ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής

Μοντέλα και Τεχνικές Διάδοσης Πληροφορίας σε Ηλεκτρονικά Κοινωνικά Δίκτυα: Η Περίπτωση του Twitter

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ειρήνης Κ. Μηλαίου

Επιβλέπων: Συμεών Χρ. Παπαβασιλείου
Καθηγητής, ΕΜΠ

Αθήνα, Μάρτιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Επικοινωνιών, Ηλεκτρονικής και Συστημάτων Πληροφορικής

Μοντέλα και Τεχνικές Διάδοσης Πληροφορίας σε Ηλεκτρονικά Κοινωνικά Δίκτυα: Η Περίπτωση του Twitter

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ειρήνης Κ. Μηλαίου

Επιβλέπων: Συμεών Χρ. Παπαβασιλείου
Καθηγητής, ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Μαρτίου 2015.

.....	.....	.....
Συμεών Παπαβασιλείου	Μιχαήλ Θεολόγου	Ευστάθιος Συκάς
Καθηγητής, ΕΜΠ	Καθηγητής, ΕΜΠ	Καθηγητής, ΕΜΠ

Αθήνα, Μάρτιος 2015.

.....

Ειρήνη Κ. Μηλαίου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Copyright© Ειρήνη Κ. Μηλαίου , 2015
Με επιφύλαξη παντός δικαιώματος. *All rights reserved.*

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας διπλωματικής εργασίας εξ' ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς το συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Τα κοινωνικά δίκτυα διαδραματίζουν σπουδαίο ρόλο στην επικοινωνία και στη διάδοση της πληροφορίας, καθώς επιτυγχάνουν την ενημέρωση και την αλληλεπίδραση των ανθρώπων με πολύ ταχύτερους ρυθμούς σε σχέση με τις παραδοσιακές μεθόδους. Ταυτόχρονα, η συνεχής αύξηση της δημοτικότητας τους οδηγεί όλο και περισσότερους οργανισμούς και επιχειρήσεις να τα χρησιμοποιούν για εμπορικούς σκοπούς. Για τους παραπάνω λόγους συνεχώς αυξάνεται το ενδιαφέρον για τον τρόπο και το ρυθμό διάδοσης των πληροφοριών μέσω αυτών καθώς και το εύρος στο οποίο έχουν τη δυνατότητα να διασκορπίσουν τις νέες αναδυόμενες πληροφορίες.

Στην παρούσα εργασία γίνεται μελέτη της δυναμικής των ηλεκτρονικών κοινωνικών δικτύων και κυρίως του δικτύου του Twitter. Στόχος της είναι να διαμορφωθεί ένα μοντέλο που να αποτυπώνει τη διάχυση της πληροφορίας στο συγκεκριμένο κοινωνικό δίκτυο και γίνεται επαλήθευση αυτού μέσω των πειραματικών αποτελεσμάτων με δεδομένα που έχουν συλλεχθεί χάρη στις υπηρεσίες που προσφέρει το Twitter. Παράλληλα, αναπτύσσεται και αναλύεται θεωρητικά ένα επιδημιολογικό μοντέλο διάχυσης της πληροφορίας, ώστε να υπολογιστεί ο αριθμός των χρηστών που ενημερώνονται για συγκεκριμένα θέματα. Επιπρόσθετα, εξετάζεται ο ρυθμός με τον οποίο οι χρήστες ενημερώνονται για νέες ειδήσεις που δημοσιεύονται στο δίκτυο και εξετάζεται η κατανομή που ακολουθούν οι χρόνοι που μεσολαβούν μεταξύ των μηνυμάτων που δέχονται στην κεντρική τους σελίδα. Μέσω πειραματικών αποτελεσμάτων δείχνεται ότι η πιο κατάλληλη κατανομή είναι η generalized pareto. Τέλος, συνοψίζονται τα αποτελέσματα της διπλωματικής και δίνονται κατευθύνσεις για μελλοντική μελέτη.

Λέξεις Κλειδιά κοινωνικά δίκτυα, σύνθετα δίκτυα, Twitter, διάχυση πληροφορίας, επιδημιολογικό μοντέλο

Abstract

Social networks play an important role in communications and in the diffusion of information, given that they succeed in informing people in a faster way than the way in which traditional means of information do. Additionally, the increasing popularity that they experience leads more and more organisations and businesses to use them for commercial purposes. For all these reasons, there is an increasing interest in the way in which information can be diffused and the extent in which new emerging trends can be spread.

This diploma thesis focuses on the study of the dynamic nature of the online social networks and especially, it focuses on the network of Twitter. The objective is to form an epidemic model describing the diffusion of the information in social networks such as Twitter and to verify this model via software based experimentation with data collected from Twitter. Furthermore, we develop and analyse a model of the number of users of Twitter that get informed about new trends and we obtain knowledge on how trending topics can be spread throughout Twitter. Additionally, this work provides information on the rate with which the users of the network get informed about new posts and emerging trends and studies the distribution of the inter arrival times of the incoming posts. The distribution which fits best the data is the generalized pareto. Finally, summing up the results of the theses and giving directions for future study.

Key Words social networks, complex ntworks, Twitter, information diffusion, epidemic model

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω θερμά τον Καθ. κ. Συμεών Παπαβασιλείου που μου έδωσε τη δυνατότητα να πραγματοποιήσω την παρούσα διπλωματική εργασία υπό την επίβλεψη και καθοδήγησή του.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τον Δρ. Βασίλειο Καρυώτη, τη Δρ. Έλενα Στάη και τη Βασιλική Πούλη για την αμέριστη βοήθειά τους και τη συνεχή καθοδήγησή τους καθόλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας.

Θα ήθελα ακόμα να ευχαριστήσω την οικογένεια μου και τους φίλους μου για την υποστήριξη και τη συμπαράσταση τους αυτά τα χρόνια φοίτησης μου στο ΕΜΠ.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Κεφάλαιο
Εισαγωγή

