι κόμβοι, με πράσινο οι θεραπευμένοι και με λευκό αυτοί που ούτε έχουν μολυνθεί ούτε θεραπευθεί.



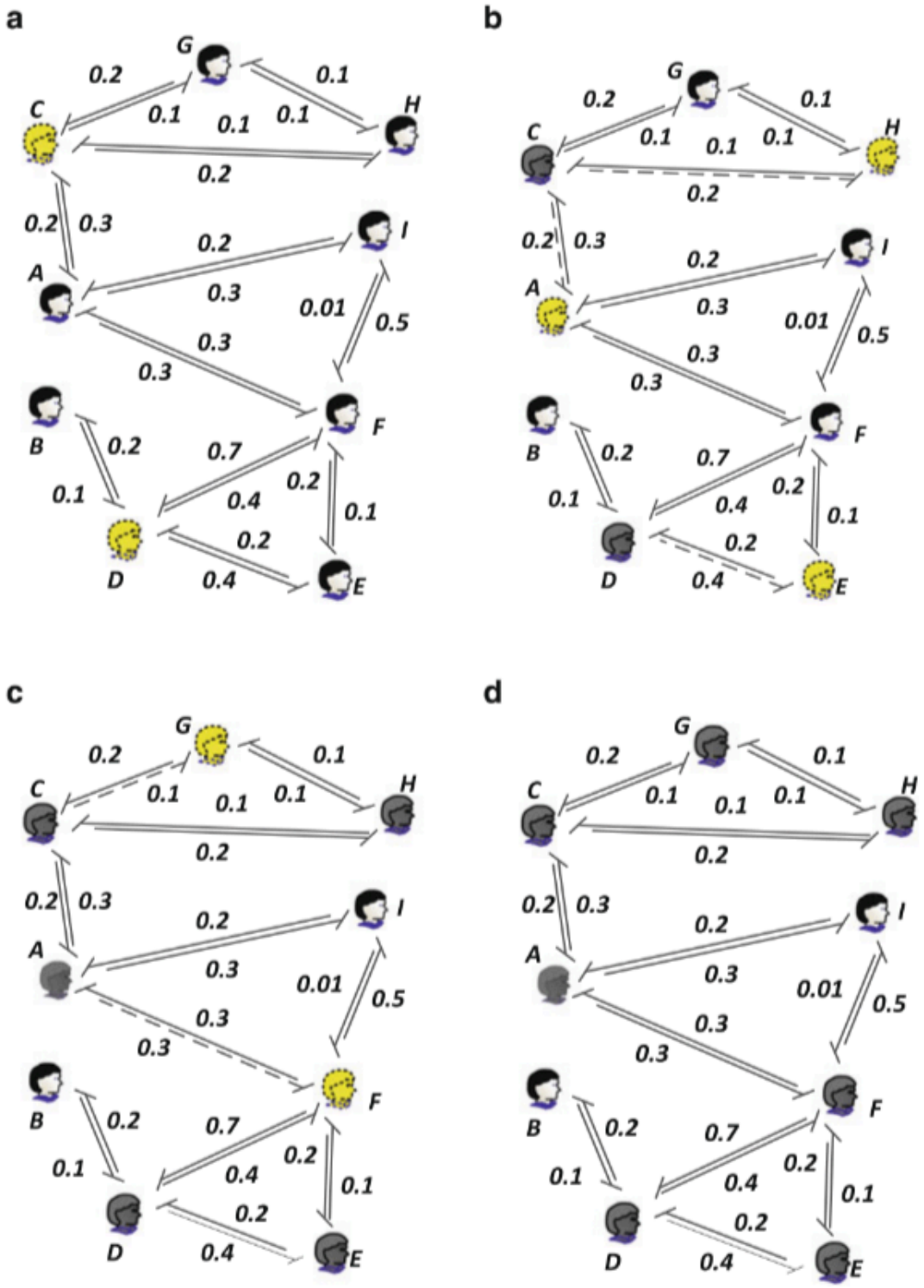
Μοντέλο Independent Cascade (IC)

Το μοντέλο Independent Cascade αποτελεί μία γενίκευση του μοντέλου SIR όπως περιγράφηκε ανωτέρω. Αντί για μία συνολική πιθανότητα μόλυνσης, υπάρχει διαφορετική πιθανότητα μόλυνσης σε συσχέτιση με κάθε ακμή του γράφου. Η πιθανότητα $P_{u,v}$ αποτελεί

την πιθανότητα να μολύνει ο κόμβος u τον κόμβο v. Η ανάθεση της συγκεκριμένης πιθανότητας μπορεί να γίνει με βάση τη συχνότητα αλληλεπιδράσεων, την απόσταση, ή παρελθοντικά δείγματα μόλυνσης. Κάθε κόμβος που μολύνεται έχει τη δυνατότητα να μολύνει έναν γείτονά του την επόμενη χρονική στιγμή με βάση την πιθανότητα που έχει ανατεθεί στην ακμή που τους συνδέει.

Επομένως, η διάδοση της πληροφορίας εντός ενός δικτύου με βάση το μοντέλο Independent Cascade ορίζεται ως εξής: Σε κάθε χρονική στιγμή t όπου I_{t-1} είναι το σύνολο των κόμβων οι οποίοι μολύνθηκαν την προηγούμενη χρονική στιγμή (t-1), κάθε κόμβος u που ανήκει στο I_{t-1} μολύνει τους μη μολυσμένους γείτονές του με πιθανότητα $P_{u,v}$.

Ένα παράδειγμα σχετικό με το συγκεκριμένο μοντέλο διάδοσης πληροφορίας παρουσιάζεται ακολούθως. Οι μολυσμένοι χρήστες παρουσιάζονται με κίτρινο. Τη χρονική στιγμή t = 0 οι χρήστες C, D είναι μολυσμένοι. Την επόμενη χρονική στιγμή, οι δύο αυτοί χρήστες έχουν τη δυνατότητα να μολύνουν τους A, G, H και B, E, F αντίστοιχα. Όπως φαίνεται, τη χρονική στιγμή t = 1 μολύνθηκαν μόνο οι A, H, E και οι C, D που μολύνθηκαν αρχικά γίνονται πλέον ανενεργοί. Για t = 2, μολύνονται οι G, F και οι προηγουμένως μολυσμένοι A, E, H γίνονται ανενεργοί. Τη χρονική στιγμή t = 3 ο χρήστης G δεν έχει γείτονα που να μην έχει μολυνθεί και ο χρήστης F έχει την ευκαιρία να μολύνει το χρήστη I, ωστόσο αποτυγχάνει. Αφού δεν υπάρχουν πλέον μολυσμένοι χρήστες, η διαδικασία διάχυσης της πληροφορίας σταματά[41].



Εικόνα: Μόλυνση μελών δικτύου με χρήση του μοντέλου independent cascade[41].

4

Αποδοτική καταγραφική κάλυψη και συμπερασμός του δικτύου διάχυσης πληροφορίας σε πλατφόρμες κοινωνικής δικτύωσης

Όπως αναφέρθηκε στην παράγραφο 3.1.2 ο συμπερασμός ενός δικτύου διάχυσης πληροφορίας αποτελεί ένα περίπλοκο πρόβλημα εξέχουσας σημασίας. Λαμβάνοντας υπόψη πως η καταγραφή και παρακολούθηση της διάχυσης πληροφορίας στα OSNs χρειάζεται σημαντικές δαπάνες για καταγραφή, μία προσέγγιση συμπερασμού για μία διαδικασία διάχυσης πληροφορίας μέσα από αποδοτική παρακολούθηση προτείνεται στη συγκεκριμένη εργασία. Η πληροφορία θεωρείται πως ανήκει όλη σε μία κατηγορία. Στο πλαίσιο της εργασίας αξιοποιούνται μετρικές της ανάλυσης κοινωνικών δικτύων με σκοπό να μειωθούν το κόστος για παρακολούθηση και οι πόροι (resources) που θα απαιτούνταν για μία εξαντλητική παρακολούθηση, ενώ παράλληλα χρησιμοποιείται και πιθανοτικός συμπερασμός για την ακρίβεια της ιχνηλάτησης πληροφορίας.

Η πληροφορία θεωρείται πως διαδίδεται στο δίκτυο μέσω του μοντέλου Independent Cascade και στη συνέχεια αξιοποιείται η τεχνική του backtracking σε συνδυασμό με τις πιθανότητες διάδοσης προκειμένου να συμπεραθεί το δίκτυο διάχυσης πληροφορίας. Ως backtracking ορίζεται η brute force προσέγγιση κατά την οποία αν η παρούσα λύση δεν καλύπτει την επίλυση του προβλήματος, δοκιμάζονται άλλες πιθανές λύσεις οι οποίες προκύπτουν με αναδρομή.

Η αποτελεσματικότητα αυτής της μεθόδου παρουσιάζεται με την εφαρμογή της σε Small-World(SW) γράφους, οι οποίοι προσομοιάζουν online social networks [40][18]. Για την αναπαράσταση των OSNs χρησιμοποιήθηκαν Small-World (SW) γράφοι στους οποίους κάθε κόμβος αναπαριστά ένα χρήστη και κάθε ακμή αναπαριστά τη φιλία δύο κόμβων-χρηστών στα πλαίσια του OSN. Συγκεκριμένα, χρησιμοποιήθηκε το μοντέλο Watts-Strogatz, το οποίο ανήκει στην ευρύτερη κατηγορία των δικτύων μικρού κόσμου.
Επομένως, η ύπαρξη ακμής που συνδέει δύο κόμβους-χρήστες, υποδεικνύει τη δυνατότητα να περάσει η πληροφορία από τον έναν στον άλλον, χωρίς όμως να το καθιστά δεδομένο ότι θα συμβεί.

Η επιλογή των κόμβων που θα χρησιμοποιηθούν ως resources γίνεται σε συνάρτηση με το βαθμό του κάθε κόμβου. Πιο συγκεκριμένα, κόμβοι με πολλούς γείτονες είναι πιο πιθανό να επιλεχθούν για monitoring σε σχέση με κόμβους με μικρότερες γειτονιές. Η εύρεση του καλύτερου συνόλου προς επιλογή για monitoring, ανάγεται σε μία εκδοχή του προβλήματος του minimum dominating set, το οποίο παρουσιάστηκε στο κεφάλαιο 2.1.4.

Για το συμπερασμό της διάχυσης πληροφορίας, καταγράφονται οι γειτονιές (cascades) του πραγματικού δικτύου (ground truth network) και στη συνέχεια οι monitored cascades, δηλαδή αυτές που περιλαμβάνουν έστω έναν monitor.
Επίσης, για το συμπερασμό αξιοποιείται ένας κατευθυνόμενος γράφος με βάρη στις ακμές του w(u,v)=rel(v) που αντιπροσωπεύουν την σχετικότητα των χρηστών στην πληροφορία.

Τέλος, χρειάζεται να αξιολογηθεί κατά πόσο ο συμπερασμός(ή αλλιώς και η πρόβλεψη) του δικτύου διάχυσης πληροφορίας ανταποκρίνεται στην πραγματικότητα.
Αυτό επιτυγχάνεται συγκρίνοντας τον αριθμό των cascades του πραγματικού δικτύου διάχυσης με αυτόν των cascades που προέκυψαν από τη διαδικασία του συμπερασμού(inference).

Ακολούθως, παρουσιάζονται αναλυτικότερα οι αλγόριθμοι με βάση τους οποίους επιχειρήθηκε η αντιμετώπιση των ανωτέρω προκλήσεων:

4.1 Διάχυση πληροφορίας

Για τη διάχυση της πληροφορίας στο δίκτυο χρησιμοποιήθηκε το μοντέλο διάχυσης πληροφορίας "Independent Cascade":

- Independent Cascade (IC)

Σε αυτό το μοντέλο διάδοσης πληροφορίας, ένας χρήστης u έχει τη δυνατότητα να διαδώσει την πληροφορία σε περισσότερους από έναν από τους γείτονές του σε κάθε μονάδα χρόνου (timestamp), ο αριθμός των οποίων επιλέγεται τυχαία.
Κάθε κόμβος χαρακτηρίζεται από τη μετρική relevance (r), που αναπαριστά το κατά πόσο ο εν λόγω κόμβος σχετίζεται με την πληροφορία που διαχέεται στο δίκτυο.
Κάθε ζευγάρι γειτονικών κόμβων χαρακτηρίζεται από τη μετρική influence f(u,v), η οποία αναπαριστά την επιρροή που ασκεί το κόμβος u στον γείτονά του v.
Κάθε γείτονας v του u έχει πιθανότητα να επιλεγεί, ίση με

$$p(u,v) = \frac{s(u,v)}{\sum_{v:(u,v)\in E} s(u,v)}.$$

όπου s (u,v) = r(v) + f(u,v) γραμμική συνάρτηση του relevance και του influence.

Ωστόσο, οι χρήστες που έχουν λάβει την πληροφορία δε μπορούν να ανταλλάξουν ξανά σχετικό περιεχόμενο, και άρα θα υπάρχουν μόνο μονοπάτια διάδοσης.

Η μετρική που αναπαριστά τη συσχέτιση του κάθε κόμβου u με την πληροφορία (relevance) ορίζεται για κάθε κόμβο ως ένας τυχαίος αριθμός, όπου $0 < \text{Relevance}_u < 1$.

Algorithm 1: Relevance

```
Input  : Number of nodes as usernum, seedno
Output: One list rel, having the relevance of each node at rel[node]
for for every node in range of usernum do
    rel[node] = random number, between 0 and 1;
end
```

Η μετρική που αναπαριστά την επιρροή που ασκεί ένας κόμβος u σε ένα γείτονά του v (influence) ορίζεται για κάθε ζευγάρι κόμβων ως ένας τυχαίος αριθμός, όπου
$0 < \text{Influence}_{u,v} < 1$.

Algorithm 2: Influence

```
Input  : A network topology G(V,E), seedno
Output: One 2D Array infl, having the influence node i to node j at infl[i][j]
for for node in G.nodes do
    for for neighbor in node's neighbors do
        infl[node][neighbor] = random number, between 0 and 1;
        infl[neighbor][node] = random number, between 0 and 1;
    end
end
```

Για την αναπαράσταση των πιθανοτήτων διάδοσης από κόμβο - χρήστη σε έναν άλλον, χρησιμοποιήθηκε ένας γράφος με βάρη. Συγκεκριμενα, λαμβάνοντας ως δεδομένο το γράφο G(V,E) του δικτύου που μελετάται, προσθέτουμε σε αυτόν βάρη στις ακμές του οι οποίες αντιστοιχούν στις πιθανότητες διάδοσης p(u,v) που αναφέρθηκαν ανωτέρω. Αυτός ο γράφος δημιουργείται ως εξής:

Algorithm 3: Propagation Probabilities

```
Input  : 2D array or probagation probabilities p
Output: A network topology pG(V,E)
for neighbor in node's neighbors do
    add edge at pG with weight = p(node,neighbor);
end
```

Τέλος, παρουσιάζεται σε ψευδοκώδικα ο αλγόριθμος που χρησιμοποιήθηκε για την υλοποίηση του μοντέλου διάδοσης independent cascade:

Algorithm 4: Independent Cascade

```
Input  : The weighted network topology pG(V,E), percentage of seeds as perSeed,
         percentage to terminate as terminatePer, seedno, percentile
Output: list of infected, The diffusion network difG(V,E)
infected=[seeds];
difG = a graph difG(V,E) ;
add monitors, publics, privates in difG;
while terminatePer not reached do
    for i in seeds do
        for j in neighbors of i do
            if j not infected then
                x= probability interval based on the percentile of the infection prob distr. ;
                if p[i][j] greater than x then
                    add j to infected ;
                    add edge(i,j) to pG with timestamp ;
                    add j as seed ;
                else
            else
        end
    end
end
```

4.2 Ορισμός monitors

Σε κάθε OSN, υπάρχουν χρήστες με δημόσιο προφίλ, οι οποίοι επιτρέπουν σε χρήστες εκτός της λίστας φίλων τους να έχουν πρόσβαση στο περιεχόμενό τους, και κόμβοι-χρήστες με ιδιωτικό προφίλ, που δεν το επιτρέπουν.
Στη συγκεκριμένη αναπαράσταση του προβλήματος, για όσους χρήστες έχουν δημόσιο προφίλ είναι δυνατή η γνώση της χρονικής στιγμής (timestamp) που έλαβαν την πληροφορία. Αξιοποιώντας αυτό τους το χαρακτηριστικό ορίζουμε κάποιους εξ αυτών ως παρατηρητές (monitors).
Για την κατηγοριοποίηση των χρηστών στην κατηγορία των χρηστών με δημόσιο προφίλ ή στην κατηγορία αυτών με ιδιωτικό, χρησιμοποιήθηκαν δύο σχήματα. Και στα δύο αυτά σχήματα πρώτα επιλέγεται αν ένας χρήστης θα είναι monitor -και άρα θα έχει δημόσιο προφίλ- ή όχι. Στη συνέχεια, από τους non-monitors επιλέγονται αυτοί που θα οριστεί να έχουν δημόσιο προφίλ.

- Ντετερμινιστική κατηγοριοποίηση (deterministic classification):

Στη συγκεκριμένη περίπτωση πραγματοποιείται η εύρεση ενός minimum maximal independent set. Συγκεκριμένα, όσο υπάρχουν κόμβοι οι οποίοι δεν έχουν κατηγοριοποιηθεί ως monitored ή non-monitored, επιλέγεται σε κάθε επανάληψη ως monitor ο κόμβος που γειτονεύει με τους περισσότερους "non-monitored" κόμβους και δεν είναι γείτονας με monitor.

Algorithm 5: Deterministic classification of monitors

```
Input  : A network topology G(V,E)
Output: Two lists, one of monitors,one of non monitors
while Not classified nodes exist do
    sort all nodes by degree;
    classify first node A as monitor;
    remove it from the list of the non-classified nodes ;
    for for every neighbor of A do
        classify it as non-monitor;
        remove it from the list of the non-classified nodes ;
    end
end
```

- Κατηγοριοποίηση βάσει πιθανοτήτων (probabilistic classification):

Λαμβάνοντας ως δεδομένο το πλήθος των χρηστών του δικτύου το οποίο επιθυμείται να έχει δημόσιο λογαριασμό (monitors), επιλέγεται συγκεκριμένος αριθμός χρηστών που θα έχει αυτό το χαρακτηριστικό. Σε αυτήν την περίπτωση, μεγαλύτερος αριθμός γειτόνων συνεπάγεται μεγαλύτερη πιθανότητα να επιλεγεί ένας κόμβος-χρήστης ως χρήστης-monitor, και δεν υπάρχει η δυνατότητα δύο χρήστες-monitors να είναι γείτονες.