Η έννοια της κοινωνικής δικτύωσης, αν και υφίσταται πολύ πριν από την ύπαρξη του Διαδικτύου, σήμερα γνωρίζει μεγάλη άνθιση λόγω της πλήρους εκμετάλλευσης των δικτύων γενικά, π.χ. το Διαδίκτυο. Ξεκινά από τα τέλη του 1890, όπου οι Émile Durkheim και Ferdinand Tönnies σκιαγράφησαν την ιδέα της κοινωνικής δικτύωσης για τη μελέτη των κοινωνικών ομάδων. Επισήμαναν ότι οι κοινωνικές ομάδες μπορούν να υπάρχουν μέσω των προσωπικών συνδέσεων και να ενώνουν άτομα με κοινά χαρακτηριστικά και ενδιαφέροντα. Επιπλέον παρατηρήθηκε εξέλιξη, καθώς διαφορετικά επιστημονικά πεδία, όπως η ψυχολογία, η ανθρωπολογία και η επιστήμη των μαθηματικών, ασχολήθηκαν και μελέτησαν αυτή την ιδέα. Σήμερα η κοινωνική δικτύωση αναφέρεται κυρίως στην εξάπλωση των επαγγελματικών και κοινωνικών δεσμών ενός ατόμου μέσω του σχηματισμού δεσμών με άλλα άτομα. Στην ανάλυση των κοινωνικών δικτύων επιδιώκεται η μελέτη των χαρακτηριστικών των δικτύων, όπως η συμπεριφορά των χρηστών και οι σχέσεις μεταξύ των χρηστών, ώστε να εξαχθούν υφιστάμενα μοτίβα δομής των δικτύων, να περιγραφεί και να μελετηθεί η ροή της πληροφορίας και να ανακαλυφθεί η επιρροή των διασυνδέσεων πάνω στους ανθρώπους και στους διάφορους οργανισμούς, καθώς και το αντίστροφο.

Η ενασχόληση με τα ηλεκτρονικά (online) κοινωνικά δίκτυα αποτελεί ένα σημαντικό κομμάτι καθημερινής ενασχόλησης για εκατομμύρια ανθρώπους σε όλο τον κόσμο. Κατά συνέπεια δημιουργήθηκαν πολλές ιστοσελίδες που εξυπηρετούν την κοινωνική δικτύωση μέσω του διαδικτύου. Μέσω αυτών παρέχεται η δυνατότητα σε άτομα να διευρύνουν τις κοινωνικές τους σχέσεις και να δημιουργήσουν φιλίες τόσο στα πλαίσια της τοπικής τους κοινωνίας όσο και σε διαφορετικές κοινωνίες σε όλο τον κόσμο. Επιπλέον μέσω των ηλεκτρονικών κοινωνικών δικτύων γίνεται η ενημέρωση των χρηστών για πρόσφατα γεγονότα και για σημαντικές ή λιγότερο σημαντικές πληροφορίες, τις οποίες άλλοι κοινοποιούν.

Παραδείγματα κοινωνικών δικτύων, τα οποία αποτελούν αναπόσπαστο κομμάτι της καθημερινής ζωής, τόσο γιατί συνεισφέρουν στην επικοινωνία, όσο και στην διαφήμιση και στο εμπόριο, είναι το Facebook, Linkedin και το Twitter [7], [8], [9].

Η μελέτη των κοινωνικών δικτύων έχει απασχολήσει την ερευνητική ιδιαίτερα τα τελευταία χρόνια, καθώς αποτελούν ένα σημαντικό τομέα στην ενημέρωση, στην επικοινωνία, στο εμπόριο και στη διαφήμιση. Η μελέτη της διάχυσης της πληροφορίας σε τέτοια δίκτυα αποτελεί ιδιαίτερα ενδιαφέρον και χρήσιμο αντικείμενο, εφόσον μπορούν εξαχθούν συμπεράσματα που θα βοηθήσουν στην κατανόηση της δυναμικής που ακολουθεί η πληροφορία και το εύρος που μπορεί να καλύψει. Τα παραπάνω αποτελέσματα αποτελούν σημαντικά στοιχεία για τη διαδικτυακή διαφήμιση και το marketing μέσω κοινωνικών δικτύων καθώς και για την αποφυγή εξάπλωσης κακόβουλης πληροφορίας. Γνωρίζοντας τη συμπεριφορά του δικτύου και των χρηστών μπορεί να προβλεφθεί η διάχυση πληροφορίας και να παρθούν μέτρα για την ασφάλεια του δικτύου και την προστασία από τυχούσες επιθέσεις σε τρωτά σημεία του δικτύου.

1.1 Συμβολή

Στην παρούσα διπλωματική εργασία μελετάται το κοινωνικό δίκτυο Twitter. Ενδιαφέρον τμήμα της μελέτης του online κοινωνικού δικτύου Twitter είναι η διάδοση των πληροφοριών μέσα σε αυτό από χρήστη σε χρήστη καθώς και η έκταση στην οποία μπορεί να γίνει γνωστή μια πληροφορία. Με βάση τα αντίστοιχα μοντέλα στον τομέα της επιδημιολογίας δημιουργείται ένα μοντέλο για τη διάδοση της πληροφορίας από 'μολυσμένους' σε 'υγιείς' χρήστες λαμβάνοντας υπόψιν τα χαρακτηριστικά του δικτύου και των χρηστών που το απαρτίζουν[1]. 'Μολυσμένοι' χρήστες θεωρούνται εκείνοι που έχουν ενημερωθεί για κάποια πληροφορία, κάποιο θέμα, το οποίο αποτελεί και την 'ασθένεια'. Αντίστοιχα οι υγιείς χρήστες αποτελούνται από το κομμάτι του πληθυσμού που παραμένει ανενημέρωτο τη στιγμή που μελετάται το φαινόμενο. Για

[1] Εκτενέστερη περιγραφή των όρων 'μολυσμένος' και 'υγιής' γίνεται στο κεφάλαιο 4

την ανάπτυξη του μοντέλου θεωρήθηκε ότι η πληροφορία που διαδίδεται στο δίκτυο αποτελεί ένα ευρύτερο θέμα το οποίο με όρους του δικτύου μπορεί να αντιστοιχηθεί σε ένα γενικότερο hashtag[2]. Στη συγκεκριμένη περίπτωση μελετάται ο ρυθμός που δέχεται ένας χρήστης tweets[3] ανεξαρτήτως περιεχομένου και συσχέτισης μεταξύ τους. Για το παραπάνω μοντέλο γίνεται η προσαρμογή του στα πραγματικά δεδομένα που έχουν συλλεχθεί από το δίκτυο του Twitter και με χρήση των παραμέτρων που προκύπτουν γίνεται η αξιολόγηση των θεωρητικών αποτελεσμάτων, τα οποία τελικά πλησιάζουν αρκετά τα πειραματικά.

Ταυτόχρονα μελετάται ο ρυθμός με τον οποίο οι χρήστες δέχονται καινούρια πληροφορία και ενημερώνονται για τα θέματα που προκύπτουν στο δίκτυο. Συγκεκριμένα, εξετάζοντας τα tweets που δέχονται οι χρήστες από τους χρήστες που ακολουθούν, εξάγονται αποτελέσματα για την κατανομή που ακολουθεί η εισερχόμενη σε αυτούς πληροφορία. Σε αυτή τη διαδικασία η μελέτη έγινε με βάση μεμονωμένα μηνύματα και όχι ομάδες μηνυμάτων με κοινό χαρακτηριστικό στο θέμα τους (hashtag).