Για τη συγκεκριμένη επιλογή, δοκιμάστηκε η επίλυση μιας παραλλαγής του προβλήματος εύρεσης ενός budgeted independent set, καθώς είναι επιθυμητο να αφιερωθούν όσο το δυνατόν λιγότεροι πόροι (resources) για παρακολούθηση του δικτύου(monitoring). Όπως φαίνεται στον αλγόριθμο 6 (Algorithm 6: Probabilistic classification of monitors) σε κάθε επανάληψη ακολουθείται η εξής λογική. Όσο υπάρχουν κόμβοι οι οποίοι δεν έχουν κατηγοριοποιηθεί και το πλήθος των κόμβων που επιθυμείται να είναι monitors δεν έχει μηδενίσει, οι κόμβοι ταξινομούνται με βάση το βαθμό τους και αυτός με το μεγαλύτερο βαθμό ορίζεται ως monitor. Παράλληλα, οι γείτονές του κατηγοριοποιούνται ως non-monitors. Τέλος, αφαιρείται ένας κόμβος από το πλήθος που απομένει να κατηγοριοποιηθεί ως monitors.

Algorithm 6: Probabilistic classification of monitors

```
Input  : A network topology G(V,E), desired number of monitors as budget
Output: Three lists, one of monitors,one of non monitors,one of unclassified nodes
while Not classified nodes exist and budget greater than zero do
    sort all nodes by degree;
    classify first node A as monitor;
    remove it from the list of the non-classified nodes ;
    for for every neighbor of A do
        classify it as non-monitor;
        remove it from the list of the non-classified nodes ;
    end
    budget= budget - 1 ;
end
```

Για τον ορισμό των public-private profiles, επιλέγεται ένα ποσοστό των non-monitors το οποίο θα οριστούν ως χρήστες με δημόσιο προφίλ. Οι υπόλοιποι non-monitors, θα έχουν ιδιωτικό προφίλ.

Algorithm 7: Classify non-monitors as public or not

```
Input  : A list of non monitors, percentage of public nodes as per, seedno
Output: Two lists, one of public nodes,one of private nodes
generate random number based on seedno;
get random sample of publics based on per;
classify rest nodes of list as privates;
```

4.3 Καταγραφή πληροφορίας

Κατα τη διάχυση της πληροφορίας ανάμεσα στους κόμβους - χρήστες του δικτύου καταγράφεται η χρονική στιγμή κατά την οποία κάθε χρήστης-monitor έλαβε την πληροφορία. Ωστόσο, δεν υπάρχει πρόσβαση στις χρονικές στιγμές κατά τις οποίες οι private χρήστες έλαβαν την πληροφορία.
Επίσης, θεωρείται γνωστή η μορφολογία του δικτύου και κατ' επέκταση η γειτονιά κάθε κόμβου-χρήστη.

4.4 Συμπερασμός δικτύου διάχυσης

Χρησιμοποιήθηκαν δύο σχήματα για το συμπερασμό του δικτύου διάχυσης πληροφορίας.
Από τους seeds (είναι οι κόμβοι που ξεκινούν τη διάδοση της πληροφορίας στο δίκτυο και έχουν επιλεγεί να είναι και monitors), γνωρίζουμε πόσοι κόμβοι (public και private)

συμμετείχαν στα cascades τους, δηλαδή πόσοι έλαβαν έμμεσα ή άμεσα την πληροφορία από αυτούς. Για κάθε κόμβο με public profile, που έχει λάβει την πληροφορία από seed-monitor, γνωρίζουμε ποιοι άλλοι κόμβοι με public profiles έλαβαν την πληροφορία από αυτόν. Το ντετερμινιστικό inference σχήμα επιλέγει το μονοπάτι που μεγιστοποιεί το συνολικό relevance των κόμβων που συμμετέχουν σε αυτό.

Το probabilistic inference σχήμα, επιλέγει με μεγαλύτερη πιθανότητα το μονοπάτι με το μεγαλύτερο συνολικό relevance. Βασική υπόθεση του συμπερασμού: Κάθε χρήστης λαμβάνει την πληροφορία από τον γείτονά του που τη δημοσίευσε πιο πρόσφατα.

Algorithm 8: Inference

```
Input  : A network topology iG(V,E), list of possible paths in the graph
Output: Two lists, one for deterministic and one for probabilistic inference
for every cascade do
    if ending node is not private then
        keep specific path ;
    end
    else
        paths to investigate are all paths in the graph with start node of this cascade as
         source ;
    end
    keep only the paths with fitting length ;
    count public users in the paths ;
    create list of candidate paths ;
    choose the path of the highest relevance ;
end
```

Όπως φαίνεται στον αλγόριθμο 8 (Algorithm 8: Inference) λαμβάνονται ως δεδομένα μία λίστα με όλα τα πιθανά μονοπάτια στο γράφο καθώς και ο γράφος iG(V,E) ο οποίος περιλαμβάνει όλα αυτά τα πιθανά μονοπάτια. Για κάθε μονοπάτι ακολουθείται η εξής διαδικασία: Αν το τελευταίος κόμβος δεν είναι private τότε σταματά η αναζήτηση σε αυτό το μονοπάτι και αυτό κρατείται προκειμένου να αξιολογηθεί στο τέλος σε σχέση με τα υπόλοιπα μονοπάτια ως προς το συνολικό του relevance, αλλιώς συνεχίζεται η αναζήτηση στο ίδιο μονοπάτι. Στη δεύτερη περίπτωση υποδηλώνεται ότι τα μονοπάτια προς εξερεύνηση είναι τα μονοπάτια τα οποία ξεκινούν με αρχικό τον κόμβο από τον οποίο ξεκίνησε το συγκεκριμένο μονοπάτι. Τελικά, κρατούνται προς αξιολόγηση μόνο τα μονοπάτια που έχουν μήκος που να ταιριάζει με τη διαφορά στα timestamps στα οποία οι χρήστες-monitors τους έλαβαν την πληροφορία και μετρώνται οι public χρήστες σε αυτά. Αυτά τα μονοπάτια συγκροτούν τη λίστα με τα μονοπάτια τα οποία θεωρούνται υποψήφια για την αναπαράσταση της πραγματικής κίνησης της πληροφορίας και από την οποία επιλέγεται ένα, αυτό που έχει το μεγαλύτερο συνολικό relevance ως προς την πληροφορία

(ντετερμινιστικό σχήμα συμπερασμού) ή ένα από αυτά που έχουν το μεγαλύτερο relevance ως προς την πληροφορία (πιθανοτικό σχήμα συμπερασμού).
Το συνολικό relevance κάθε μονοπατιού προκύπτει ως το άθροισμα του επιμέρους relevance κάθε κόμβου-χρήστη του μονοπατιού.

5

Αποτελέσματα τεχνικών καταγραφής και συμπερασμού

5.1 Αποτελέσματα για το πρώτο σχήμα monitor selection - Ντετερμινιστική κατηγοριοποίηση

Για τη συλλογή των αποτελεσμάτων χρησιμοποιήθηκαν Small World δίκτυα με διαφορετικές παραμέτρους όσον αφορά το μέγεθος του δικτύου (network size : s), τον αριθμό των γειτόνων με τους οποίους συνδέεται κάθε κόμβος (neighbour number : k) και την πιθανότητα να αλλάξει ο κόμβος στον οποίο συνδέεται μια υπάρχουσα ακμή (rewiring probability : p).

Ακολούθως, παρουσιάζονται οι μέσοι όροι ανά διαφορετικό συνδυασμό των ανωτέρω παραμέτρων. Για κάθε συνδυασμό, χρησιμοποιήθηκαν 100 δίκτυα.

100 δίκτυα με s : 100								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	25.52	27.34	19.80	21.90	16.21	18.17	14.04	15.95
Private profiles	71.48	69.96	76.50	74.91	79.79	77.88	81.96	80.05
Infected users	85.18	86.41	91.42	91.92	91.13	92.25	93.34	98.30
Accuracy of deterministic inference	53.87	54.70	32.28	30.35	21.90	21.73	15.78	16.05
Accuracy of probabilistic inference	51.35	52.73	29.27	28.05	19.14	20.33	13.94	13.92

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 30.83%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 28.59%

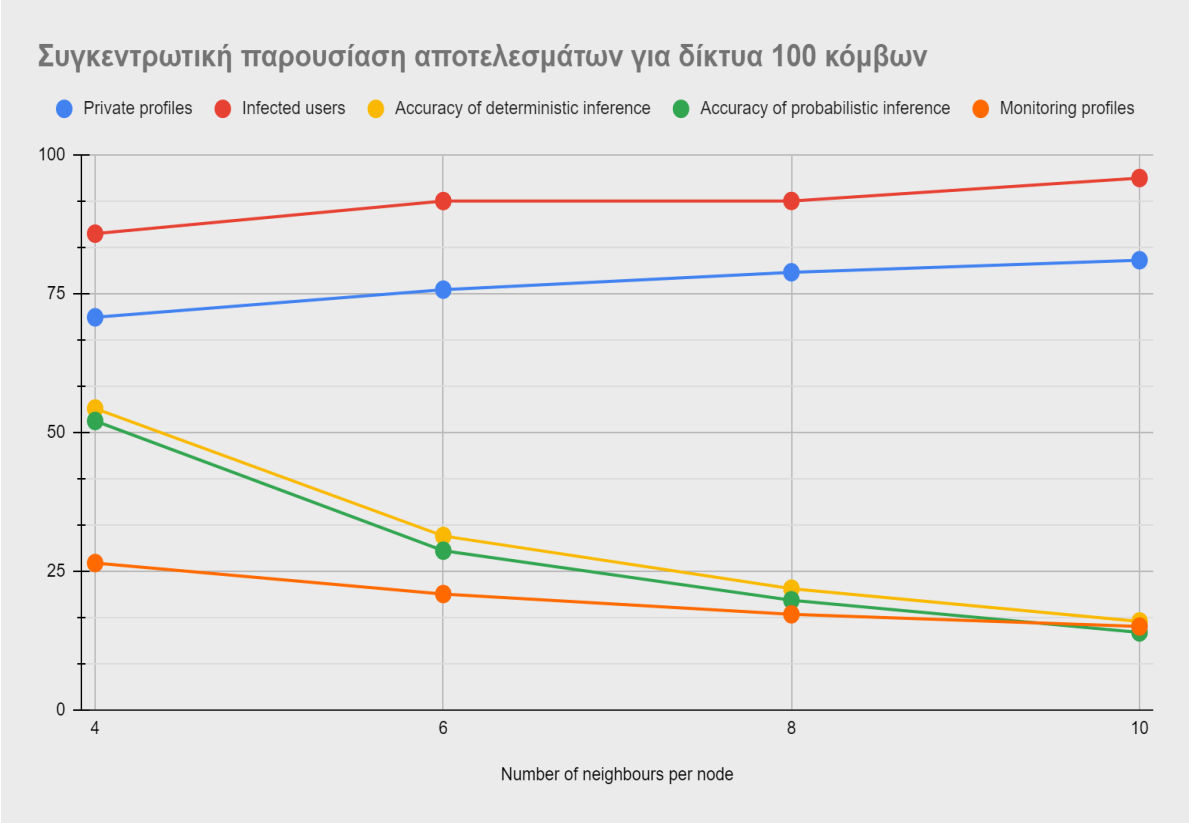
100 δίκτυα με s : 200								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	50.85	54.83	39.56	29.79	32.56	36.61	28.02	31.62
Private profiles	142.15	138.19	152.73	162.21	159.44	155.40	163.98	160.38
Infected users	170.12	170.52	177.19	187.03	183.81	182.42	178.15	194.95
Accuracy of deterministic inference	57.32	56.91	34.27	18.17	23.97	23.68	19.27	17.83
Accuracy of probabilistic inference	54.09	55.09	31.30	16.38	21.42	21.26	16.78	15.96

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 31.42%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 29.04%

100 δίκτυα με s : 300								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	76.17	82.30	59.18	65.30	49.26	55.20	41.65	46.28
Private profiles	212.86	207.41	229.10	223.68	238.74	232.83	246.06	241.52
Infected users	255.53	253.37	261.54	282.24	283.09	267.45	264.58	295.31
Accuracy of deterministic inference	58.35	60.30	35.96	35.58	24.08	25.29	18.77	17.34
Accuracy of probabilistic inference	55.38	58.18	33.34	33.96	22.27	23.66	16.92	15.21

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 31.96%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 32.37%

Στη συνέχεια παρουσιάζεται μία σύγκριση των διαφόρων μετρικών για τις διαφορετικές τιμές των παραμέτρων που εξετάζονται. Εν γένει, φαίνεται πως το μέγεθος του δικτύου δεν οδηγεί σε αναλογικά μεγάλες διαφορές καθώς ποσοστιαία τα μεγέθη παραμένουν παρεμφερή. Διαφορές παρατηρούνται κυρίως με τις διακυμάνσεις των τιμών p,k.
Πιο συγκεκριμένα:

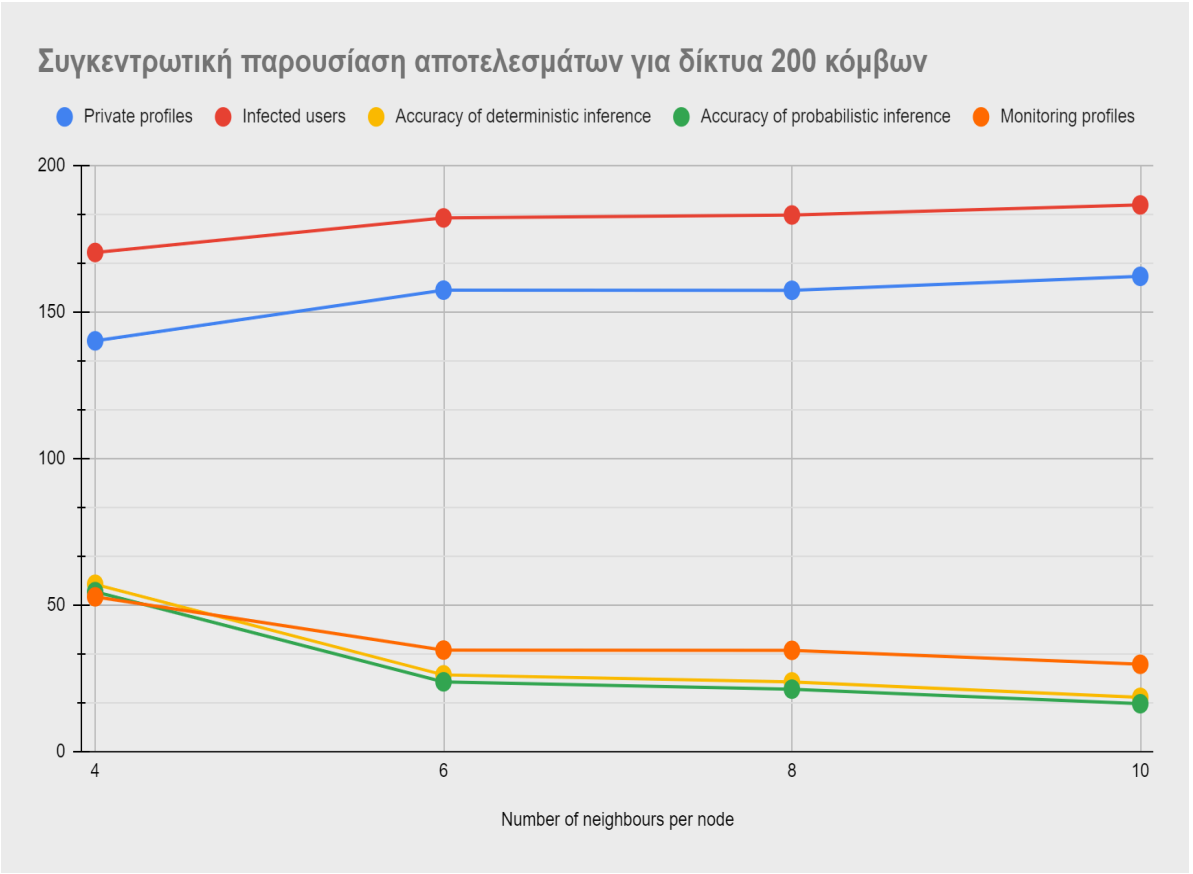


Αριθμός monitoring profiles

Από τα παραπάνω αποτελέσματα φαίνεται πως αύξηση της τάξης του 150% (0.2 → 0.5) στην τιμή της πιθανότητας να γίνει rewiring μίας ακμής κρατώντας σταθερές τις τιμές s,k οδηγεί σε πολύ μικρή αύξηση – της τάξης του 10% - του αριθμού των monitoring profiles. Αντίθετα, μία αύξηση 50% στον αριθμό των γειτόνων με τους οποίους συνδέεται ο κάθε κόμβος οδηγεί σε μείωση κατά 10-20%. Αυτή η μείωση προκύπτει από τον περιορισμό πως δε μπορεί να οριστεί ως monitor ένας κόμβος που γειτονεύει με κόμβο-monitor. Συγκεκριμένα, αύξηση στον αριθμό των κόμβων με τους οποίους γειτονεύει κάθε κόμβος συνεπάγεται μεγαλύτερο αριθμό κόμβων οι οποίοι συνδέονται απευθείας με κάποιον monitor και άρα δε μπορούν να οριστούν ως monitors.

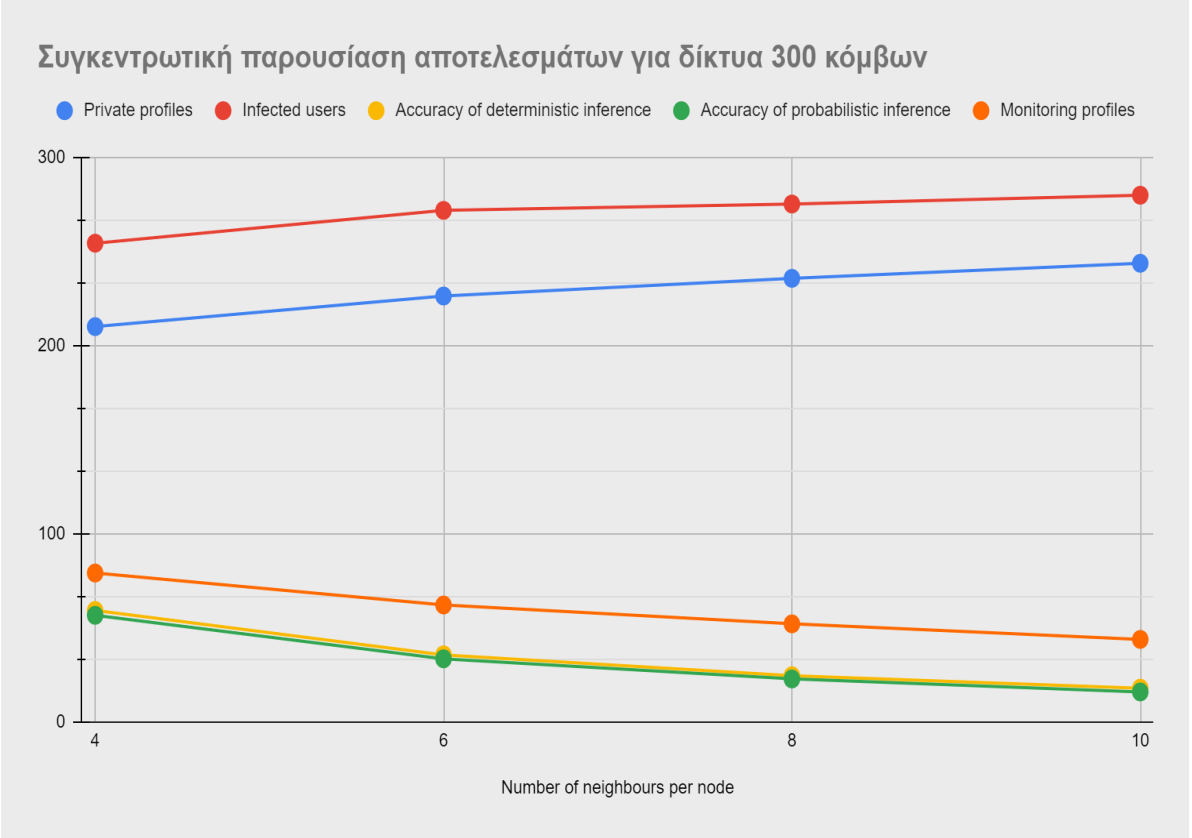
Αριθμός private profiles

Όσον αφορά τον αριθμό των private profiles, για την αντίστοιχη αύξηση του p διαφαίνεται μείωση 2-3%. Αντίστοιχα, με την αντίστοιχη αύξηση του k παρατηρείται αύξηση 3-7% στον αριθμό των ιδιωτικών λογαριασμών. Για τον ίδιο λόγο για τον οποία παρατηρείται μείωση των monitoring profiles, παρατηρείται η συγκεκριμένη αύξηση των private profiles για την περίπτωση της αύξησης του k. Μεγαλύτερος αριθμός non-monitoring κόμβων θα οδηγήσει αυτόματα και σε μεγαλύτερο αριθμό private κόμβων, αφού οι private κόμβοι αποτελούν υποσύνολο των non-monitoring κόμβων.



Αριθμός χρηστών στους οποίους έφτασε η πληροφορία

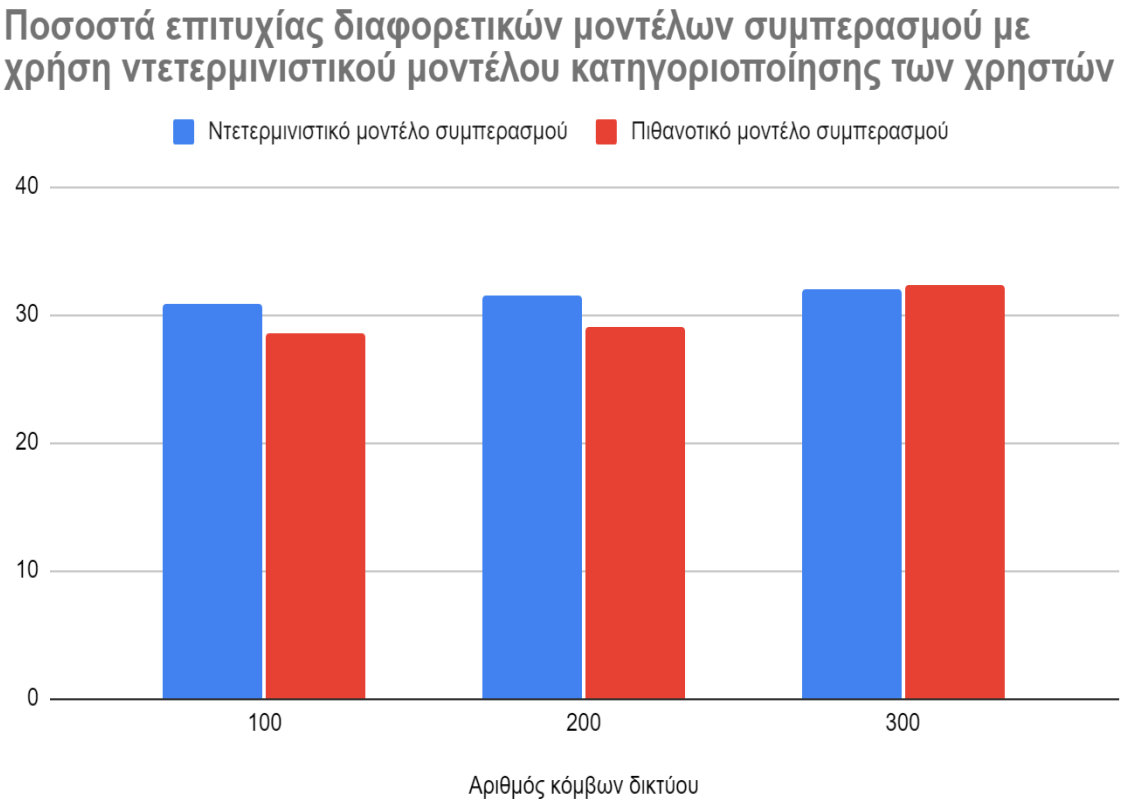
Στις περισσότερες περιπτώσεις φαίνεται να μην ακολουθεί μία σταθερή συμπεριφορά αυτή η μετρική από την πιθανότητα να αλλάξει προορισμό μια υπάρχουσα ακμή, με εξαίρεση την περίπτωση για k = 10, όπου και παρατηρείται αύξηση 5%. Αντίθετα, η αύξηση αυτή φαίνεται να επαναλαμβάνεται σε κάθε περίπτωση αύξησης του k, για p σταθερό. Αυτό φαίνεται λογικό, καθώς είναι ευκολότερο για την πληροφορία να κινηθεί σε πιο πυκνά δίκτυα (δίκτυα στα οποία ο κάθε κόμβος που μολύνεται έχει πληθώρα γειτόνων στους οποίους μπορεί να μεταδώσει την πληροφορία) παρά σε πιο αραιά δίκτυα.



Ποσοστό ακρίβειας συμπερασμού

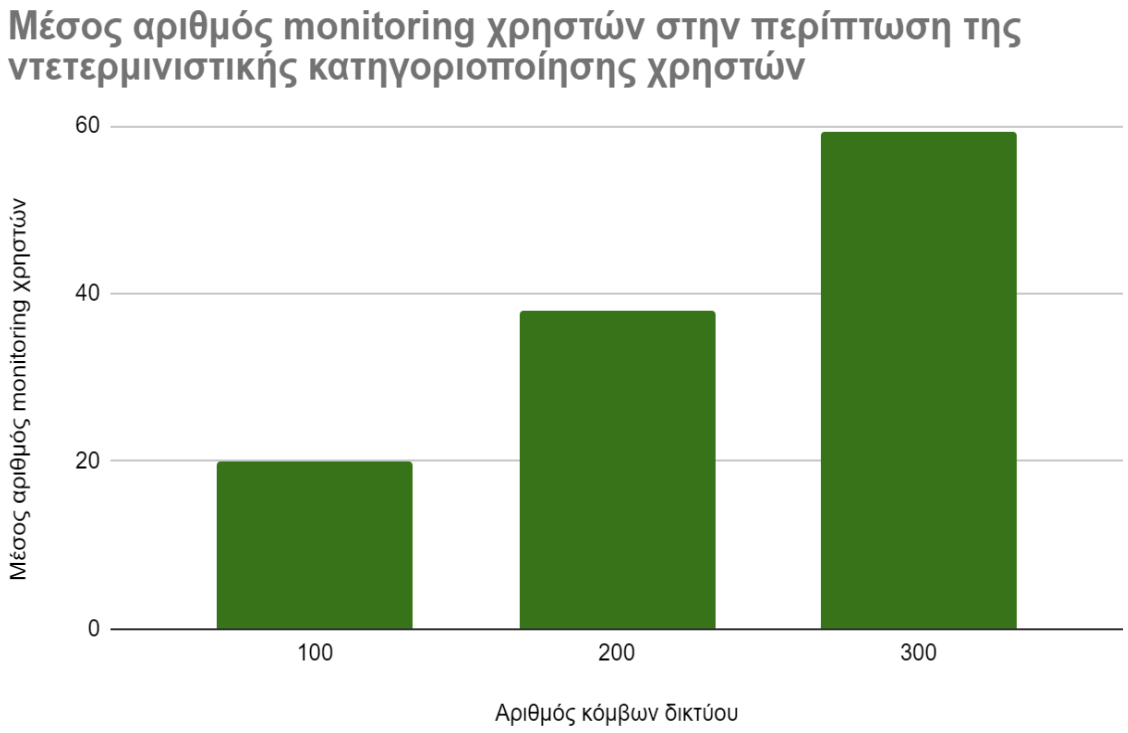
Παρατηρείται παρόμοια συμπεριφορά όσον αφορά τις δύο διαφορετικές μεθόδους συμπερασμού. Συγκεκριμένα, δε σημειώνονται σημαντικές διαφορές κατά τη μεταβολή του p κρατώντας σταθερά τα s, k. Αντιθέτως, η αύξηση των γειτόνων κάθε κόμβου από τους 4 στους 6 σχεδόν υποδιπλασιάζει τα δύο ποσοστά, και περαιτέρω αύξηση του συγκεκριμένου μεγέθους συνεχίζει να τα μειώνει. Αιτία αυτής της μείωσης είναι το ότι με την αύξηση του αριθμού των γειτόνων κάθε κόμβου αυξάνονται αναλογικά και τα πιθανά μονοπάτια μέσω των οποίων κινήθηκε η πληροφορία μέσα στο δίκτυο. Έτσι, τελικά θα υπάρχει μεγαλύτερος αριθμός υποψήφιων μονοπατιών προς αξιολόγηση, και κατ' επέκταση μεγαλύτερη δυσκολία στην επιτυχημένη αξιολόγηση αυτών των μονοπατιών.

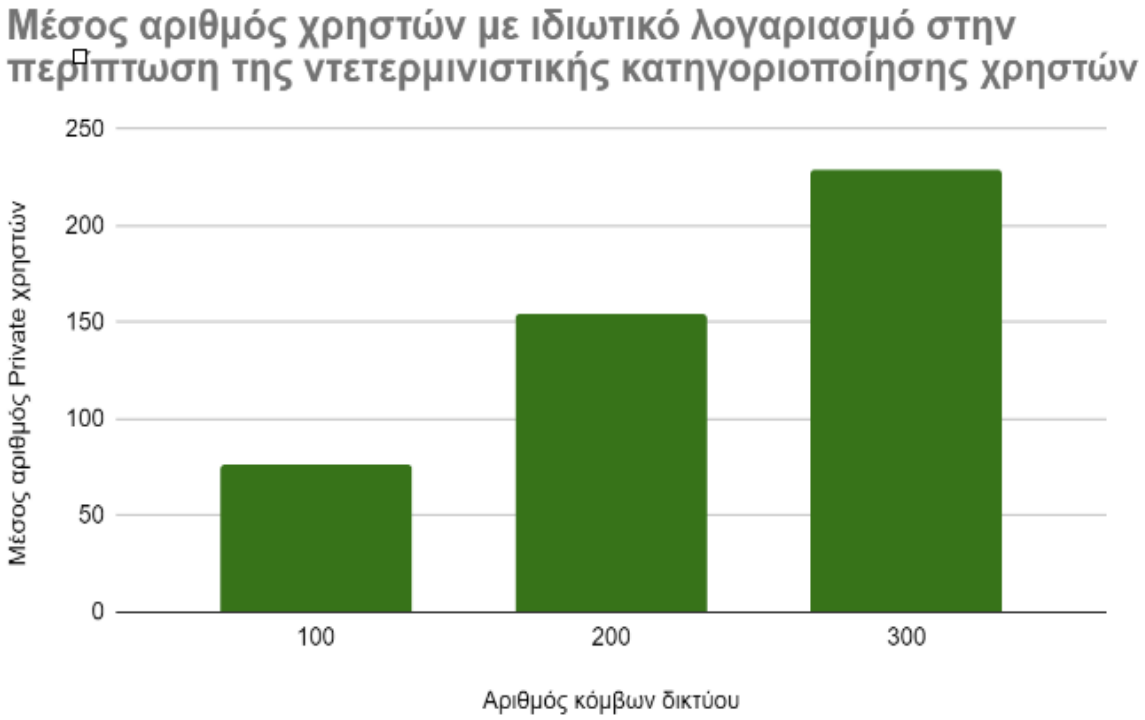
Συγκεντρωτικά, οι μέσες τιμές του ποσοστού ακρίβειας που σημειώθηκε με χρήση των διαφορετικών μοντέλων συμπερασμού παρουσιάζονται στο ακόλουθο γράφημα για δίκτυα 100, 200, 300 χρηστών.

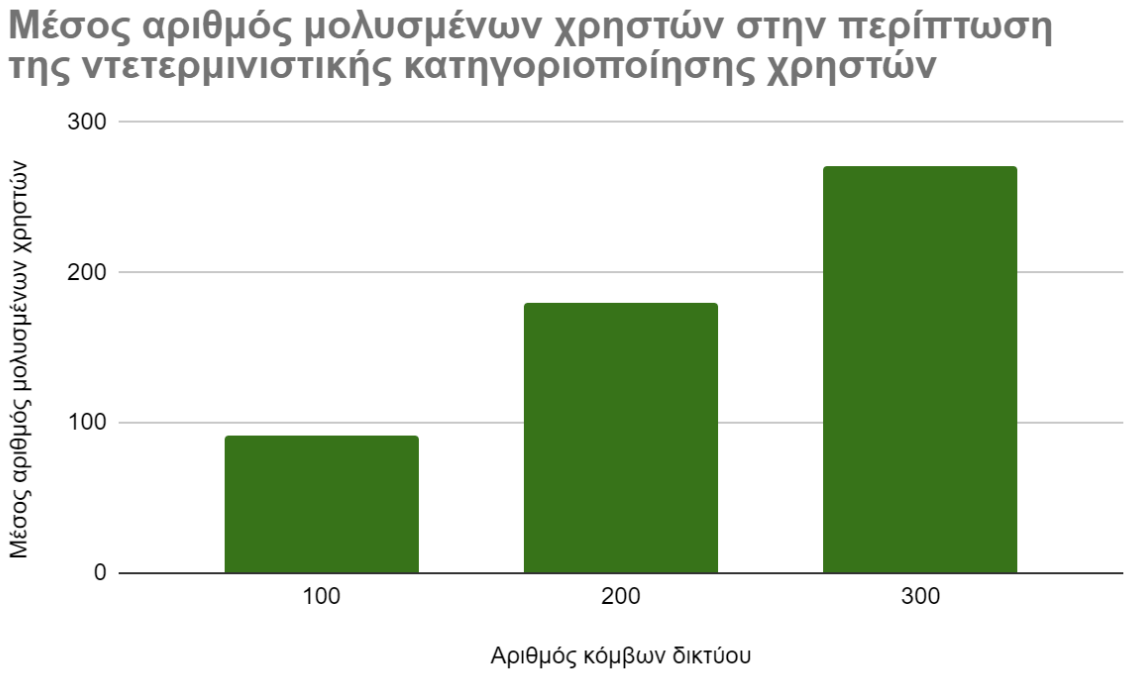


Φαίνεται πως το ντετερμινιστικό μοντέλο συμπερασμού οδηγεί σε καλύτερα αποτελέσματα για δίκτυα 100, 200 κόμβων, ενώ το πιθανοτικό υπερισχύει ελαφρά στα δίκτυα 300 κόμβων. Η αποτυχία του πιθανοτικού στις πρώτες δύο περιπτώσεις οφείλεται στο γεγονός πως με βάση αυτό το μοντέλο δεν επιλέγεται πάντα το μονοπάτι με το μεγαλύτερο συνολικό relevance, απλά έχει περισσότερες πιθανότητες να επιλεγεί. Αυτή η αδυναμία του πιθανοτικού μοντέλου συμπερασμού αμβλύνεται και γίνεται πλεονέκτημα στην περίπτωση των δικτύων 300 κόμβων καθώς υπάρχουν περισσότερα υποψήφια μονοπάτια με αποτέλεσμα να υπάρχει μείωση στις πιθανότητες το πραγματικό μονοπάτι στο οποίο κινήθηκε η πληροφορία να είναι αποκλειστικά αυτό με το μεγαλύτερο relevance.

Στη συνέχεια, παρουσιάζονται συγκεντρωτικά τα αποτελέσματα για τις μέσες τιμές στην περίπτωση του πρώτου σχήματος monitor selection (deterministic classification):


Μέσος αριθμός monitoring χρηστών στην περίπτωση της ντετερμινιστικής κατηγοριοποίησης χρηστών
60
40
20
0
Μέσος αριθμός monitoring χρηστών
100
200
300
Αριθμός κόμβων δικτύου


Μέσος αριθμός χρηστών με ιδιωτικό λογαριασμό στην περίπτωση της ντετερμινιστικής κατηγοριοποίησης χρηστών
250
200
150
100
50
0
Μέσος αριθμός Private χρηστών
100
200
300
Αριθμός κόμβων δικτύου



Φαίνεται πως υπάρχει ένα μοτίβο στους αριθμούς των monitoring, των private και των μολυσμένων χρηστών. Συγκεκριμένα, όσο αυξάνεται ο αριθμός των συνολικών χρηστών του δικτύου, αναλογικά μεγαλώνουν και οι τρεις αυτοί αριθμοί. Αναλυτικά ο διπλασιασμός ή τριπλασιασμός του αρχικού αριθμού των χρηστών οδηγεί σε έναν αριθμό πολύ κοντά ή ακριβώς στον διπλασιασμό ή τριπλασιασμό και του αριθμού των χρηστών που σχετίζεται με τις προαναφερθείσες μετρικές.

5.2 Αποτελέσματα για το δεύτερο σχήμα monitor selection - Πιθανοτική κατηγοριοποίηση

Για τη συλλογή των αποτελεσμάτων χρησιμοποιήθηκαν Small World δίκτυα με διαφορετικές παραμέτρους όσον αφορά το μέγεθος του δικτύου (network size : s), τον αριθμό των γειτόνων με τους οποίους συνδέεται κάθε κόμβος (neighbour number : k) και την πιθανότητα να αλλάξει ο κόμβος στον οποίο συνδέεται μια υπάρχουσα ακμή (rewiring probability : p). Επίσης, η πρόσθετη παράμετρος που λαμβάνεται υπόψη στη συγκεκριμένη περίπτωση είναι το ποσοστό των χρηστών του δικτύου οι οποίοι θα χρησιμοποιηθούν ως monitors (monitor percentage: mp).

Για την τιμή του ποσοστού των χρηστών του δικτύου που θα χρησιμοποιηθούν ως monitors επιλέχθηκαν τα ποσοστά 15%, 30%, 40% καθώς παρατηρήθηκε πως με την επιλογή των monitors μέσω της ντετερμινιστικής μεθόδου, αυτοί συντελούσαν σε κάθε περίπτωση περίπου το 25% της δύναμης του δικτύου.