1.2 Διάρθρωση

Η παρούσα διπλωματική εργασία ακολουθεί την εξής δομή:

Το Κεφάλαιο 2 αποτελεί μια εισαγωγή στη θεωρία γραφημάτων (Graph Theory), όπου παρατίθενται τα βασικά χαρακτηριστικά αυτών. Γίνεται εισαγωγή στις έννοιες των γράφων και τα στοιχεία που τους απαρτίζουν. Περιγράφεται η διαδικασία με την οποία μελετώνται καθώς και βασικές εφαρμογές, όπως η διάσχιση τους. Γίνεται επίσης αναφορά σε συγκεκριμένο είδος γραφημάτων, τα δένδρα, στη μορφή τους και στους κανόνες που τα διέπουν. Τέλος αναφέρονται οι τρόποι αναπαράστασης των γραφημάτων, δηλαδή οι πίνακες γειτνίασης και οι λίστες γειτνίασης. Συνήθως, στη μελέτη των δικτύων βοηθά η αναπαράσταση τους ως γραφήματα και για αυτό το λόγο η γνώση των μεθόδων για το χειρισμό και τη μελέτη των γράφων είναι απαραίτητη για την επιπλέον μελέτη και κατανόηση των δικτύων.

[2]Μια μέθοδος για να γίνει η συλλογή όλων των μηνυμάτων που είναι σχετικά με ένα συγκεκριμένο θέμα.

[3]Τα μηνύματα με μέγεθος μέχρι 140 χαρακτήρες που δημοσιεύουν οι χρήστες του Twitter.

Στο Κεφάλαιο 3 γίνεται μια εισαγωγή στα Σύνθετα Δίκτυα (Complex Networks). Περιγράφονται εκτενώς ορισμένες από τις κατηγορίες των Σύνθετων Δικτύων, όπως τα Κοινωνικά Δίκτυα και τα Κινητά Κοινωνικά Δίκτυα. Επίσης γίνεται περιγραφή των δικτύων Μικρού-Κόσμου (Small-World Networks), καθώς και των δικτύων Χωρίς-Κλίμακα (Scale-free Networks). Εφόσον η παρούσα εργασία επικεντρώνεται στη μελέτη ενός κοινωνικού δικτύου το οποίο παρουσιάζει τα χαρακτηριστικά και τις συμπεριφορές των σύνθετων δικτύων, είναι απαραίτητη η εισαγωγή σε αυτά τα δίκτυα και η κατανόηση ορισμένων στοιχείων τους.

Στο Κεφάλαιο 4 περιγράφεται η έννοια της διάχυσης της πληροφορίας στα δίκτυα και το επιδημιολογικό μοντέλο για την μελέτη της. Επιπρόσθετα περιγράφονται εκτενώς διάφορα είδη του επιδημιολογικού μοντέλου και παρουσιάζεται το βασικό υπόβαθρο για χρήση στην περιγραφή του προβλήματος που αναλύεται στη διπλωματική.

Ακολουθεί το Κεφάλαιο 5, που γίνεται η παρουσίαση του προβλήματος στο οποίο επικεντρώνεται η παρούσα εργασία. Παρουσιάζεται το μοντέλο που αναπτύχθηκε για τη μελέτη της διάδοσης της πληροφορίας στο δίκτυο του Twitter και η επίλυσή του, ενώ πραγματοποιείται η αξιολόγηση του με βάση τα εμπειρικά δεδομένα που συλλέχθηκαν από το Twitter. Επιπλέον, περιγράφεται ο τρόπος που έγινε η συλλογή των δεδομένων με τη βοήθεια της υπηρεσίας του Twitter. Τέλος, αναλύεται ο ρυθμός με τον οποίο ενημερώνονται μέσω λήψης νέας πληροφορίας οι χρήστες του δικτύου και μελετάται η κατανομή της εισερχόμενης πληροφορίας στους χρήστες.

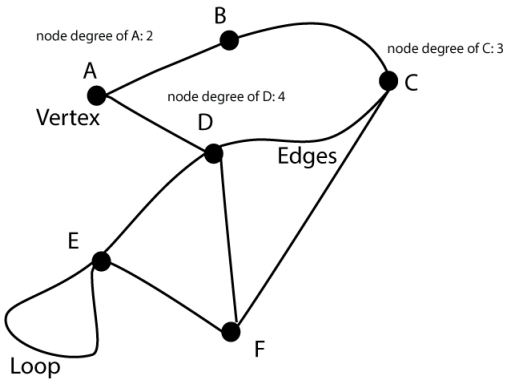
Τέλος, στο Κεφάλαιο 6 γίνεται μια σύνοψη αυτών που παρουσιάστηκαν στα παραπάνω κεφάλαια, των αποτελεσμάτων που προέκυψαν καθώς και μελλοντικές επεκτάσεις της παρούσας εργασίας.

2 Κεφάλαιο
Θεωρία Γραφημάτων

2.1 Γενικά Στοιχεία των Γραφημάτων

Στη μελέτη των δικτύων χρησιμοποιούνται μηχανισμοί και δομές για την αποδοτικότερη περιγραφή και ανάλυση τους. Μία από τις πιο θεμελιώδεις και εκφραστικές δομές είναι το γράφημα (graph). Το γράφημα, $G(V, E)$, είναι ένας τρόπος κωδικοποίησης των σχέσεων ανά ζεύγη μεταξύ των αντικειμένων ενός συνόλου, καθώς αποτελείται από ένα σύνολο V κόμβων (nodes, vertices) και ένα σύνολο E ακμών (edges). Κάθε ακμή 'συνδέει' δύο κόμβους. Έτσι μπορεί μια ακμή $e \in E$ να αναπαρασταθεί ως ένα υποσύνολο του V με δύο στοιχεία, δηλαδή $e = (u, v)$ για κάποια $u, v \in V$, όπου τα u και v ονομάζονται άκρα (ends) της ακμής e.