Ακολούθως, παρουσιάζονται οι μέσοι όροι ανά διαφορετικό συνδυασμό των ανωτέρω παραμέτρων. Για κάθε συνδυασμό, χρησιμοποιήθηκαν 100 δίκτυα.

5.2.1 Πιθανοτική κατηγοριοποίηση με ποσοστό monitors επί του δικτύου: 15%

100 δίκτυα με s : 100, mp: 15%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	15	15	15	15	14.94	14.99	13.94	14.85
Private profiles	63.85	64.21	75.99	74.28	79.79	77.86	81.96	80.05
Infected users	85.19	85.47	90.18	92.00	92.20	91.89	92.94	97.97
Accuracy of deterministic inference	61.73	63.35	33.10	31.10	21.28	21.82	15.57	15.86
Accuracy of probabilistic inference	59.29	60.16	29.20	28.75	18.78	19.21	13.78	14.02

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 32.98%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 30.40%

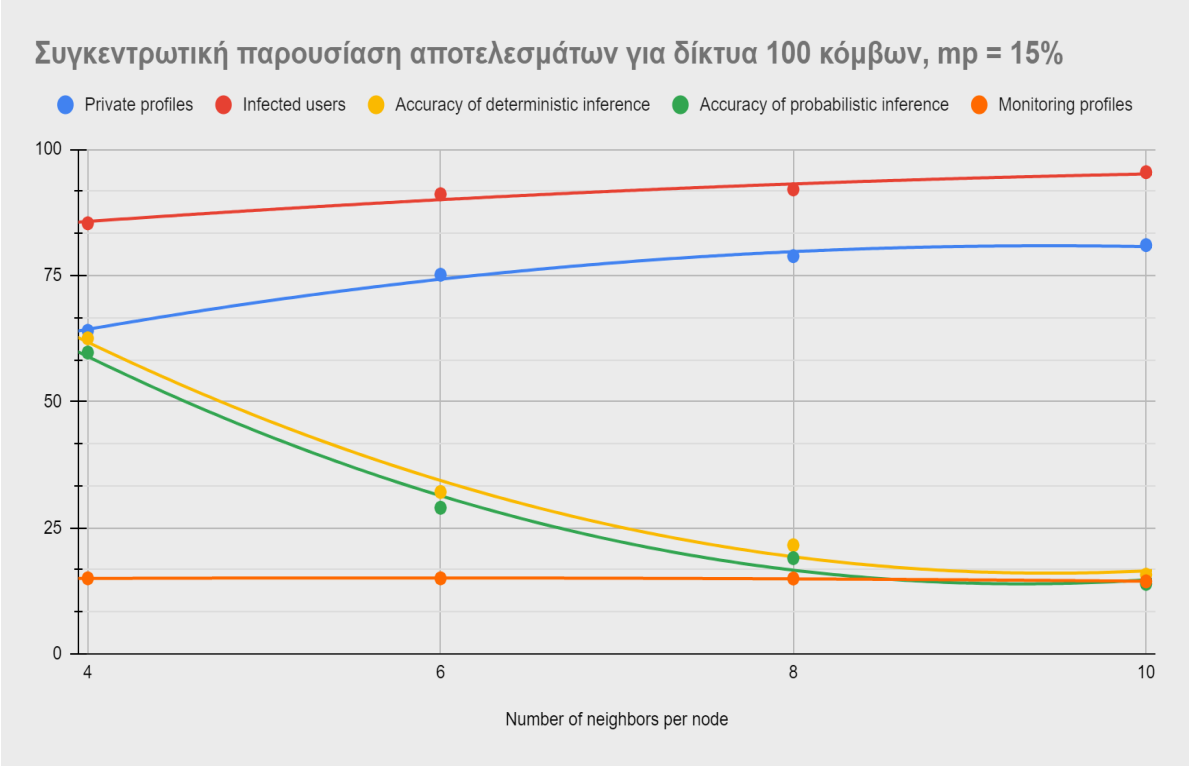
100 δίκτυα με s : 200, mp: 15%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	30	30	30	30	29.93	30	27.93	29.76
Private profiles	127.63	128.22	151.39	148.27	159.44	155.39	163.98	160.38
Infected users	169.69	170.64	182.05	179.83	187.91	177.01	177.63	192.06
Accuracy of deterministic inference	66.57	65.61	36.63	36.35	24.27	24.43	19.25	18.34
Accuracy of probabilistic inference	64.33	63.49	33.24	33.72	21.74	22.09	16.69	16.62

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 36.43%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 33.99%

100 δίκτυα με s : 300, mp: 15%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	45	45	45	45	45	45	41.65	44.91
Private profiles	190.81	191.99	227.31	221.49	238.74	232.82	246.06	240.01
Infected users	254.28	255.32	266.94	273.99	282.85	263.02	264.58	288.46
Accuracy of deterministic inference	66.65	65.45	36.01	36.39	24.18	24.70	18.77	18.67
Accuracy of probabilistic inference	63.65	64.17	33.65	34.09	22.27	22.80	16.92	17.12

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 36.35%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 34.33%

Παρατίθεται ο σχολιασμός των αποτελεσμάτων για σταθερή την τιμή του mp = 15%.

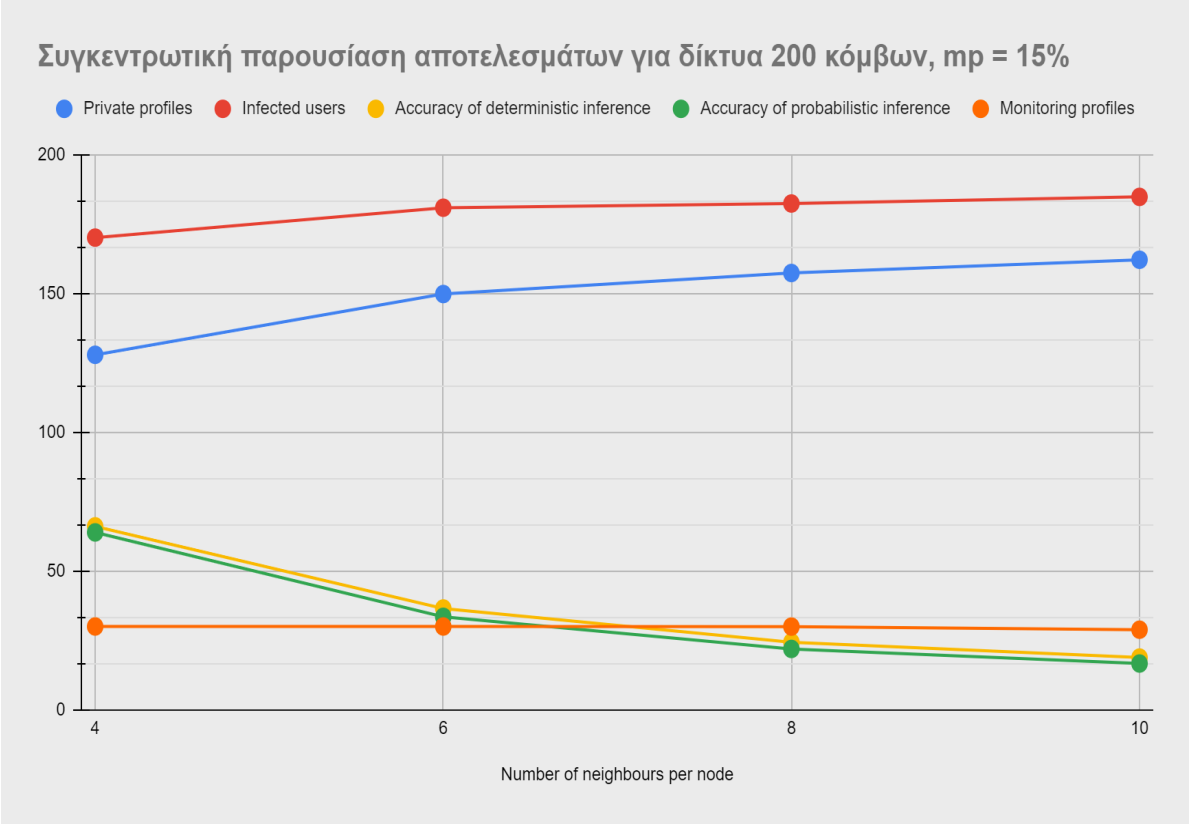


Αριθμός monitoring profiles

Στη συγκεκριμένη περίπτωση δεν κρίνεται σκόπιμο να σχολιαστεί κάτι πέραν του ότι σε όλα τα παραπάνω σενάρια ο αριθμός των monitors είναι συνεχώς πολύ κοντά ή ακριβώς στο επιλεγμένο ποσοστό.

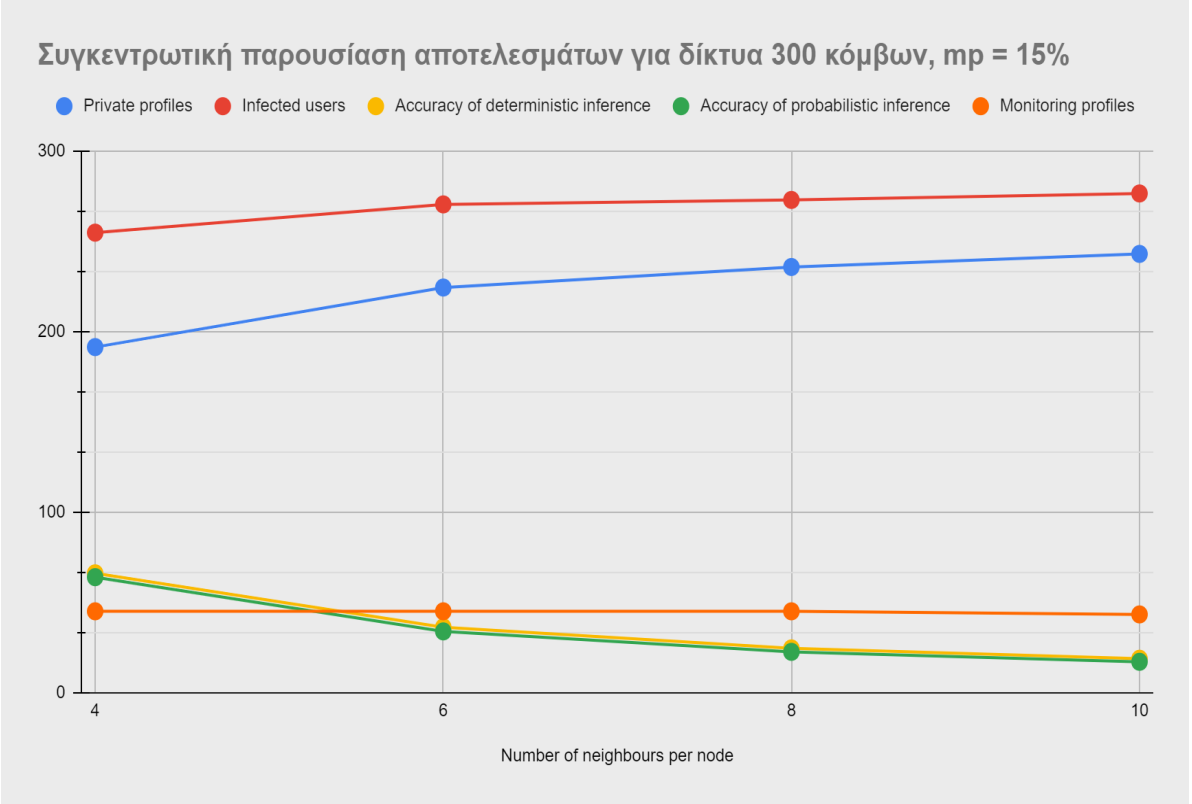
Αριθμός private profiles

Όσον αφορά τον αριθμό των private profiles, για την αντίστοιχη αύξηση του p διαφαίνεται μείωση 2-3%. Αντίστοιχα, με την αντίστοιχη αύξηση του k παρατηρείται αύξηση στον αριθμό των ιδιωτικών λογαριασμών. Η αύξηση αυτή οφείλεται στο γεγονός πως κάθε ένας από τους κόμβους - monitors γειτονεύει με περισσότερους κόμβους οι οποίοι θα κατηγοριοποιηθούν ως non-monitors και άρα περισσότερους κόμβους οι οποίοι θα κατηγοριοποιηθούν ως private. Αξιοσημείωτο είναι και το γεγονός πως στην περίπτωση για k = 10 δεν υπάρχει αρκετά μεγάλος αριθμός μη κατηγοριοποιημένων κόμβων ώστε να επιλεγούν οι monitors με βάση το επιθυμητό πλήθος monitors.



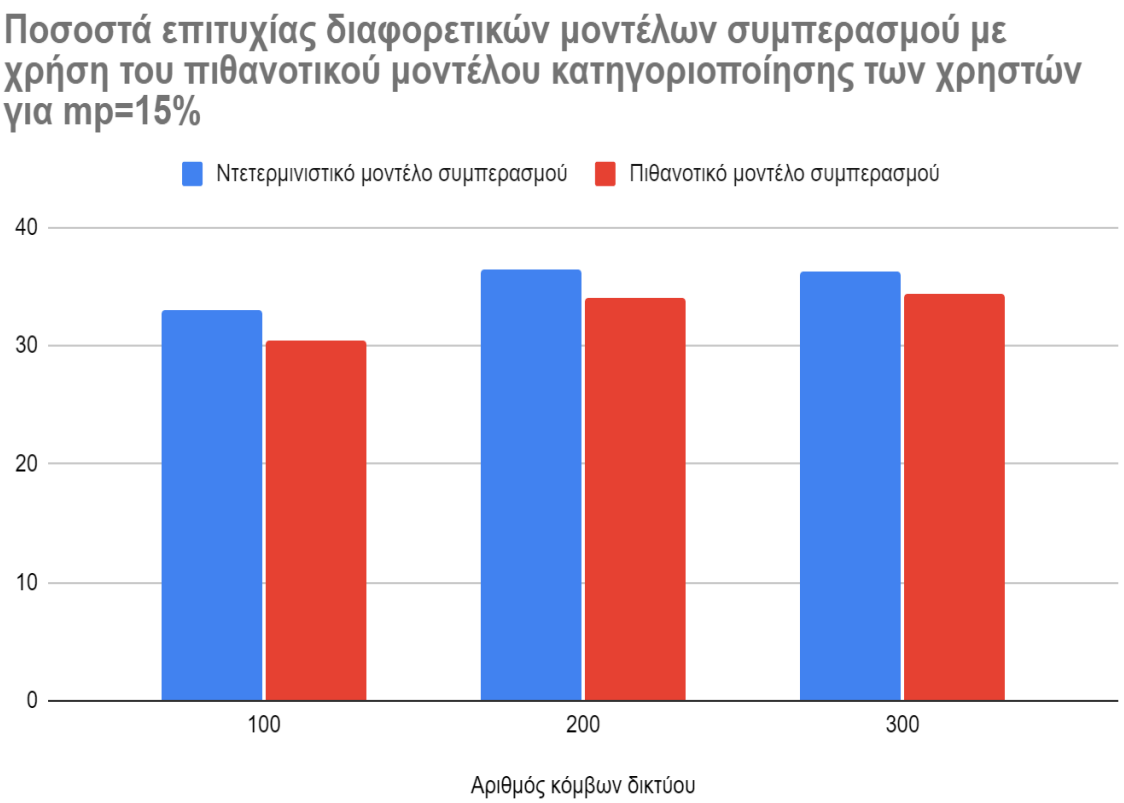
Αριθμός χρηστών στους οποίους έφτασε η πληροφορία

Στις περισσότερες περιπτώσεις φαίνεται πως αυτή η μετρική δεν κρατάει κάποια σταθερή συμπεριφορά σε σχέση με την αύξηση ή μείωση της πιθανότητας να αλλάξει προορισμό μια υπάρχουσα ακμή. Ωστόσο παρατηρείται ένα μοτίβο όσον αφορά τη συγκεκριμένη μετρική σε περίπτωση αύξησης του αριθμού των γειτόνων κάθε κόμβου, στην οποία και ο αριθμός των "μολυσμένων" χρηστών αυξάνεται. Κάτι απολύτως λογικό, καθώς αύξηση στον αριθμό των γειτόνων κάθε χρήστη συνεπάγεται μεγαλύτερο αριθμό εναλλακτικών διαδρομών στις οποίες μπορεί να κινηθεί η πληροφορία και μείωση των απομονωμένων χρηστών του δικτύου.