Οι ακμές σε ένα μη-κατευθυνόμενο γράφο υποδεικνύουν μια συμμετρική σχέση μεταξύ των κόμβων. Για την απεικόνιση μιας ασύμμετρης σχέσης χρησιμοποιείται μια συναφής έννοια, εκείνη του κατευθυνόμενου ή προσανατολισμένου γράφου (directed graph). Το κατευθυνόμενο γράφημα αποτελείται από ένα σύνολο κόμβων V και από ένα σύνολο κατευθυνόμενων ακμών (directed edges) E'. Για κάθε ακμή $e' \in E'$ με άκρα u, v δεν μπορεί να γίνει εναλλαγή των άκρων, καθώς έχουν συγκεκριμένους ρόλους. Το v είναι η ουρά (tail) της ακμής ενώ το u αποτελεί την αρχή (head) της. Πρακτικά κάθε κατευθυνόμενη ακμή αποτελεί ένα διατεταγμένο ζεύγος κόμβων με αποτέλεσμα να μην επιτρέπεται η εναλλαγή τους. Συνηθέστερα για την ακμή e' χρησιμοποιείται η έκφραση *εξέρχεται* από τον κόμβο u και *εισέρχεται* στον κόμβο v. Αντίστοιχα ένα απλό γράφημα με συμμετρικές ακμές θα ονομάζεται *μη κατευθυνόμενο* (undirected graph). Μια ακμή μπορεί να ενώνει τον κόμβο με τον εαυτό του, δηλαδή να σχηματίζει έναν βρόγχο (loop). Ταυτόχρονα ένα ζευγάρι κόμβων μπορούν να ενώνονται με παραπάνω από μια ακμές, όπου τότε ο γράφος καθίσταται κυκλικός. Ένα γράφημα θεωρείται απλό, όταν δεν περιέχει κανένα από τα χαρακτηριστικά που αναφέρθηκαν παραπάνω [24], [18].



Σχήμα 2.1: Βασικό γράφημα με στοιχεία όπως οι κόμβοι, οι ακμές, οι κύκλοι και ο βαθμός κόμβου.

Στη συνέχεια θα μελετηθούν ορισμένα χαρακτηριστικά των γραφημάτων που επιτρέπουν την αποδοτικότερη και ευκολότερη μελέτη τους. Σε ένα μη κατευθυνόμενο δίκτυο ο βαθμός (degree) ενός κόμβου, u, συμβολίζεται με $d(u)$ και αποτυπώνει τον αριθμό των ακμών που προσπίπτουν στον κόμβο u, δηλαδή τις άμεσες συνδέσεις του κόμβου με τους άλλους κόμβους μέσα στο γράφο. Εάν υπάρχει κόμβος του δικτύου που έχει $d(u) = 0$, τότε θεωρείται απομονωμένος, καθώς δεν υπάρχει ακμή που να τον ενώνει με οποιοδήποτε άλλο κόμβο του γραφήματος. Σε ένα κατευθυνόμενο γράφημα ορίζονται δύο είδη βαθμών για τους κόμβους, ο 'μέσα-βαθμός' (in-degree) και ο 'έξω-βαθμός' (out-degree). Ο μέσα-βαθμός δηλώνει τον αριθμό των ακμών που εισέρχονται στον συγκεκριμένο κόμβο. Αντίστοιχα ο έξω-βαθμός αντιπροσωπεύει τον αριθμό των εξερχόμενων ακμών από τον κόμβο.

Τα σταθμισμένα γραφήματα (Weighted graphs) αποτελούν μία περιοχή της θεωρίας των γραφημάτων, η οποία χρησιμοποιείται σε μεγάλο βαθμό για

τη μελέτη των πραγματικών δικτύων. Σε αυτά τα γραφήματα οι ακμές έχουν ένα επιπλέον γνώρισμα, έναν πραγματικό αριθμό που καλείται βάρος (weight). Η έννοια του συντομότερου μονοπατιού μεταξύ κόμβων είναι λίγο διαφορετική στους γράφους με βάρη, καθώς πλέον λαμβάνεται υπόψιν το βάρος των ακμών και όχι το πλήθος που συμμετέχουν σε αυτό. Γενικά στον τομέα των δικτύων η εύρεση του συντομότερου μονοπατιού μεταξύ τυχαίων κόμβων είναι ένα από τα πιο ενδιαφέροντα και κρίσιμα ζητήματα [24], [12], π.χ στα οδικά δίκτυα ή στα ασύρματα δίκτυα και τα δίκτυα τηλεπικοινωνιών. Υπάρχει ένα πλήθος αλγορίθμων για εύρεση συντομότερων μονοπατιών σε ποικίλα γραφήματα π.χ Dijkstra, BFS, Bellman Ford, [24].

Για γραφήματα με βάρη στις ακμές μελετάται επίσης η δύναμη (strength) κάθε κόμβου, μέγεθος που δίνει μία πλουσιότερη περιγραφή του γραφήματος. Για ένα κόμβο s_i το μέτρο της δύναμης του δίνεται :

$$s_i = \sum\nolimits_{j \in V(i)} w_{ij} \tag{1}$$

όπου w_{ij} είναι όλα τα βάρη των ακμών που προσπίπτουν στον κόμβο s_i. Αντιστοίχως για κατευθυνόμενα γραφήματα με βάρη στις ακμές υπάρχουν δύο μεγέθη δύναμης που χαρακτηρίζουν τους κόμβους. Η 'μέσα-δύναμη' (in-strength) και η 'έξω-δύναμη' (out-strength) δίνονται :

$$s_i^{in} = \sum\nolimits_{j \in V(i)} w_{ij} \tag{2}$$

$$s_i^{out} = \sum\nolimits_{j \in V(i)} w_{ji} \tag{3}$$

Ως τάξη ενός γραφήματος θεωρείται ο αριθμός των κόμβων, $|V|$, από τους οποίους αποτελείται. Το μέγεθος του γραφήματος είναι ίσο με τον αριθμό των ακμών, $|E|$. Βασιζόμενοι στο βαθμό των κόμβων υπάρχει περαιτέρω διαχωρισμός των γραφημάτων. Ως πλήρης γράφος θεωρείται ο απλός, μη κατευθυνόμενος γράφος που περιέχει όλες τις πιθανές ακμές του δικτύου, δηλαδή αν n ο αριθμός των κόμβων τότε ο αριθμός των ακμών είναι $\frac{n(n-1)}{2}$. Κ-κανονικό (K-regular) ονομάζεται το απλό γράφημα, όπου όλοι οι κόμβοι έχουν βαθμό ίσο με Κ. Ένα κατευθυνόμενο γράφημα θεωρείται ισορροπημένο (balanced) εάν ισχύει $d_{in}(v) = d_{out}(v)$ για όλους τους κόμβους.

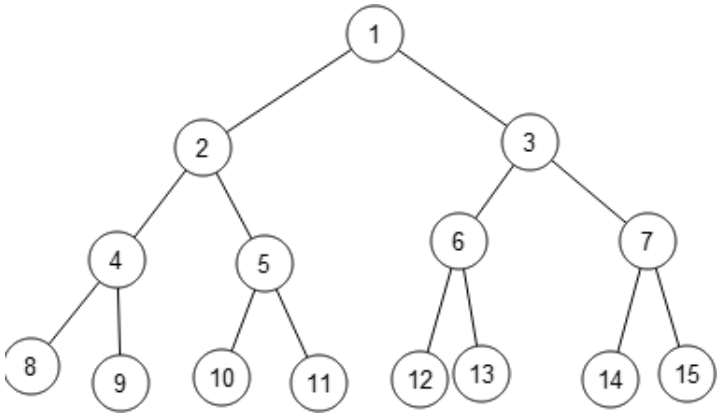
Μια από τις θεμελιώδεις εφαρμογές σε ένα γράφημα είναι η διάσχιση του μέσω μίας ακολουθίας διακριτών κόμβων που συνδέονται με ακμές. Για την παραπάνω λειτουργία απαραίτητος είναι ο ορισμός της έννοιας της διαδρομής ή μονοπατιού (path). Σε ένα μη κατευθυνόμενο γράφημα $G(V, E)$ η διαδρομή P ορίζεται από την ακολουθία κόμβων $v_1, v_2, ..., v_k$ και κάθε ζευγάρι διαδοχικών κόμβων που εμφανίζεται, v_i, v_{i+1}, θα πρέπει να συνδέεται με μια ακμή στο G. Η ακολουθία αυτή ονομάζεται διαδρομή από το v_i στο v_k. Ένα μη κατευθυνόμενο γράφημα ονομάζεται συνεκτικό, αν για κάθε ζευγάρι κόμβων u, v υπάρχει διαδρομή μεταξύ τους. Για τους κατευθυνόμενους γράφους ο ορισμός της συνεκτικότητας είναι λίγο πιο δύσκολος, καθώς πρέπει να ληφθεί υπόψιν και η κατεύθυνση των ακμών. Ισχυρά συνεκτικό ονομάζεται το γράφημα για το οποίο υπάρχει διαδρομή από τον κόμβο u στον κόμβο v και το αντίστροφο. Ορισμένες φορές είναι επιθυμητή και η γνώση της πιο σύντομης διαδρομής μεταξύ δύο τυχαίων κόμβων. Ορίζεται η απόσταση (distance) μεταξύ δύο κόμβων u, v ως ο ελάχιστος αριθμός ακμών σε όλες τις διαδρομές μεταξύ u, v.

2.2 Δένδρα

Στην κατηγορία των γραφημάτων ανήκουν τα δένδρα. Πρόκειται για γράφους οι οποίοι δεν περιέχουν κύκλους. Κάθε δένδρο αποτελείται από έναν αρχικό κόμβο, τη ρίζα (root), τους ενδιάμεσους κόμβους και τους κόμβους φύλλα. Τα φύλλα έχουν βαθμό κόμβου 1, διότι αποτελούν το τελευταίο επίπεδο του δένδρου, ενώ οι ενδιάμεσοι κόμβοι έχουν μεγαλύτερο βαθμό εξαρτώμενο από τα γνωρίσματα του κάθε δένδρου. Στις δομές των δένδρων δύο κόμβοι ενώνονται μέσω ενός μοναδικού μονοπατιού, καθώς υπάρχει έλλειψη κύκλων.

2.3 Αναπαραστάσεις Γραφημάτων

Υπάρχουν δύο καθιερωμένοι τρόποι αναπαράστασης ενός γραφήματος $G = (V, E)$. Ο πρώτος γίνεται υπό τη μορφή ενός συνόλου από καταλόγους γειτνίασης και ο δεύτερος γίνεται υπό τη μορφή ενός πίνακα γειτνίασης. Αμφότερες οι αναπαραστάσεις μπορούν να εφαρμοστούν τόσο σε κατευθυνόμενα όσο και σε μη κατευθυνόμενα γραφήματα. Συνήθως προτιμάται η αναπαράσταση



Σχήμα 2.2: Αναπαράσταση γράφου με μορφή δένδρου

μέσω καταλόγων γειτνίασης, διότι προσφέρεται για πυκνή αναπαράσταση αραιών γραφημάτων, στα οποία το $|E|$ είναι πολύ μικρότερο του $|V|^2$. Όταν το γράφημα είναι πυκνό -δηλαδή, όταν το $|E|$ είναι συγκρίσιμο με το $|V|^2$ - τότε πιθανόν η αναπαράσταση μέσω πίνακα γειτνίασης να είναι προτιμότερη.

Η αναπαράσταση μέσω καταλόγων γειτνίασης ενός γραφήματος $G = (V, E)$ συνίσταται σε μια συστοιχία, Adj, από $|V|$ καταλόγους, έναν για κάθε κόμβο του V. Για κάθε $u \in V$, ο κατάλογος γειτνίασης Adj[u] περιέχει όλους τους κόμβους v για τους οποίους υπάρχει η ακμή $(u, v) \in E$. Δηλαδή ο Adj[u] αποτελείται από όλους τους κόμβους που γειτνιάζουν με τον u στο G. Η διάταξη των κόμβων σε κάθε κατάλογο γειτνίασης κατά κανόνα είναι αυθαίρετη. Ένα πιθανό μειονέκτημα της αναπαράστασης των καταλόγων γειτνίασης είναι ότι ο ταχύτερος δυνατός τρόπος να προσδιοριστεί εάν υπάρχει στο γράφημα μια δεδομένη ακμή (u, v) είναι μέσω της αναζήτησης του v στον κατάλογο γειτνίασης του Adj[u]. Το μειονέκτημα αυτό μπορεί να αρθεί με την αναπαράσταση του γραφήματος σε πίνακα γειτνίασης αλλά με τίμημα τη δέσμευση περισσότερης μνήμης.

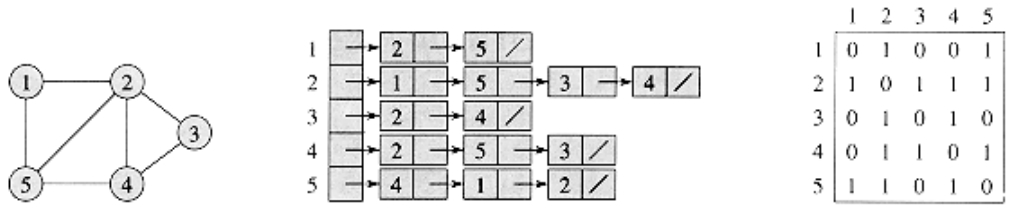
Για την αναπαράσταση μέσω πίνακα γειτνίασης ενός γραφήματος $G = (V, E)$ θεωρείται ότι οι κόμβοι είναι αριθμημένοι από το 1 έως το $|V|$ κατά

αυθαίρετο τρόπο. Στην περίπτωση αυτή, η αναπαράσταση ενός γραφήματος G με χρήση του πίνακα γειτνίασης συνίσταται από έναν $|V|\times|V|$ πίνακα $A = (a_{ij})$ με στοιχεία