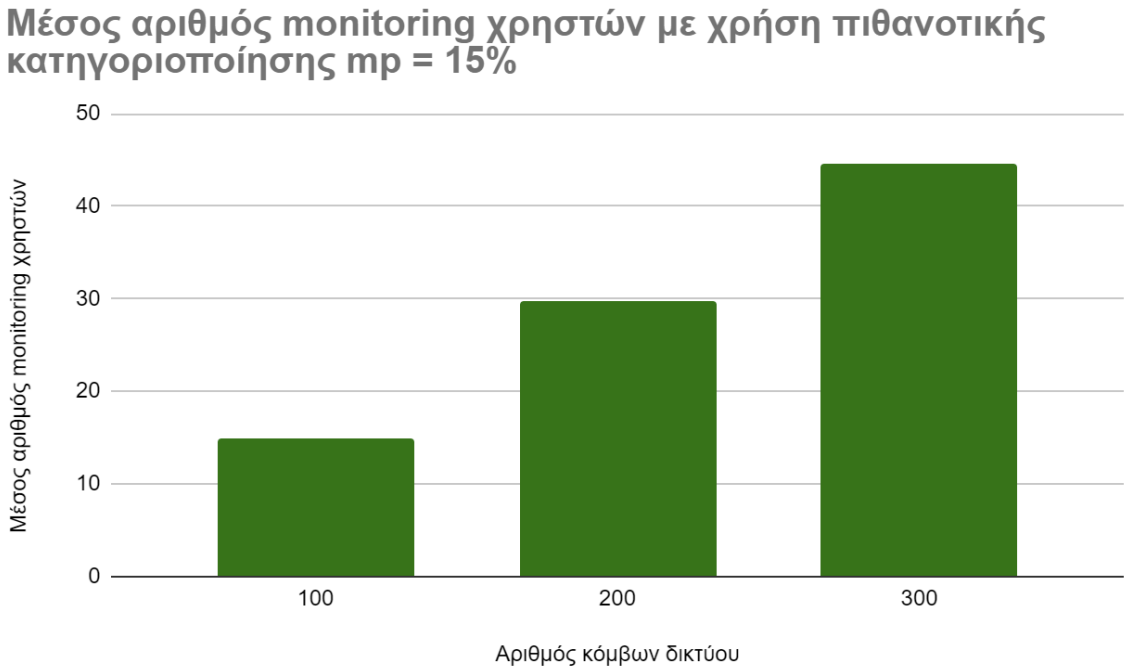


Ποσοστό ακρίβειας συμπερασμού

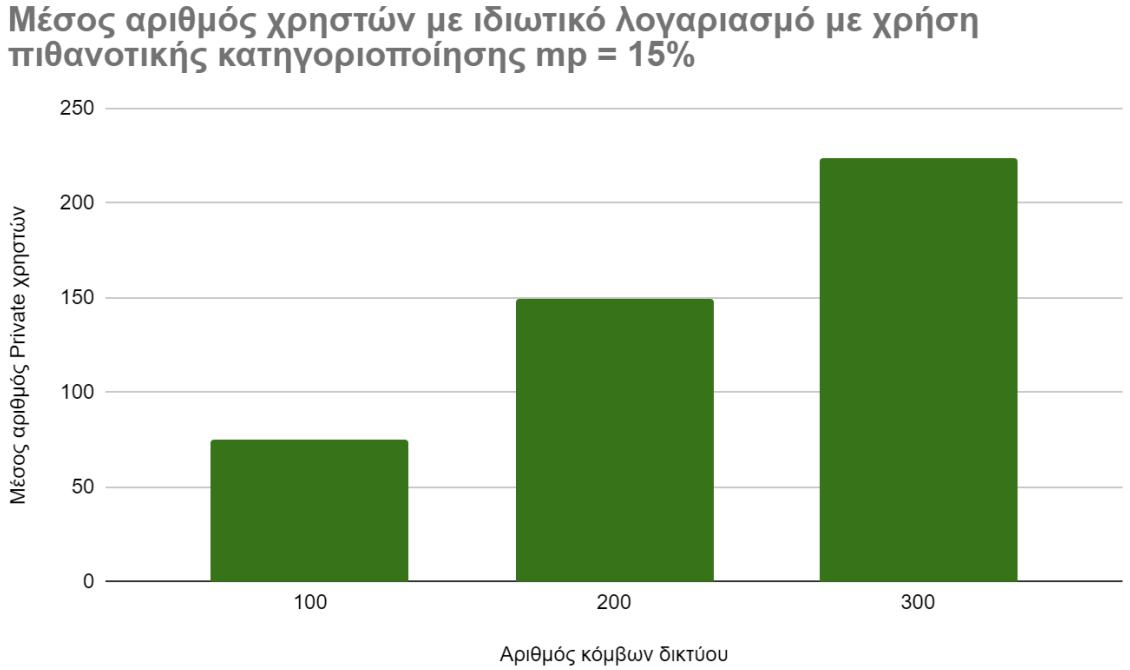
Παρατηρείται παρόμοια συμπεριφορά όσον αφορά τις δύο διαφορετικές μεθόδους συμπερασμού. Συγκεκριμένα, δε σημειώνονται σημαντικές διαφορές κατά τη μεταβολή του p κρατώντας σταθερά τα s, k. Αντιθέτως, η αύξηση των γειτόνων κάθε κόμβου από τους οδηγεί σε σημαντική μείωση των ποσοστών επιτυχίας, το οποίο οφείλεται στο γεγονός της αύξησης των υποψήφιων μονοπατιών για την κίνηση της πληροφορίας.
Άξιο αναφοράς φαίνεται και το γεγονός της βελτίωσης της απόδοσης του συμπερασμού, όσο αυξάνεται το μέγεθος του δικτύου.

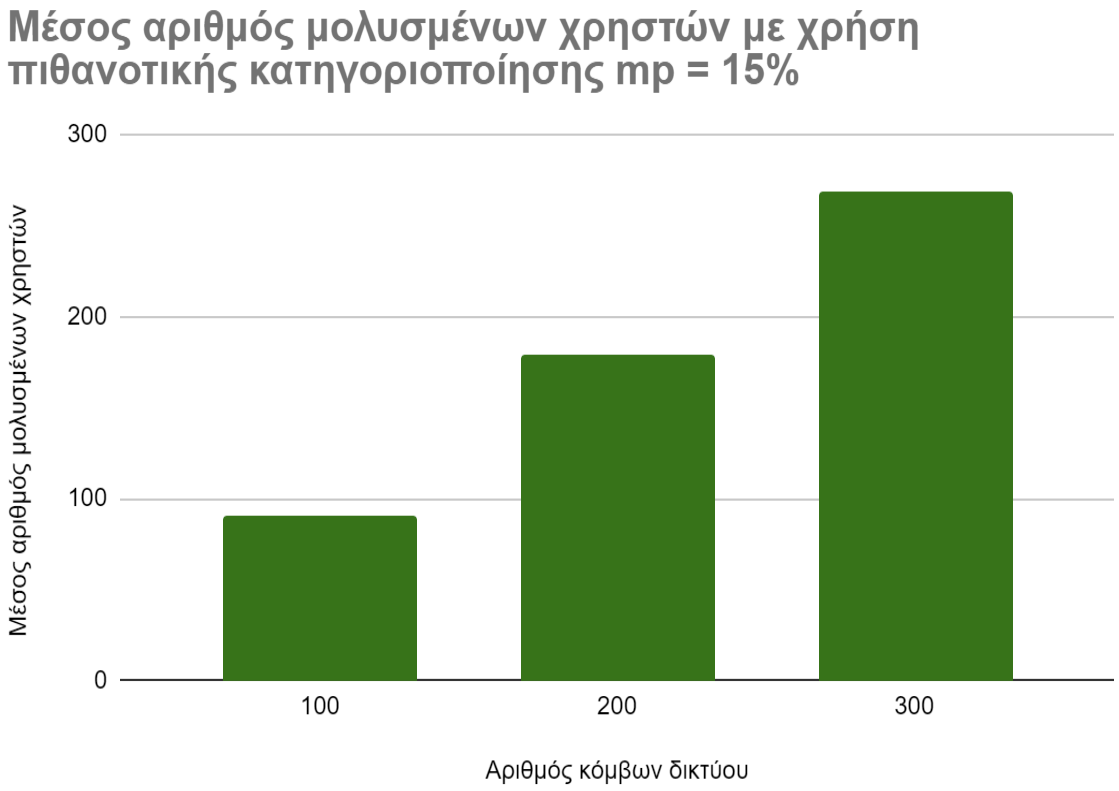


Και στις τρεις περιπτώσεις αριθμού κόμβων δικτύου, φαίνεται πως υπερισχύει το ντετερμινιστικό μοντέλο συμπερασμού και άρα η επιλογή του μονοπατιού με το μεγαλύτερο relevance. Αυτό οφείλεται εν μέρει στο γεγονός πως οι κόμβοι - monitors ισως δεν είναι ομοιόμορφα κατανεμημένοι στο δίκτυο, αφού ως monitors δεν ορίζονται απαραίτητα εκείνοι οι κόμβοι με το μεγαλύτερο βαθμό.



Για το μέσο αριθμό monitoring χρηστών δεν έχει ιδιαίτερη αξία κάποιος σχολιασμός, καθώς αυτός κάθε φορά ισούται με το 15% του συνολικού αριθμού κόμβων του δικτύου.





Όσον αφορά το μέσο αριθμό των χρηστών με ιδιωτικό προφίλ αυτός φαίνεται να διατηρείται επίσης ποσοστιαία σταθερός, κάτι που φαντάζει λογικό λαμβάνοντας υπόψη τη συμπεριφορά των monitoring χρηστών.

Τέλος, φαίνεται και πως ο μέσος αριθμός μολυσμένων χρηστών διατηρεί μία παρόμοια συμπεριφορά, με το συγκεκριμένο ποσοστό να κυμαίνεται περίπου στο 90% των συνολικών χρηστών του δικτύου.

5.2.2 Πιθανοτική κατηγοριοποίηση με ποσοστό monitors επί του δικτύου: 30%

100 δίκτυα με s : 100, mp: 30%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	25.52	26.23	19.80	21.90	16.21	18.17	14.04	15.95
Private profiles	71.48	72.45	76.50	74.91	79.79	77.88	81.96	80.05
Infected users	85.18	87.14	91.42	91.92	91.13	92.25	93.34	98.30
Accuracy of deterministic inference	53.87	54.09	32.28	30.35	21.90	21.73	15.78	16.05
Accuracy of probabilistic inference	51.35	51.97	29.27	28.05	19.14	20.33	13.94	13.92

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 30.76%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 28.50%

100 δίκτυα με s : 200, mp: 30%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	50.85	54.81	39.56	43.31	32.56	26.61	28.02	31.62
Private profiles	142.15	138.19	152.73	149.57	159.44	155.40	163.98	160.38
Infected users	170.12	170.27	177.19	187.67	183.81	182.42	178.15	194.95
Accuracy of deterministic inference	57.32	57.01	34.27	33.33	23.97	23.68	19.27	17.83
Accuracy of probabilistic inference	54.09	55.15	31.30	31.03	21.42	21.26	16.78	15.96

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 33.34%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 30.87%

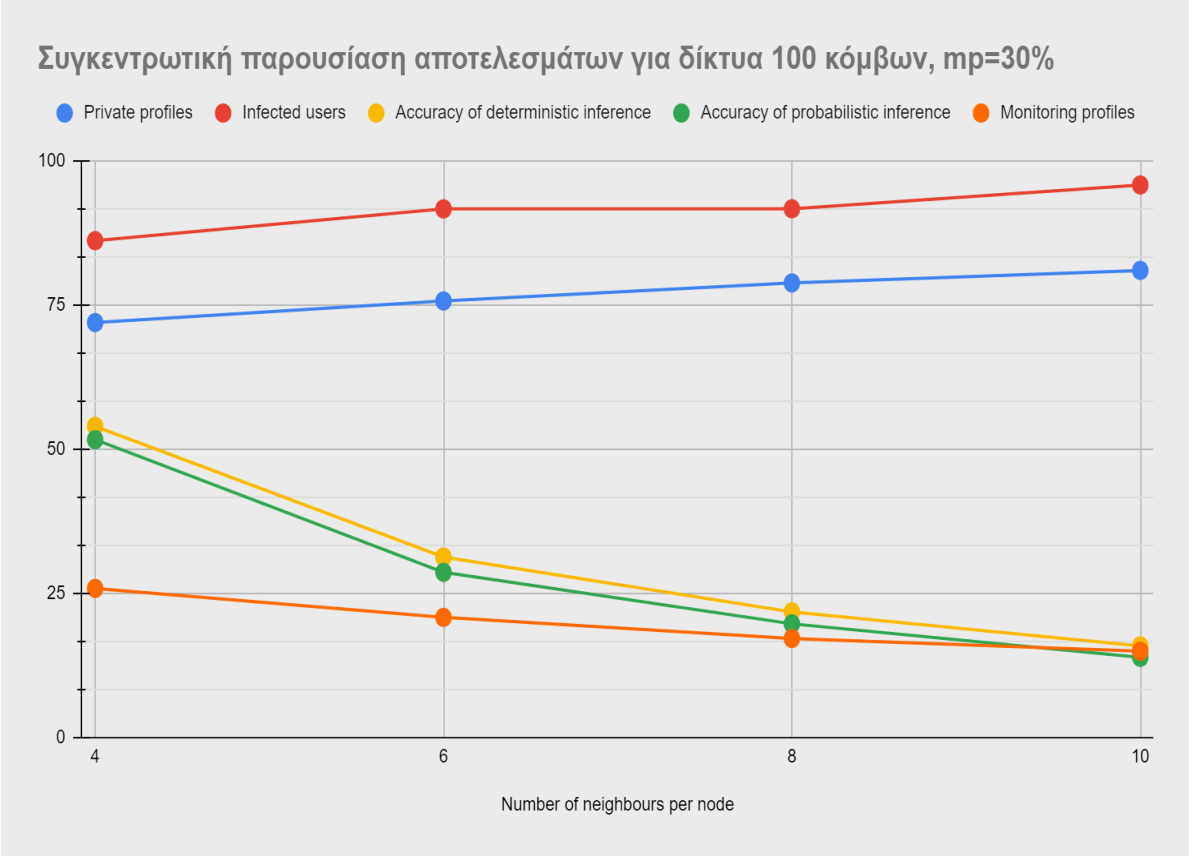
100 δίκτυα με s : 300, mp: 30%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	76.17	82.29	59.18	65.30	49.26	55.20	41.65	47.99
Private profiles	212.86	207.41	229.10	223.68	238.74	232.83	246.06	240.01
Infected users	255.53	253.23	261.54	282.24	283.09	267.45	264.58	288.79
Accuracy of deterministic inference	58.35	60.30	35.96	35.58	24.08	25.29	18.77	18.92
Accuracy of probabilistic inference	55.39	58.19	33.34	33.95	22.27	23.66	16.92	17.02

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 34.66%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 32.59%

Παρατίθεται ο σχολιασμός των αποτελεσμάτων για σταθερή την τιμή του mp = 40%.

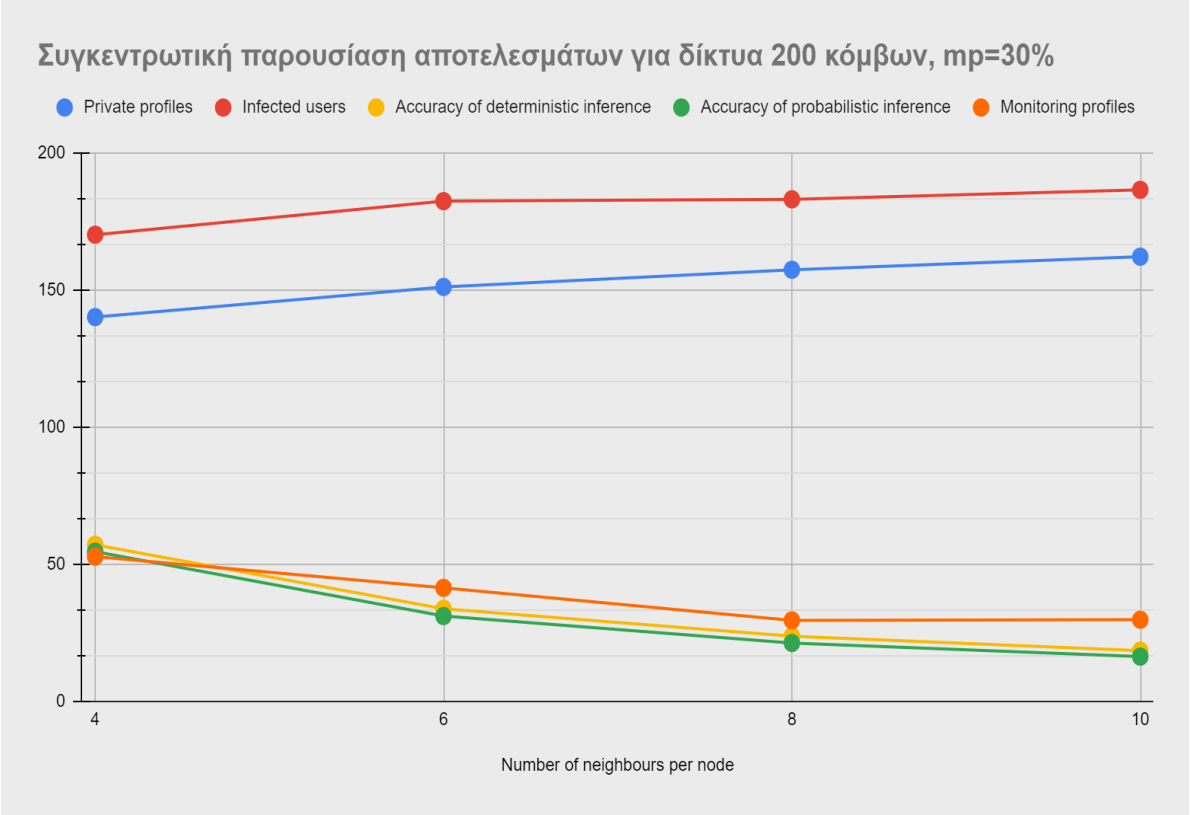
Αριθμός monitoring profiles

Στη συγκεκριμένη περίπτωση παρατηρείται πως ο αριθμός των κόμβων - monitors φτάνει κοντά στο ζητούμενο ποσοστό μόνο για την περίπτωση k = 4, p = 0.5, για οποιαδήποτε τιμή του s. Η απόκλιση αυτή οφείλεται στον περιορισμό ο οποίος επιτρέπει την κατηγοριοποίηση ενός κόμβου ως monitor, μόνο στην περίπτωση που δε συνδέεται ήδη με κάποιο κόμβο - monitor.



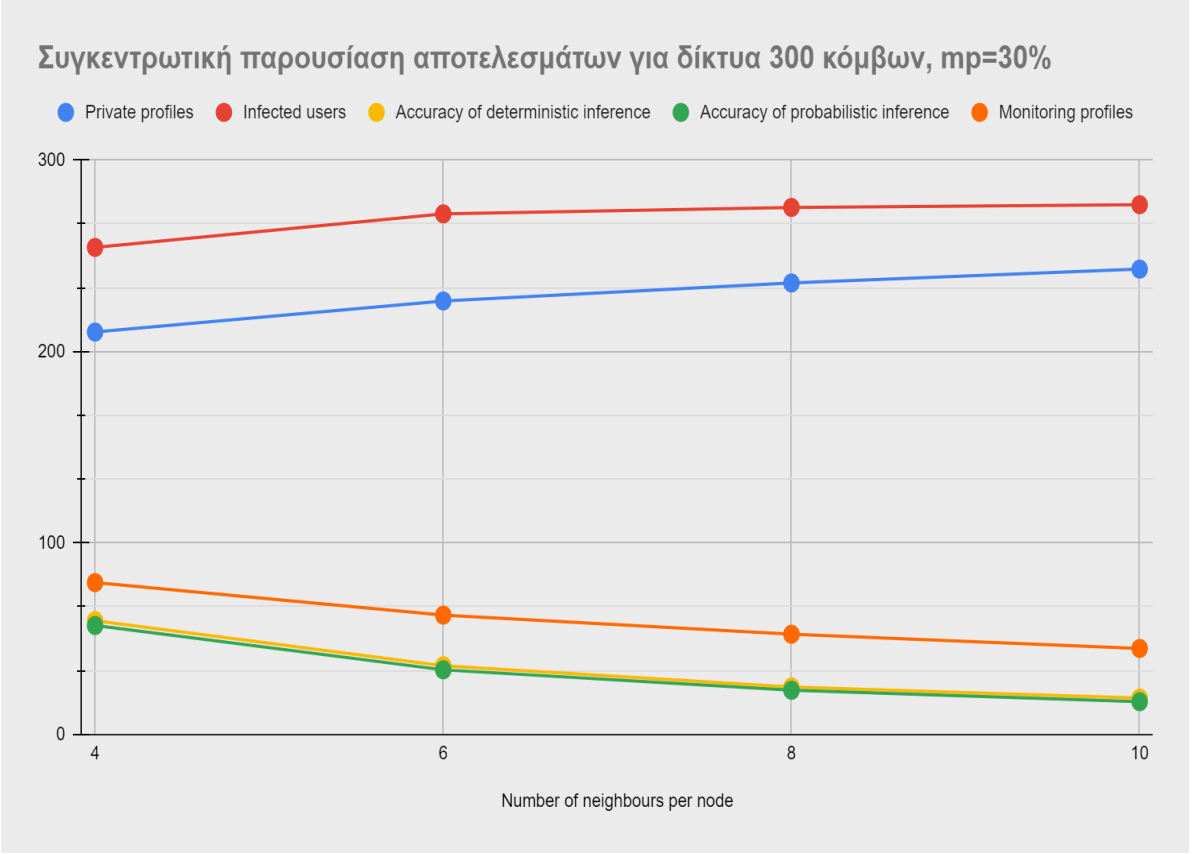
Αριθμός private profiles

Όσον αφορά τον αριθμό των private χρηστών, για την αντίστοιχη αύξηση του p διαφαίνεται μείωση, κυρίως αισθητή για s = 200, s = 300. Αντίστοιχα, με την αντίστοιχη αύξηση του k παρατηρείται αύξηση στον αριθμό των ιδιωτικών λογαριασμών. Όπως και στις προηγούμενες περιπτώσεις, όσο αυξάνεται ο αριθμός των κόμβων με τους οποίους συνδέεται κάθε κόμβος αυξάνεται και ο αριθμός αυτών που συνδέονται με κόμβους οι οποίοι ήδη έχουν κατηγοριοποιηθεί ως monitors. Κατ' επέκταση, είναι μεγαλύτερος ο αριθμός των χρηστών που μπορούν να οριστούν ως private.