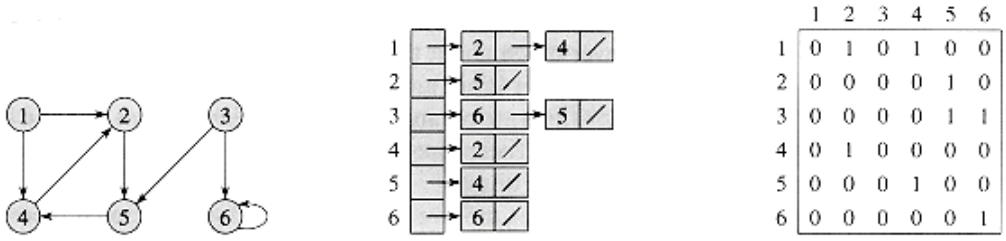
$$a_{ij} = \begin{cases} 1 & (i,j) \in E \\ 0 & (i,j) \notin E. \end{cases}$$

Για ένα μη κατευθυνόμενο γράφημα, όπου οι (u,v) και (v,u) αναπαριστούν την ίδια ακμή, ο πίνακας γειτνίασης του γραφήματος θα είναι συμμετρικός ως προς την κύρια διαγώνιο. [24]

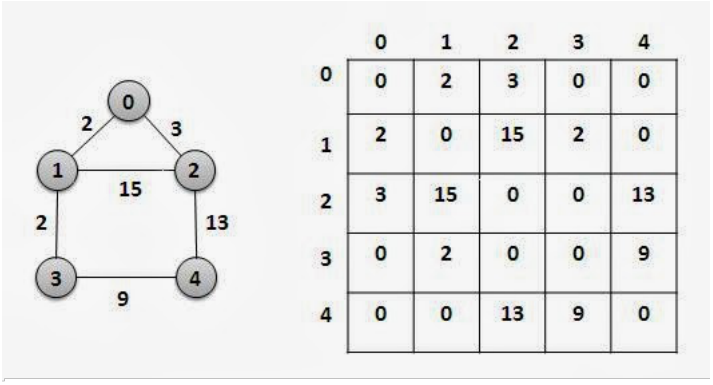
Η αναπαράσταση μέσω πίνακα γειτνίασης μπορεί να χρησιμοποιηθεί και για γραφήματα με βάρη στις ακμές. Π.χ., το $G = (V,E)$ είναι γράφημα με βάρη, με συνάρτηση βάρους ακμών w, το βάρος $w(u,v)$ της ακμής $(u,v) \in E$ αποθηκεύεται απλώς ως το στοιχείο της γραμμής u και της στήλης v του πίνακα γειτνίασης. Στη συνέχεια ακολουθούν παραδείγματα αναπαράστασης γραφήματος για κατευθυνόμενους, μη κατευθυνόμενους και γράφους με βάρη στις ακμές τους.



Σχήμα 2.3: Μη κατευθυνόμενο γράφημα, ο αντίστοιχος κατάλογος γειτνίασης και ο αντίστοιχος πίνακας γειτνίασης, [24].



Σχήμα 2.4: Κατευθυνόμενο γράφημα, ο αντίστοιχος κατάλογος γειτνίασης και ο αντίστοιχος πίνακας γειτνίασης, [24].



Σχήμα 2.5: Γράφημα με βάρη στις ακμές και ο αντίστοιχος σταθμισμένος πίνακας γειτνίασης, [24].

3 Κεφάλαιο
Σύνθετα Δίκτυα - Complex Networks

3.1 Γενικά στοιχεία των Σύνθετων Δικτύων

Ένα δίκτυο είναι ένα σύνολο σχέσεων. Ειδικότερα, ένα δίκτυο αποτελείται από αντικείμενα (κόμβοι) και μια περιγραφή των σχέσεων μεταξύ των κόμβων. Το απλούστερο δίκτυο αποτελείται από δύο αντικείμενα, 1 και 2, και μία σχέση που τα ενώνει. Οι κόμβοι 1 και 2, για παράδειγμα, μπορεί να αποτελούν ανθρώπους και η σχέση που τους ενώνει να είναι 'βρίσκονται στο ίδιο δωμάτιο' .Υπάρχουν και κατευθυνόμενα δίκτυα στα οποία οι συνδέσεις έχουν κατεύθυνση μεταξύ των κόμβων. Οι συνδέσεις σε ένα δίκτυο δεν αντιπροσωπεύουν μόνο κοινά γνωρίσματα, αλλά μπορούν να αποτυπώνουν μια ροή μεταξύ των κόμβων.

Η μελέτη των σύνθετων δικτύων (complex networks) [23] γνωρίζει ταχεία ανάπτυξη καθώς αυτά συναντώνται συχνά και στην καθημερινή ζωή. Τα σύνθετα δίκτυα έχουν μη τετριμμένα τοπολογικά χαρακτηριστικά τα οποία δεν εμφανίζονται σε άλλα δίκτυα όπως π.χ. δίκτυα πλέγματα και τυχαία δίκτυα. Χαρακτηριστικά παραδείγματα των σύνθετων δικτύων είναι το Internet, τα νευρωνικά συστήματα καθώς και τα κοινωνικά δίκτυα.

Μια βασική ιδιότητα των complex networks είναι ότι εκτίθενται σε συμπεριφορές που δεν μπορούν να προβλεφθούν a priori. Για τη μοντελοποίηση των σύνθετων δικτύων γίνονται ορισμένες παραδοχές. Για την απεικόνιση των αλληλεπιδράσεων και των σχέσεων μεταξύ των αντικειμένων του δικτύου χρησιμοποιούνται οι ακμές (σύνδεσμοι). Οι κόμβοι μπορούν να ανταλλάσσουν δεδομένα διαφόρων τύπων μέσω των παραπάνω συνδέσμων. Τέλος, όσον αφορά την επικοινωνία μεταξύ των κόμβων θεωρείται ότι δύο κόμβοι θα πρέπει να συνδέονται άμεσα, ώστε να μπορούν να ανταλλάξουν δεδομένα. Οι άμεσες συνδέσεις μπορούν να ανήκουν στο φυσικό στρώμα π.χ. ασύρματα δίκτυα, αλλά μπορούν επίσης να είναι νοητές, όπως είναι μια αλληλεξάρτηση εντός

ενός κοινωνικού δικτύου.

Ένα από τα τυπικότερα χαρακτηριστικά των συγκεκριμένων δικτύων είναι το μέγεθός τους. Συνηθέστερα αποτελούνται από ένα μεγάλο αριθμό κόμβων πράγμα που καθιστά δύσκολη τη συνολική περιγραφή και μελέτη τους. Λόγω αυτού, πολλοί ερευνητές επικεντρώνονται στη μελέτη υποσυνόλων τους ή υποσυνόλων των χαρακτηριστικών τους. Πιο συγκεκριμένα, οι ερευνητές ενδιαφέρονται για τον αριθμό των κόμβων και τους κανόνες, βάση των οποίων γίνονται οι τοπικές συνδέσεις, οι οποίοι είναι πιθανοτικοί. Γενικότερα, ορισμένες χαρακτηριστικές συμπεριφορές που παρατηρούνται στα σύνθετα δίκτυα έχουν ως εξής:
Οι τυπικές αποστάσεις μεταξύ των κόμβων ενός δικτύου είναι μικρές. Ταυτόχρονα ο βαθμός των κόμβων του δικτύου ακολουθεί power-law κατανομή με λίγους κόμβους να έχουν μεγάλο βαθμό κόμβου (degree) και η πλειοψηφία να έχει ένα μέσο και σχετικά μικρό βαθμό κόμβου.

3.2 Σύνθετα Δίκτυα και Θεωρία Δικτύων

Όπως αναφέρθηκε και σε παραπάνω κεφάλαιο ένας δημοφιλής τρόπος για τη μελέτη των δικτύων και ειδικότερα των Σύνθετων Δικτύων, είναι μέσω της θεωρίας Γραφημάτων. Τα Σύνθετα Δίκτυα απαρτίζονται από κόμβους (nodes) και από σχέσεις που ενώνουν αυτούς, τις ακμές (edges), που τους επιτρέπουν να αλληλεπιδρούν μεταξύ τους. Για τους κόμβους, σε αντιστοιχία με τα γραφήματα χαρακτηριστικά, όπως ο βαθμός κόμβου (node degree), μετρική με μεγάλη σημασία για την περιγραφή των Σύνθετων Δικτύων. Ταυτόχρονα, συναντώνται ομοιότητες με τα γραφήματα με βάρη, καθώς οι αλληλεπιδράσεις μεταξύ κόμβων διαφέρουν ανάλογα με τη φύση της σχέσης και τις συμπεριφορές των χρηστών. Τέλος στα συγκεκριμένα δίκτυα υπάρχει διαχωρισμός σε κατευθυνόμενα και μη δίκτυα, όπως ακριβώς στα γραφήματα.

3.3 Κοινωνικά Δίκτυα - Social Networks

Τα κοινωνικά δίκτυα αποτελούν μια δομή από οντότητες (άτομα, οργανώσεις ή συστήματα) που συνδέονται μεταξύ τους μέσω μίας ή πολλών αλληλεξαρ-

τήσεων. Αυτές οι αλληλεξαρτήσεις αντιπροσωπεύουν αξίες, φυσικές επαφές, οικονομικές ανταλλαγές και ομάδες συμμετοχής. Επομένως, μέσω της κοινωνικής δικτύωσης ορίζονται οι συμπεριφορές των οντοτήτων και κατανοούνται και οι σχέσεις που επικρατούν μεταξύ τους. Αυτό επιτυγχάνεται μέσω ενός αναπτυγμένου συνόλου μεθόδων ικανών για την ανάλυση της δομής του δικτύου, καθώς και την κατανόηση των μοτίβων που επικρατούν.

Ο λόγος που τα κοινωνικά δίκτυα ξεχωρίζουν σε σχέση με τα υπόλοιπα είναι οι συνθήκες κάτω από τις οποίες σχηματίζονται ή όχι συνδέσεις μεταξύ των χρηστών τους και το γεγονός ότι οι κόμβοι μπορούν να επιδεικνύουν παρόμοια χαρακτηριστικά. Υπάρχουν ορισμένες υποθέσεις που ισχύουν για τα κοινωνικά δίκτυα, όσον αφορά τον τρόπο με τον οποίο σχηματίζουν τις συνδέσεις τους. Η εγγύτητα είναι ένας παράγοντας που καθορίζει τη δημιουργία σχέσεων μεταξύ των χρηστών. Θεωρείται πιθανότερο να ενωθούν δύο κόμβοι, όταν οι υπόλοιπες συνθήκες είναι όμοιες, εάν βρίσκονται τοπολογικά (γεωγραφικά) κοντά ο ένας στον άλλον.

Ένας δεύτερος παράγοντας διαμόρφωσης σχέσεων μεταξύ οντοτήτων σε κοινωνικά δίκτυα είναι να υπάρχουν ένα ή περισσότερα κοινά κοινωνικά χαρακτηριστικά μεταξύ των χρηστών, όπως για παράδειγμα η κοινωνική τάξη. Αυτός ο παράγοντας είναι γνωστός ως Homophily [16]. Επιπλέον η απόσταση μεταξύ δύο κόμβων μπορεί να παίξει καθοριστικό ρόλο στη δημιουργία συνδέσεων. Η απόσταση καθορίζεται από δύο παραμέτρους: (1) ο αριθμός των άμεσων 'φίλων' των κόμβων του δικτύου, (2) την έκταση στην οποία υπάρχει επικάλυψη μεταξύ των κόμβων 'φίλων' των δραστών. Γενικότερα στα δίκτυα αυτά δίνεται μεγάλη έμφαση στη δημιουργία και στη σημασία των ακμών μεταξύ των δραστών για αυτό άλλωστε και η ανάλυση των κοινωνικών δικτύων βασίζεται σε μοντέλα για τη μελέτη της διαμόρφωσης των σχέσεων και συνδέσεων μεταξύ των κόμβων.

Πλέον η έννοια του κοινωνικού δικτύου έχει εισχωρήσει και στην περιοχή των τηλεπικοινωνιών και της πληροφορίας, αποτελώντας μέσο για την αποτελεσματική ανταλλαγή δεδομένων καθώς και την παροχή υπηρεσιών. Η

θεώρηση των κοινωνικών σχέσεων σε ένα δίκτυο τηλεπικοινωνιών και πληροφορικής έχει τη δυνατότητα να βελτιστοποιεί την ακρίβεια και την αποτελεσματικότητα των υπηρεσιών, που παρέχει στους χρήστες του μέσω της μελέτης των δομών και των δεσμών μεταξύ του πληθυσμού του. Τυπικό παράδειγμα κοινωνικών δικτύων είναι τα κινητά κοινωνικά δίκτυα (mobile social networks), τα οποία αποτελούν ήδη μια ανεπτυγμένη περιοχή των δικτύων, στην οποία έχουν συνδυαστεί τα χαρακτηριστικά των κοινωνικών δικτύων με αυτά των ενσύρματων αλλά και ασύρματων δικτύων, π.χ. Internet.