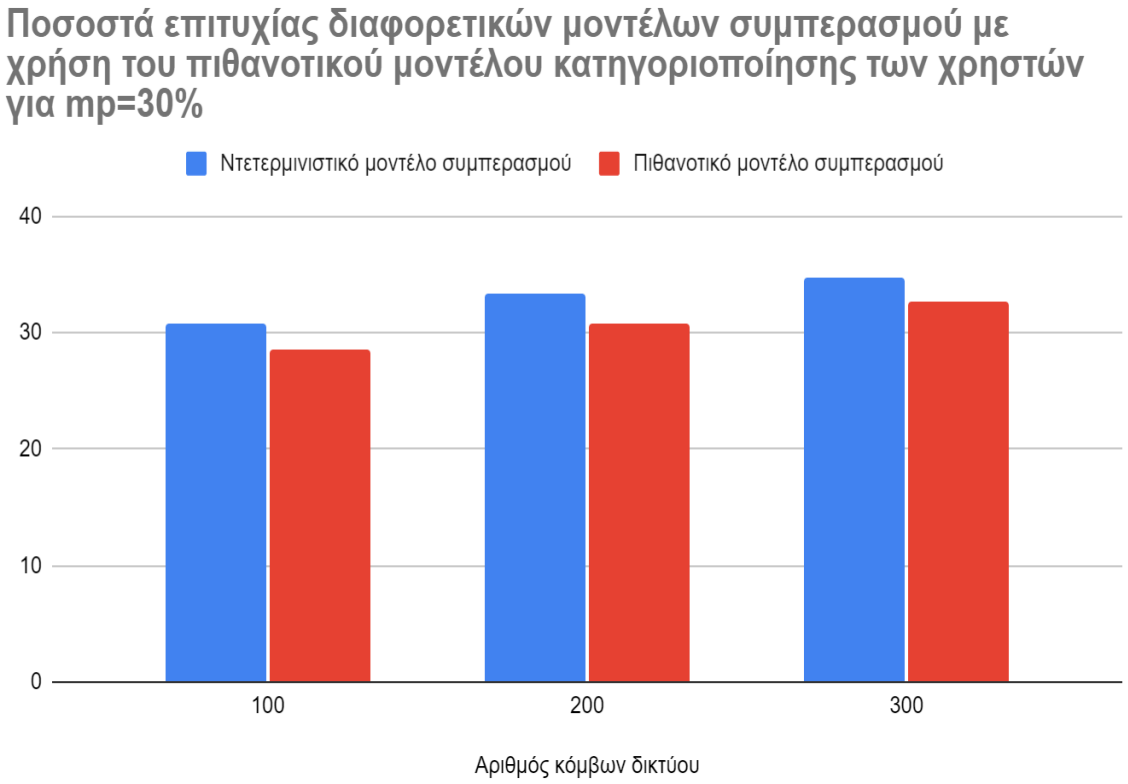
Αριθμός χρηστών στους οποίους έφτασε η πληροφορία

Στις περισσότερες περιπτώσεις φαίνεται πως αυτή η μετρική δεν κρατάει κάποια σταθερή συμπεριφορά σε σχέση με την αύξηση ή μείωση της πιθανότητας να αλλάξει προορισμό μια υπάρχουσα ακμή. Ωστόσο παρατηρείται ένα μοτίβο όσον αφορά τη συγκεκριμένη μετρική σε περίπτωση αύξησης του αριθμού των γειτόνων κάθε κόμβου, στην οποία και ο αριθμός των "μολυσμένων" χρηστών αυξάνεται, γεγονός λογικό αφού μεγαλύτερος αριθμός συνδέσεων εντός του δικτύου συνεπάγεται και περισσότερα μονοπάτια στα οποία μπορεί να κινηθεί η πληροφορία.

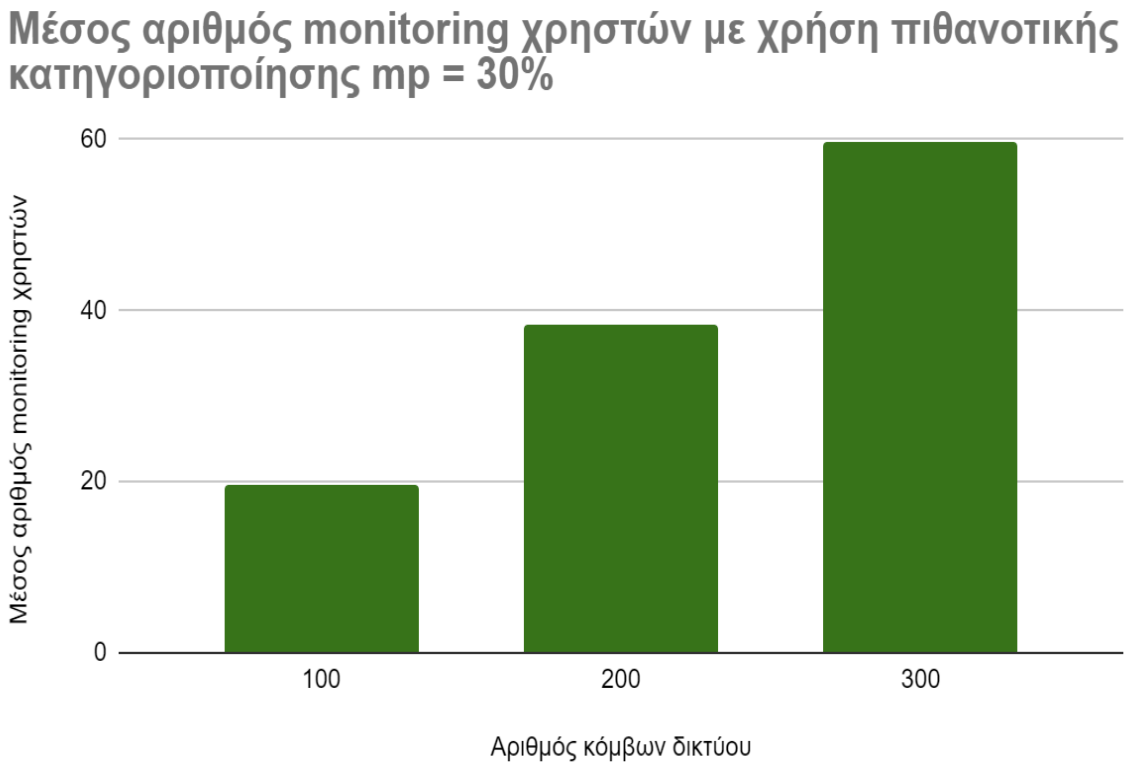


Ποσοστό ακρίβειας συμπερασμού

Παρατηρείται παρόμοια συμπεριφορά όσον αφορά τις δύο διαφορετικές μεθόδους συμπερασμού. Συγκεκριμένα, δε σημειώνονται σημαντικές διαφορές κατά τη μεταβολή του p κρατώντας σταθερά τα s, k. Αντιθέτως, η αύξηση των γειτόνων κάθε κόμβου οδηγεί σε σημαντική μείωση των ποσοστών επιτυχίας. Αυτό οφείλεται στο γεγονός πως περισσότεροι γείτονες για κάθε κόμβο συνεπάγονται και μεγαλύτερο αριθμό υποψηφίων μονοπατιών, κάτι που καθιστά δυσκολότερη τη διαδικασία του συμπερασμού.. Άξιο αναφοράς φαίνεται και το γεγονός της βελτίωσης της απόδοσης του συμπερασμού, όσο αυξάνεται το μέγεθος του δικτύου. Αυτό μπορεί να ερμηνευθεί αν ληφθεί υπόψη πως με την αύξηση του μεγέθους του δικτύου αμβλυνεται η επίδραση που έχει η αύξηση του αριθμού των γειτόνων. Πιο συγκεκριμένα, στην περίπτωση των 100 κόμβων με μετάβαση από k = 4 σε k = 10, κάθε κόμβος συνδέεται με το 4% ή το 10% του δικτύου αντίστοιχα. Αντίθετα, στην περίπτωση των δικτύων 200, 300 κόμβων τα ποσοστά είναι της τάξης του 2% και 5% και άρα η μετάβαση -και άρα και η επίδραση της αύξησης των γειτόνων- ποσοστιαία είναι πολύ μικρότερη.

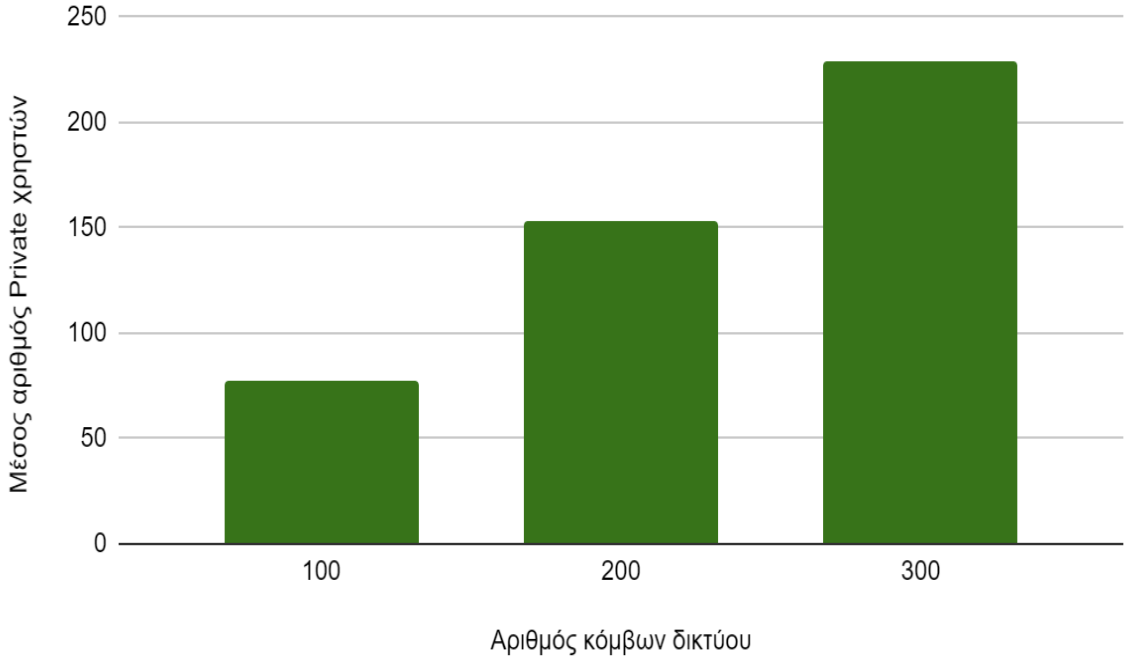


Φαίνεται πως με χρήση του πιθανοτικού μοντέλου κατηγοριοποίησης των χρηστών για mp = 30%, επιτυγχάνονται καλύτερα αποτελέσματα με χρήση του ντετερμινιστικού μοντέλου συμπερασμού. Όπως και στην περίπτωση για mp = 15%, αυτό πιθανώς οφείλεται στο γεγονός πως στην περίπτωση του ντετερμινιστικού μοντέλου επιλέγεται το μονοπάτι με το μεγαλύτερο συνολικό relevance και άρα πιθανά λάθη στα συμπερασμό που θα οφείλονταν στην ασύμμετρη τοποθέτηση των monitors στο δίκτυο, δεν επηρεάζουν το αποτέλεσμα. Επίσης, παρουσιάζεται μικρή αύξηση στα ποσοστά επιτυχίας και των δύο μεθόδων όσο αυξάνεται το μέγεθος του δικτύου, καθώς μεγαλύτερο δίκτυο συνεπάγεται και μεγαλύτερη ευελιξία στην κατηγοριοποίηση κόμβων ως monitors, αφού όσο μικρότερο είναι ένα δίκτυο τόσο πιο επιδραστικός θα είναι και ο περιορισμός που ορίζει πως δε γίνεται να υπάρχουν δύο γείτονες - monitors.

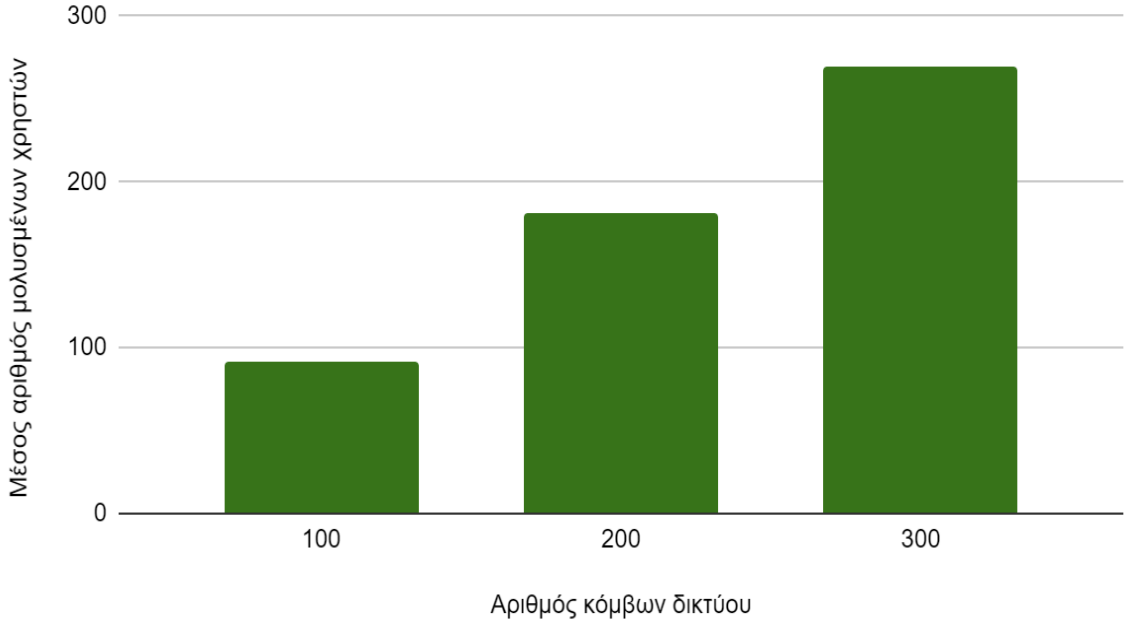


Κάτι που προκύπτει από το παραπάνω διάγραμμα, είναι πως αν και ορίζεται ποσοστό monitoring χρηστών το 30% του δικτύου, αυτό σε καμία από τις τρεις περιπτώσεις δε φαίνεται να επιτυγχάνεται. Είναι προφανές πως για αυτό οφείλεται ο περιορισμός περί μη γειτόνων monitors. Συνδυάζοντάς αυτό με το προηγούμενο διάγραμμα σχετικά με τα ποσοστά επιτυχίας των διαφορετικών μοντέλων συμπερασμού και το πώς αυτά αυξήθηκαν με την αύξηση του μεγέθους του δικτύου, γίνεται εύκολα αντιληπτή αυτή η ευελιξία που παρέχεται από τα μεγαλύτερα δίκτυα, προκειμένου να υπάρχει μεγαλύτερος αριθμός monitors και άρα καλύτερη παρακολούθηση της ροής της πληροφορίας.

Μέσος αριθμός χρηστών με ιδιωτικό λογαριασμό με χρήση πιθανοτικής κατηγοριοποίησης mp=30%



Μέσος αριθμός μολυσμένων χρηστών με χρήση πιθανοτικής κατηγοριοποίησης mp = 30%



Αντίθετα με τον αριθμό monitoring χρηστών, οι μετρικές των χρηστών με ιδιωτικό λογαριασμό και των χρηστών στους οποίους έφτασε η πληροφορία φαίνεται και για mp = 30% να μένουν ανεπηρέαστες από την αύξηση του μεγέθους του δικτύου.

5.2.3 Πιθανοτική κατηγοριοποίηση με ποσοστό monitors επί του δικτύου: 40%

100 δίκτυα με s : 100, mp: 40%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	25.52	27.34	19.80	21.90	16.21	18.17	14.04	15.95
Private profiles	71.48	69.66	76.50	74.91	79.79	77.88	81.96	80.05
Infected users	85.18	86.41	91.42	91.92	91.13	92.25	93.34	98.30
Accuracy of deterministic inference	53.87	54.70	32.28	30.35	21.90	21.73	15.78	16.05
Accuracy of probabilistic inference	51.35	52.73	29.27	28.05	19.14	20.33	13.94	13.92

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 30.83%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 28.59%

100 δίκτυα με s : 200, mp: 40%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	50.85	54.83	39.56	43.31	32.56	36.61	28.02	31.62
Private profiles	142.15	138.19	152.73	149.57	159.44	155.40	163.98	160.38
Infected users	170.12	170.52	177.19	187.67	183.81	182.42	178.15	194.95
Accuracy of deterministic inference	57.32	56.90	34.27	33.33	23.98	23.68	19.27	17.83
Accuracy of probabilistic inference	54.09	55.09	31.30	31.03	21.42	21.26	16.78	15.96

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 33.32%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 30.87%

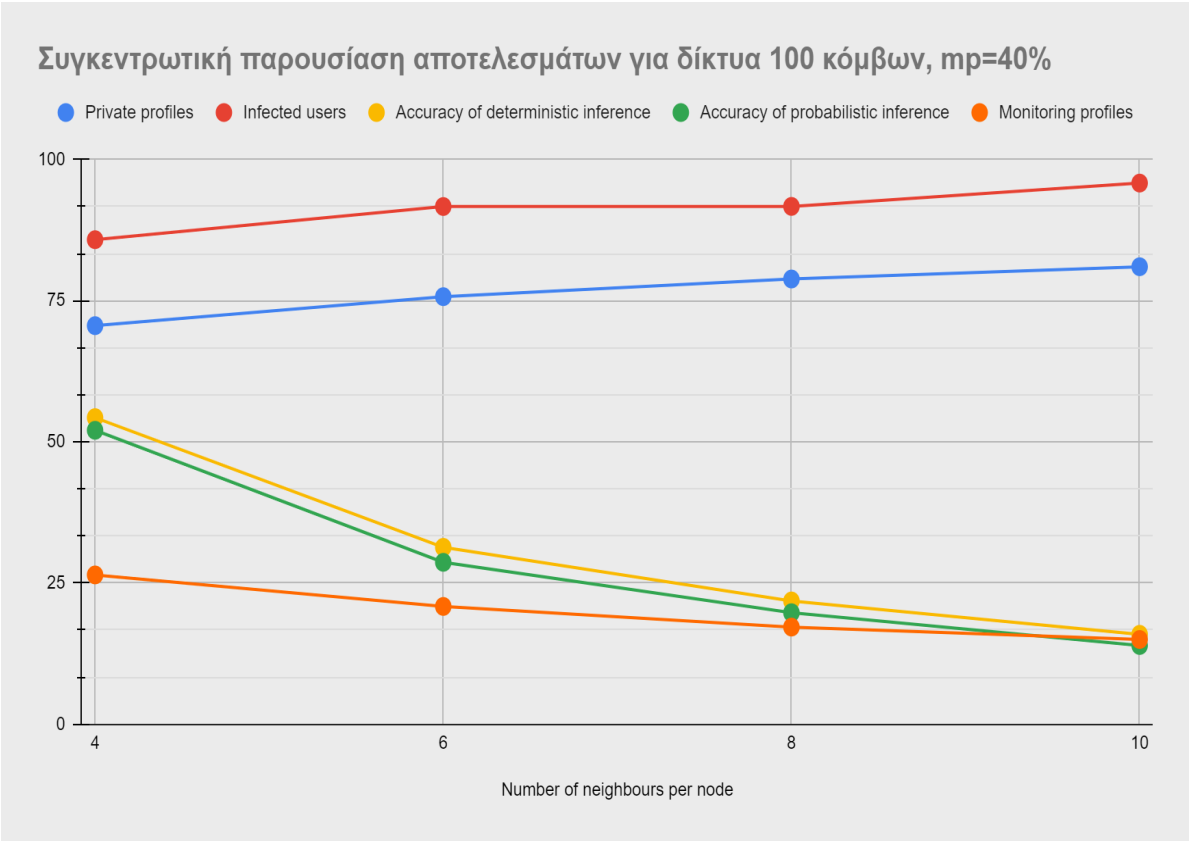
100 δίκτυα με s : 300, mp: 40%								
	k = 4		k = 6		k = 8		k = 10	
	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5	p = 0.2	p = 0.5
Monitoring profiles	76.17	82.30	59.18	65.30	49.26	55.20	41.65	47.99
Private profiles	212.86	207.41	229.10	223.68	238.74	232.83	246.06	240.01
Infected users	255.53	253.37	261.54	282.24	283.09	267.45	264.58	288.79
Accuracy of deterministic inference	58.35	60.30	35.96	35.58	24.08	25.29	18.77	18.92
Accuracy of probabilistic inference	55.38	58.18	33.34	33.95	22.27	23.66	16.92	17.02

Μέσο ποσοστό επιτυχίας ντετερμινιστικού σχήματος συμπερασμού: 34.66%
Μέσο ποσοστό επιτυχίας πιθανοτικού σχήματος συμπερασμού: 32.59%

Παρατίθεται ο σχολιασμός των αποτελεσμάτων για σταθερή την τιμή του mp = 30%.

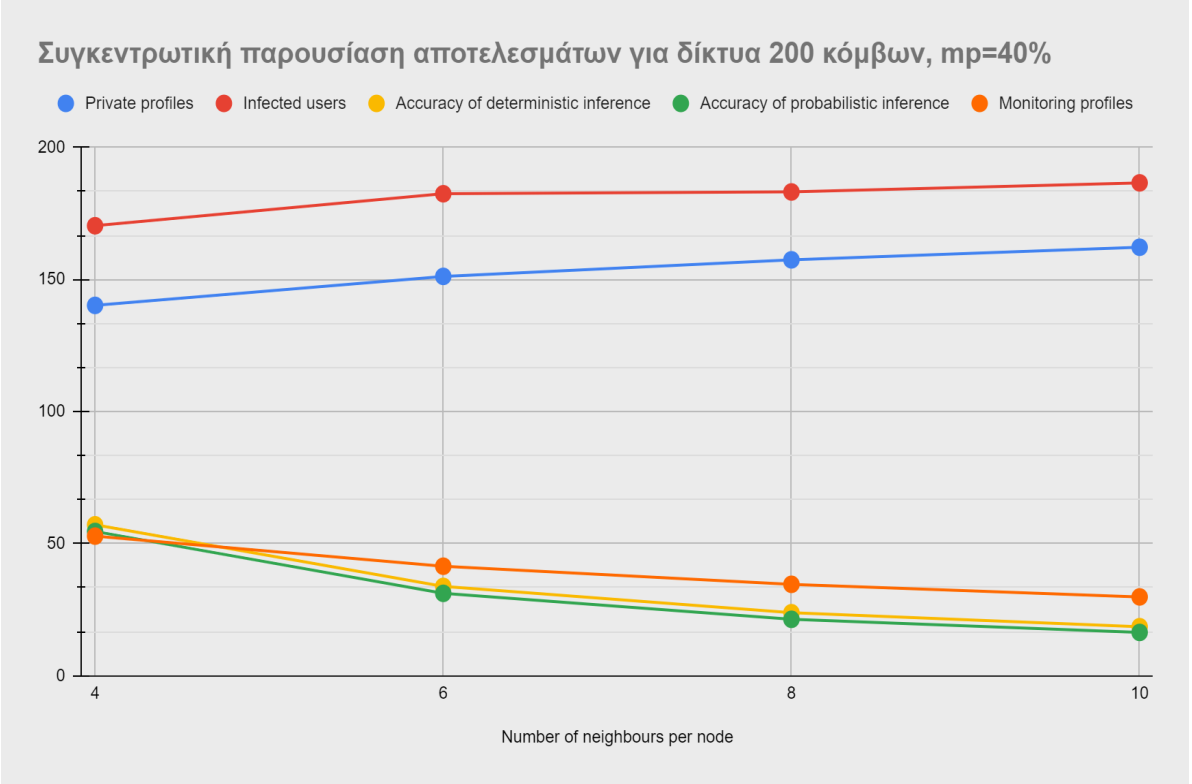
Αριθμός monitoring profiles

Στη συγκεκριμένη περίπτωση παρατηρείται πως ο αριθμός των κόμβων - monitors δε φτάνει κοντά στο ζητούμενο ποσοστό με κανένα συνδυασμό παραμέτρων s, k, p. Αυτό οφείλεται πως παρόλο που "επιτρέπεται" μεγάλο ποσοστό για monitors, το γεγονός πως υφίσταται ο περιορισμός σχετικά με τους γείτονες κάθε monitor συντελεί στη δημιουργία ενός bottleneck για τη συγκεκριμένη μετρική.



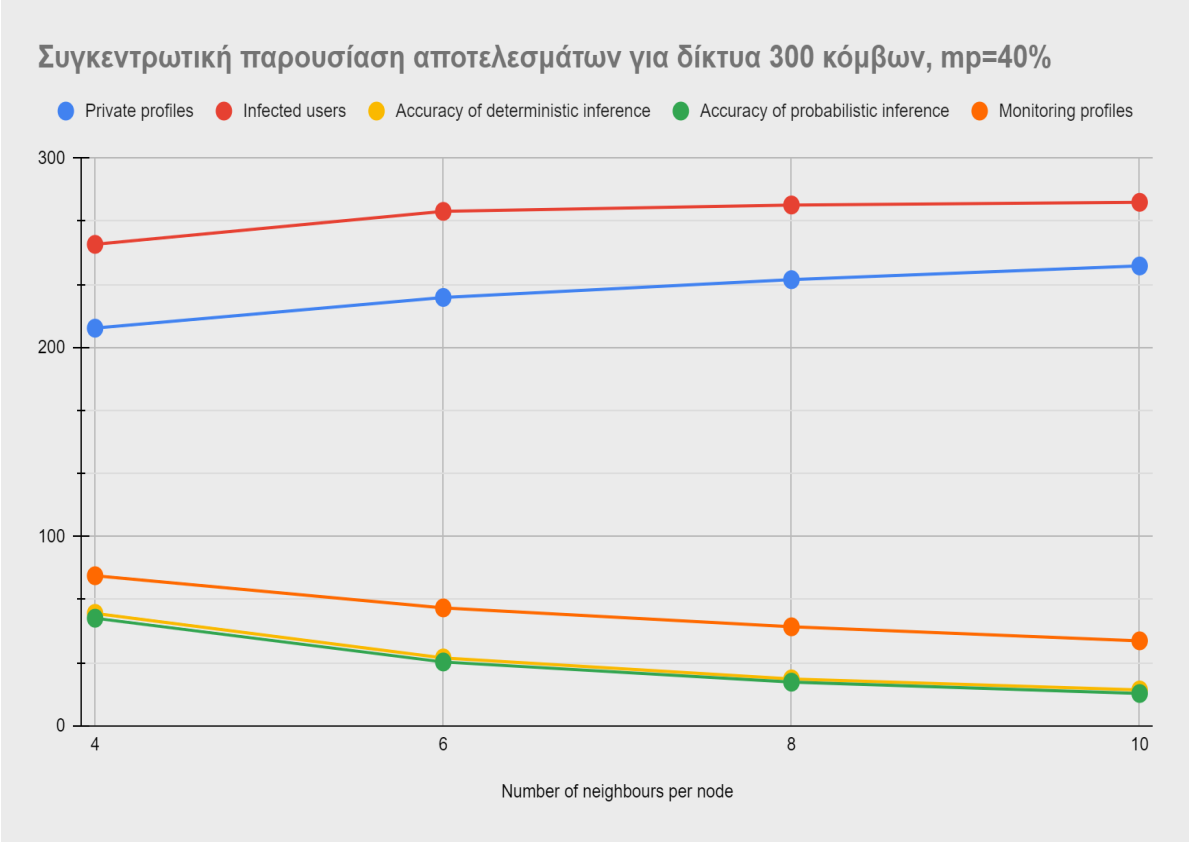
Αριθμός private profiles

Όσον αφορά τον αριθμό των private profiles, για την αντίστοιχη αύξηση του p διαφαίνεται μείωση. Αντίστοιχα, με την αντίστοιχη αύξηση του k παρατηρείται αύξηση στον αριθμό των ιδιωτικών λογαριασμών, κάτι απολύτως λογικό το οποίο συμβαδίζει με τη συμπεριφορά και όλων των προηγούμενων περιπτώσεων δικτύων που εξετάστηκαν.



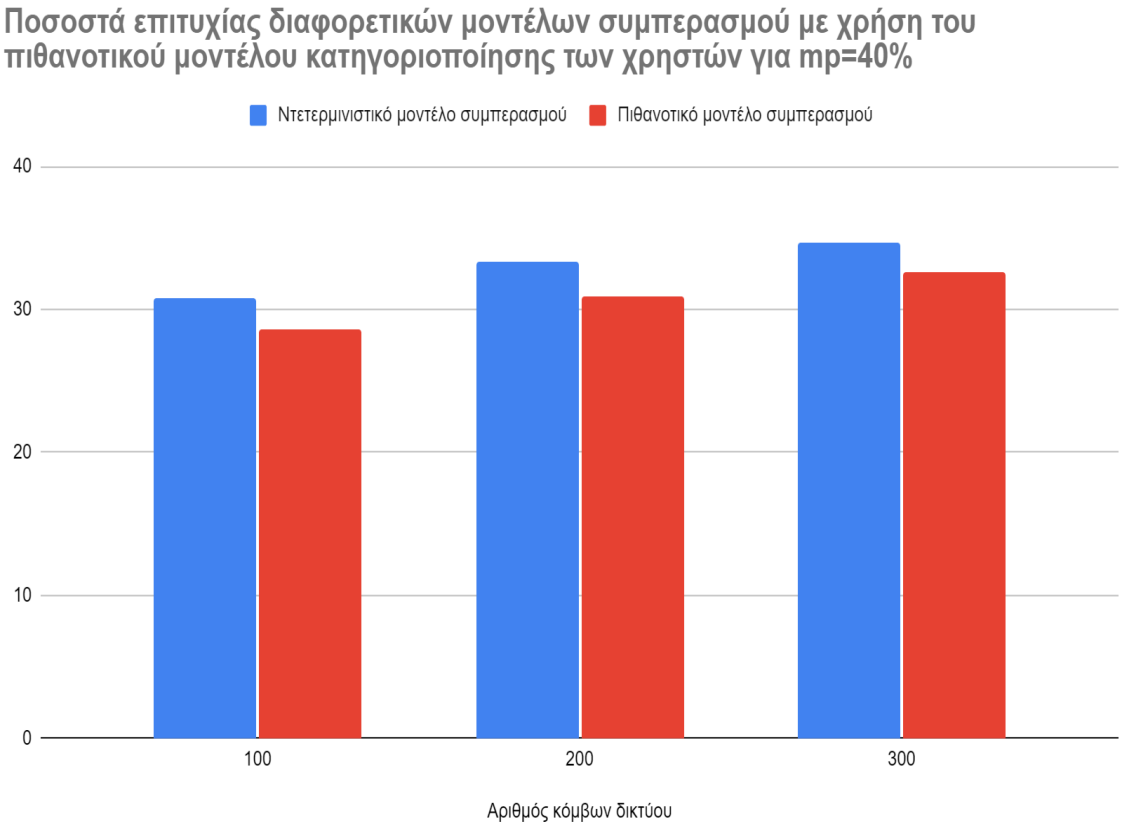
Αριθμός χρηστών στους οποίους έφτασε η πληροφορία

Στις περισσότερες περιπτώσεις φαίνεται πως αυτή η μετρική δεν κρατάει κάποια σταθερή συμπεριφορά σε σχέση με την αύξηση ή μείωση της πιθανότητας να αλλάξει προορισμό μια υπάρχουσα ακμή. Ωστόσο παρατηρείται ένα μοτίβο όσον αφορά τη συγκεκριμένη μετρική σε περίπτωση αύξησης του αριθμού των γειτόνων κάθε κόμβου, στην οποία και ο αριθμός των "μολυσμένων" χρηστών αυξάνεται. Αυτό έχει εξηγηθεί και ανωτέρω, μεγαλύτερος αριθμός γειτόνων για κάθε κόμβο συνεπάγεται και μεγαλύτερες εναλλακτικές διαδρομές για να κινηθεί η πληροφορία.

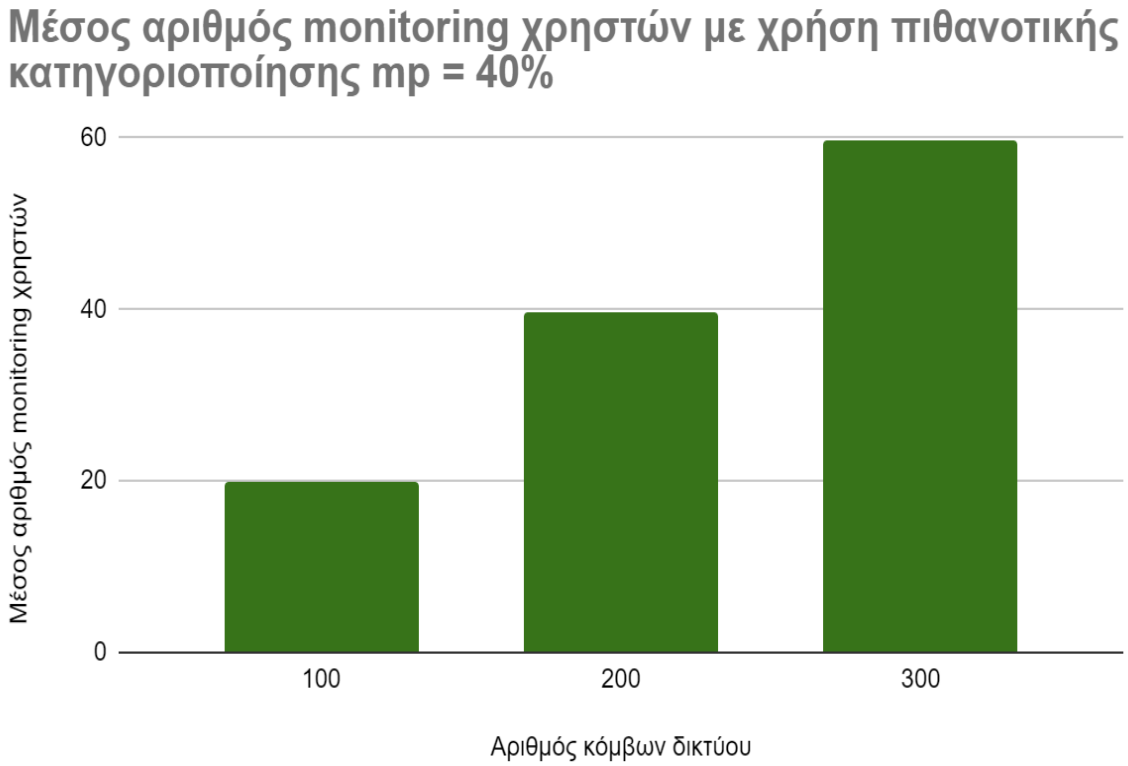


Ποσοστό ακρίβειας συμπερασμού

Παρατηρείται παρόμοια συμπεριφορά όσον αφορά τις δύο διαφορετικές μεθόδους συμπερασμού. Συγκεκριμένα, δε σημειώνονται σημαντικές διαφορές κατά τη μεταβολή του p κρατώντας σταθερά τα s, k. Αντιθέτως, η αύξηση των γειτόνων κάθε κόμβου από τους οδηγεί σε σημαντική μείωση των ποσοστών επιτυχίας. Αξιοσημείωτο είναι και το γεγονός της βελτίωσης της απόδοσης του συμπερασμού, όσο αυξάνεται το μέγεθος του δικτύου. Η συμπεριφορά αυτή είναι παρόμοια με τις προηγούμενες περιπτώσεις δικτύων οι οποίες εξετάστηκαν, και έχει εξηγηθεί ανωτέρω.

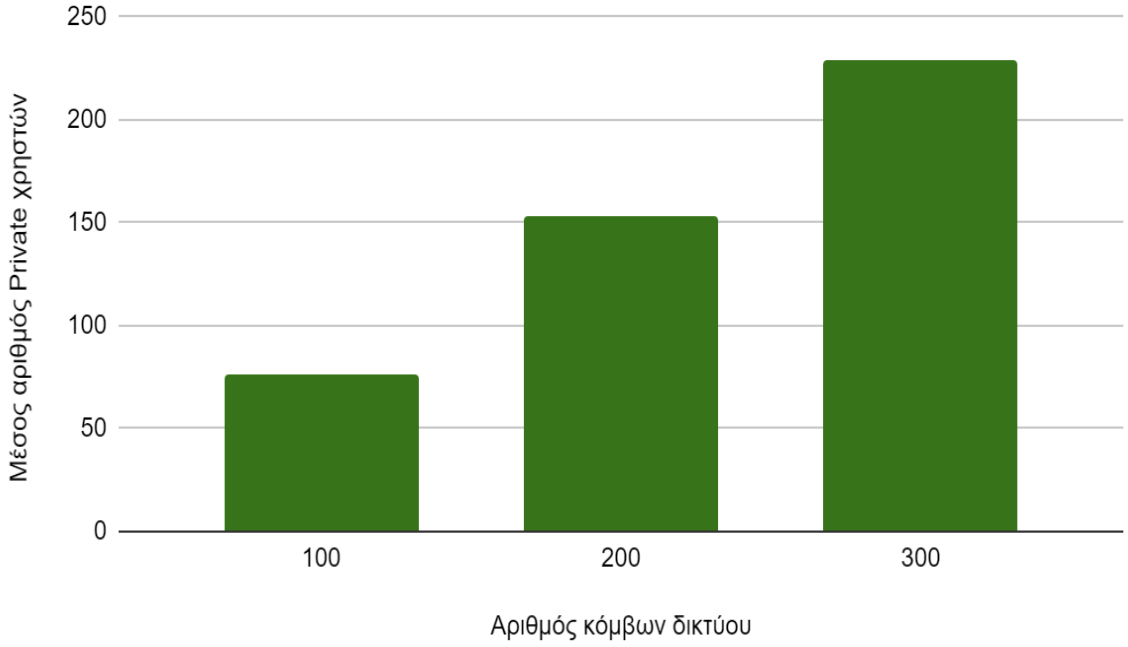


Παρατηρείται παρόμοια συμπεριφορά με την περίπτωση για mp = 40%.

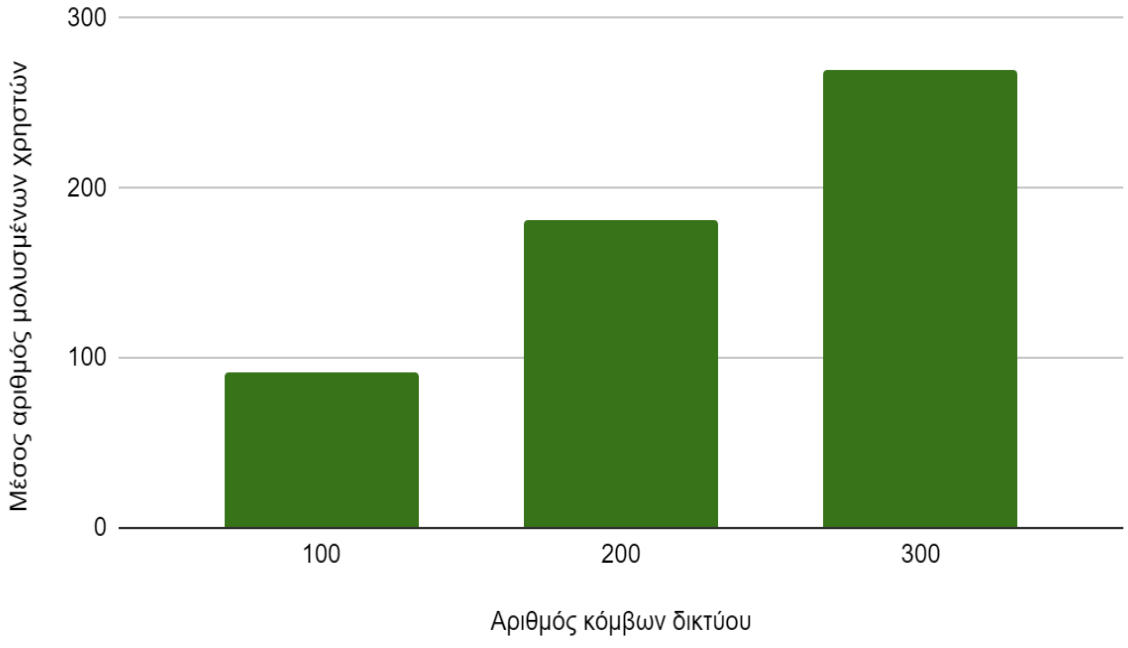


Όπως και στην περίπτωση για mp = 30%, ο αριθμός των monitors περιορίζεται από τον όρο να μη γειτονεύουν δύο κόμβοι και το ποσοστό 40% δεν επιτυγχάνεται για κανένα συνδυασμό παραμέτρων.

Μέσος αριθμός χρηστών με ιδιωτικό λογαριασμό με χρήση πιθανοτικής κατηγοριοποίησης mp = 40%



Μέσος αριθμός μολυσμένων χρηστών με χρήση πιθανοτικής κατηγοριοποίησης mp = 40%

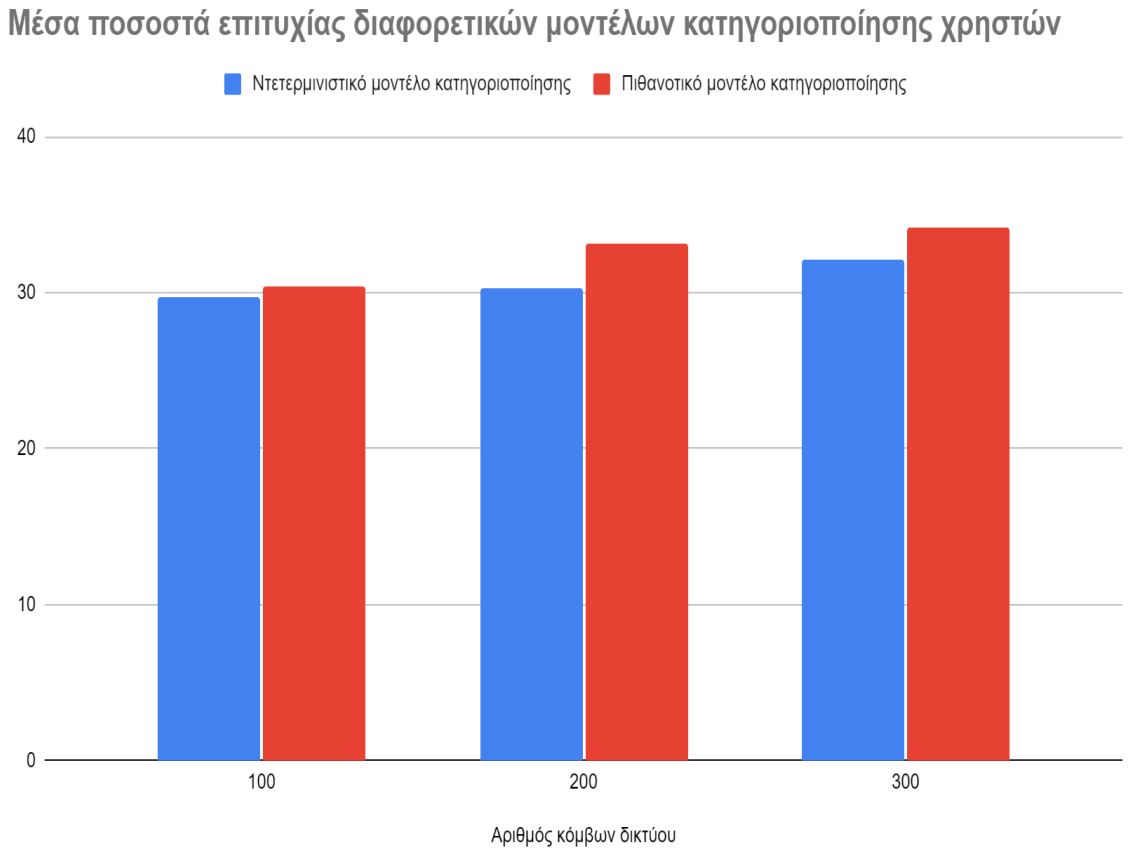


5.3 Σύγκριση των δύο μεθόδων επιλογής των κόμβων - monitors

Όσον αφορά τον αριθμό των monitoring profiles, φαίνεται πως ο περιορισμός του ποσοστού των χρηστών που θα οριστούν ως monitors έχει νόημα μόνο για ποσοστό μικρότερο του 25%, που είναι αυτό που επιτυγχάνεται με την ντετερμινιστική κατηγοριοποίηση. Στις περιπτώσεις στις οποίες δοκιμάστηκε ποσοστό μεγαλύτερο, της τάξης του 30% ή του 40%, δεν επετεύχθη ποτέ καθώς υπερίσχυε ο περιορισμός περί μη γειτονικών κόμβων - monitors.

Σχετικά με τον αριθμό των private profiles και του αριθμού των χρηστών που έλαβαν την πληροφορία, παρατηρούνται παρόμοιες συμπεριφορές και για τους δύο τρόπους κατηγοριοποίησης, με κοινό παρονομαστή την αύξηση των private χρηστών όσο μειώνεται ο αριθμός των χρηστών - monitors.

Όσον αφορά τα ποσοστά επιτυχίας του συμπερασμού οι τιμές είναι παρεμφερείς, με το ντετερμινιστικό μοντέλου συμπερασμού να σημειώνει καλύτερα αποτελέσματα κάτι που οφείλεται στο γεγονός πως επιλέγονται ως monitors οι κόμβοι οι οποίοι έχουν μεγαλύτερο βαθμό.



Τέλος, σχετικά με τα διαφορετικά μοντέλα κατηγοριοποίησης των χρηστών, φαίνεται πως επιτυγχάνονται καλύτερα αποτελέσματα μέσω του πιθανοτικού μοντέλου.

6

Επίλογος

Στο κεφάλαιο αυτό παρουσιάζονται συνοπτικά τα αποτελέσματα της εργασίας και τα συμπεράσματα που προέκυψαν για τις διαφορετικές μεθόδους κατηγοριοποίησης των χρηστών ενός δικτύου καθώς και για τις διαφορετικές μεθόδους συμπερασμού του δικτύου διάχυσης πληροφορίας. Τέλος, παρουσιάζονται πιθανές κατευθύνσεις για την επέκταση της παρούσας εργασίας.

6.1 Σύνοψη

Στην παρούσα διπλωματική εργασία αξιολογήθηκε η δυνατότητα πρόβλεψης του δικτύου διάχυσης πληροφορίας σε ένα Online Social Network (OSN) μέσω ενός σχήματος πιθανοτικής οπισθοδρόμησης, του δικτύου δηλαδή που αποτελείται από τους χρήστες που έλαβαν την πληροφορία, διασφαλίζοντας παράλληλα και την γνώση της πηγής από την οποία την έλαβαν.
Για την εξαγωγή συμπερασμάτων αξιοποιήθηκαν κατηγοριοποιήσεις των χρηστών ανάλογα με το βαθμό στον οποίο υπάρχει πρόσβαση στα δεδομένα τους, καθώς και διαφορετικές μέθοδοι για την κατηγοριοποίηση αυτή των χρηστών του OSN. Τέλος, για τη διαδικασία του συμπερασμού εξετάστηκαν δύο προσεγγίσεις, μία πιθανοτική και μια ντετερμινιστική.

6.2 Γενικά συμπεράσματα

Όσον αφορά την κατηγοριοποίηση των χρηστών σε αυτούς που επιτρέπουν πρόσβαση στα δεδομένα τους, αυτούς που δεν επιτρέπουν και αυτους που χρησιμοποιούνται ως monitors, προκύπτει πως η κατηγοριοποίησή τους με πιθανοτικό τρόπο συνεισφέρει στην πρόβλεψη του δικτύου διάχυσης πληροφορίας με μεγαλύτερη ακρίβεια.

Αντίθετα, φαίνεται πως το ντετερμινιστικό μοντέλο συμπερασμού του δικτύου διάχυσης επιστρέφει ελαφρά καλύτερα αποτελέσματα από τη μέθοδο συμπερασμού που στηρίζεται στο πιθανοτικό μοντέλο.

6.3 Μελλοντικές επεκτάσεις

Το αντικείμενο της παρούσας διπλωματικής εργασίας θα μπορούσε ενδεικτικά να επεκταθεί στις εξής κατευθύνσεις:

- Αξιοποίηση του δικτύου διάχυσης πληροφορίας για τη σύσταση ενός αποδοτικού εξατομικευμένου συστήματος συστάσεων (recommendation system).
- Προώθηση συγκεκριμένων προϊόντων και υπηρεσιών εντός ενός OSN αξιοποιώντας το δίκτυο διάχυσης πληροφορίας.
- Έλεγχος ορθότητας πληροφοριών εντός ενός OSN αξιολογώντας την πηγή από την οποία αυτές προήλθαν.

Βιβλιογραφία

[1] Albert, Réka, and Albert-László Barabási. "Statistical mechanics of complex networks." Reviews of modern physics 74.1 (2002): 47.

[2] Vitoropoulou, M., Karyotis, V. & Papavassiliou, S. Sensing and monitoring of information diffusion in complex online social networks. Peer-to-Peer Netw. Appl. 12, 604–619 (2019).

[3] Bang-Jensen, Jørgen, and Gregory Z. Gutin. Digraphs: theory, algorithms and applications. Springer Science & Business Media, 2008.

[4] Weisstein, Eric W. "Minimum spanning tree." https://mathworld. wolfram. com/ (2000).

[5] Chartrand, Gary. Introductory graph theory. Courier Corporation, 1977.

[6] Karyotis, Vasileios, Eleni Stai, and Symeon Papavassiliou. Evolutionary dynamics of complex communications networks. CRC Press, 2014.

[7] Bender, Edward A., and S. Gill Williamson. Lists, decisions and graphs. S. Gill Williamson, 2010.

[8] Bender, Edward A., and S. Gill Williamson. Lists, decisions and graphs. S. Gill Williamson, 2010.

[9] Cahn, Robert. Wide area network design: concepts and tools for optimization. Morgan Kaufmann, 1998.

[10] Bollobás, Béla. Modern graph theory. Vol. 184. Springer Science & Business Media, 1998.

[11] Bollobás, Béla. Modern graph theory. Vol. 184. Springer Science & Business Media, 1998.

[12] Bender, Edward A., and S. Gill Williamson. Lists, decisions and graphs. S. Gill Williamson, 2010.

[13] Graham, Ronald L., and Pavol Hell. "On the history of the minimum spanning tree problem." Annals of the History of Computing 7.1 (1985): 43-57.

[14] Graham, Ronald L., and Pavol Hell. "On the history of the minimum spanning tree problem." Annals of the History of Computing 7.1 (1985): 43-57.

[15] Lewis, Harry R. "Michael R. ΠGarey and David S. Johnson. Computers and intractability. A guide to the theory of NP-completeness. WH Freeman and Company, San Francisco1979, x+ 338 pp." The Journal of Symbolic Logic 48.2 (1983): 498-500.

[16] Karyotis, Vasileios, Eleni Stai, and Symeon Papavassiliou. Evolutionary dynamics of complex communications networks. CRC Press, 2014.

[17] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of ‘small-world’networks." nature 393.6684 (1998): 440-442.

[18] Vitoropoulou, Margarita. Socio-aware content allocation in complex networks via efficient monitoring of information dissemination. Diss. Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ). Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών. Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής. Εργαστήριο Διαχείρισης και Βέλτιστου Σχεδιασμού Δικτύων Τηλεματικής (NETMODE), 2021.

[19] Zhang, Huiyuan, et al. "Recent advances in information diffusion and influence maximization in complex social networks." Opportunistic Mobile Social Networks 37.1.1 (2014): 37.

[20] Cheng, Shin-Ming, et al. "Diffusion models for information dissemination dynamics in wireless complex communication networks." Journal of Complex Systems 2013 (2013).

[21] Stai, Eleni, et al. "Strategy evolution of information diffusion under time-varying user behavior in generalized networks." Computer Communications 100 (2017): 91-103.

[22] Banerjee, Suman, Mamata Jenamani, and Dilip Kumar Pratihar. "A survey on influence maximization in a social network." Knowledge and Information Systems 62.9 (2020): 3417-3455.

[23]Guille, Adrien, et al. "Information diffusion in online social networks: A survey." ACM Sigmod Record 42.2 (2013): 17-28.

[24] Kleinberg, J. "Bursty and hierarchical structure in streams, data mining and knowledge discovery." elected Papers from the 8th ACM SIGKDD Int. Conf. on Knowledge I Discovery and Data Mining? Part. Vol. 7. No. 4. 2002

[25] Elder IV, John John F., et al. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery (ACM), 2009.

[26] Lu, Rong, and Qing Yang. "Trend analysis of news topics on twitter." International Journal of Machine Learning and Computing 2.3 (2012): 327.

[27] Gomez-Rodriguez, Manuel, Jure Leskovec, and Andreas Krause. "Inferring networks of diffusion and influence." ACM Transactions on Knowledge Discovery from Data (TKDD) 5.4 (2012): 1-37.

[28] Rodriguez, Manuel Gomez, et al. "Uncovering the structure and temporal dynamics of information propagation." Network Science 2.1 (2014): 26-65.

[29] Yang, Xiao, et al. "Recommender system-based diffusion inferring for open social networks." IEEE Transactions on Computational Social Systems 7.1 (2019): 24-34.

[30] Lovász, László. "Random walks on graphs." Combinatorics, Paul erdos is eighty 2.1-46 (1993): 4.

[31] Masuda, Naoki, Mason A. Porter, and Renaud Lambiotte. "Random walks and diffusion on networks." Physics reports 716 (2017): 1-58.

[32] Karyotis, Vasileios, and M. H. R. Khouzani. Malware diffusion models for modern complex networks: theory and applications. Morgan Kaufmann, 2016.

[33] Cercel, Dumitru-Clementin, and Stefan Trausan-Matu. "Opinion propagation in online social networks: A survey." Proceedings of the 4th international conference on web intelligence, mining and semantics (WIMS14). 2014.

[34] Katz, Leo. "A new status index derived from sociometric analysis." Psychometrika 18.1 (1953): 39-43.

[35] Tsitseklis, Konstantinos, et al. "Socio-Aware Recommendations Under Complex User Constraints." IEEE Transactions on Computational Social Systems 8.2 (2021): 377-387.

[36] Domingos, Pedro, and Matt Richardson. "Mining the network value of customers." Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001.

[37] Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003.

[38] Peng, Sancheng, et al. "Influence analysis in social networks: A survey." Journal of Network and Computer Applications 106 (2018): 17-32.

[39] Zhuang, Honglei, et al. "Influence maximization in dynamic social networks." 2013 IEEE 13th International Conference on Data Mining. IEEE, 2013.

[40] Stai, Eleni, Vasileios Karyotis, and Symeon Papavassiliou. "User interest dictated information diffusion over generalized networks." 2015 IEEE International Conference on Communication Workshop (ICCW). IEEE, 2015.

[41] Shakarian, Paulo, et al. "The independent cascade and linear threshold models." Diffusion in Social Networks. Springer, Cham, 2015. 35-48.