Ένα βασικό χαρακτηριστικό των κοινωνικών δικτύων και εν γένει των σύνθετων δικτύων είναι η υψηλή ομαδοποίηση των κόμβων, γνωστό ως clustering [25]. Το μέγεθος αυτό χρησιμοποιείται ώστε να περιγράψει τη δομή ενός κοινωνικού δικτύου τόσο τοπικά, π.χ. σε επίπεδο κόμβων, όσο και γενικά, π.χ. σε επίπεδο δικτύου. Η ιδιότητα αυτή υποδηλώνει την ικανότητα των παραπάνω δικτύων να σχηματίζουν πολλούς δεσμούς μεταξύ γειτονικών κόμβων. Η ομαδοποίηση C_i ενός κόμβου i δηλώνει τον αριθμό των άμεσων συνδέσεων μεταξύ των γειτονικών κόμβων του κόμβου.

$$C_i = \frac{\text{αριθμός τριγώνων συνδεδεμένων στον κόμβο } i}{\text{αριθμός τριπλέτων με κέντρο τον κόμβο } i} \tag{4}$$

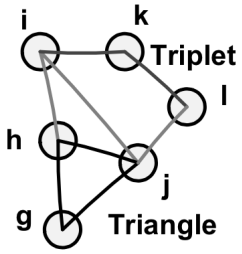
$$= \frac{\text{αριθμός ακμών μεταξύ των γειτονικών κόμβων του κόμβου } i}{\text{αριθμός όλων των πιθανών ακμών μεταξύ των γειτονικών κόμβων του κόμβου } i} \tag{5}$$

Στη θεωρία γραφημάτων, ένα τρίγωνο είναι ένας κατευθυνόμενος πλήρης 3-κόμβων υπογράφος, ενώ η τριπλέτα είναι ένας απλός συνδεδεμένος 3-κόμβων υπογράφος [25]

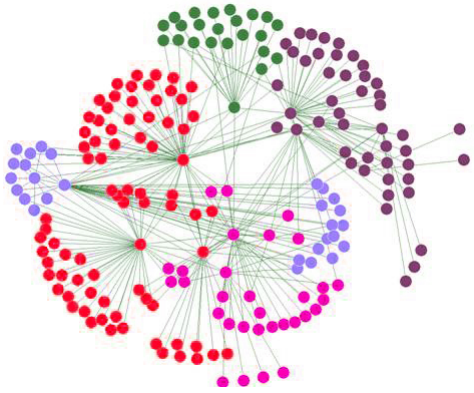
Η ομαδοποίηση ολόκληρου του γραφήματος δίνεται από :

$$C_G = \frac{\text{αριθμός τριγώνων στο δίκτυο}}{\text{αριθμός τριπλέτων στο δίκτυο}} \tag{6}$$

Στη συνέχεια υπάρχει εκτενέστερη αναφορά στα κινητά κοινωνικά δίκτυα με περισσότερες πληροφορίες σχετικές με τη δομή τους και τα χαρακτηριστικά τους.



Σχήμα 3.1: Παράδειγμα υπολογισμού της ομαδοποίησης του κόμβου [25].

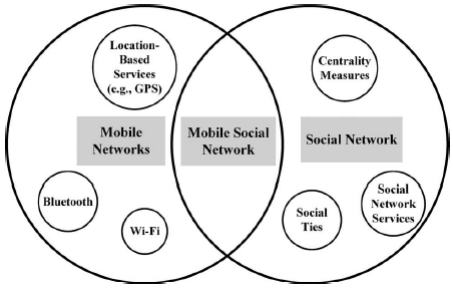


Σχήμα 3.2: Παράδειγμα Κοινωνικού Δικτύου

3.4 Κινητά Κοινωνικά Δίκτυα - Mobile Social Networks

Τα Mobile Social Networks (MSN) [13] προκύπτουν ως η τομή δύο διαφορετικών τύπων δικτύων, των κινητών τηλεπικοινωνιακών και των κοινωνικών δικτύων. Είναι μια καινούρια περιοχή στην οποία μελετάται η δημοσίευση μηνυμάτων, η ανταλλαγή δεδομένων και η παροχή υπηρεσιών, δηλαδή η κοινωνική δικτύωση μέσω των κινητών τηλεπικοινωνιακών συστημάτων. Ειδικότερα, στα συστήματα αυτά οι χρήστες μεταφέρουν κινητές συσκευές και επικοινωνούν δημοσιεύοντας πληροφορίες μεταξύ τους μέσω των ηλεκτρονικών κοινωνικών δικτύων ή ανταλλάσσοντας μηνύματα μέσω ασύρματης δικτύωσης. Αποτέλεσμα αυτών των δικτύων είναι η αλληλεπίδραση και αλληλεξάρτηση των χρηστών τόσο σε επίπεδο χωρικό όσο και κοινωνικό. Η χρήση αλγορίθμων

και μετρικών της ανάλυσης των κοινωνικών δικτύων (social network analysis) μπορεί να βοηθήσει σημαντικά τη λειτουργία των κινητών κοινωνικών δικτύων μέσα από ένα αποτελεσματικό σχεδιασμό πρωτοκόλλων. Στο Σχήμα 3.3 γίνεται φανερό ότι τα κινητά κοινωνικά δίκτυα είναι ο συνδυασμός δύο δικτύων, των κοινωνικών και των κινητών.



Σχήμα 3.3: Mobile Social Network ως ένωση των κινητών και των κοινωνικών δικτύων [13].

Τα MSN αναπτύσσονται τόσο σε κεντροποιημένα όσο και κατανεμημένα κινητά δίκτυα, χάρη την αλληλεξάρτηση μεταξύ των κινητών συσκευών. Ταυτόχρονα εκμεταλλευόμενα τη γνώση που εξάγεται από την κοινωνική δικτύωση μπορούν να επιτύχουν βελτιωμένη ποιότητα υπηρεσιών (QoS). Βασικές διαφορές των κινητών κοινωνικών με των απλών κοινωνικών δικτύων είναι ότι η κινητικότητα των χρηστών μπορεί να χρησιμοποιηθεί ως επιπλέον πληροφορία για την ανάλυση των κοινωνικών σχέσεων μεταξύ των χρηστών. Επιπλέον ο ρόλος των χρηστών στα MSN δεν περιορίζεται μόνο στο να μεταδίδουν δεδομένα αλλά και στο να προσφέρουν ανάδραση σχετικά με την ποιότητα των υπηρεσιών στοχεύοντας στη βελτιστοποίηση τους. [13]

Τα MSNs μπορούν να διακριθούν σε δύο τύπους, τα κεντροποιημένα (Web-based MSNs) και τα αποκεντρωμένα (decentralized MSNs). Το πρώτο είδος χρησιμοποιεί τις υπηρεσίες των κοινωνικών δικτύων για την απόκτηση πληροφοριών μέσω των κινητών τερματικών. Η ύπαρξη πλήθους εφαρμογών

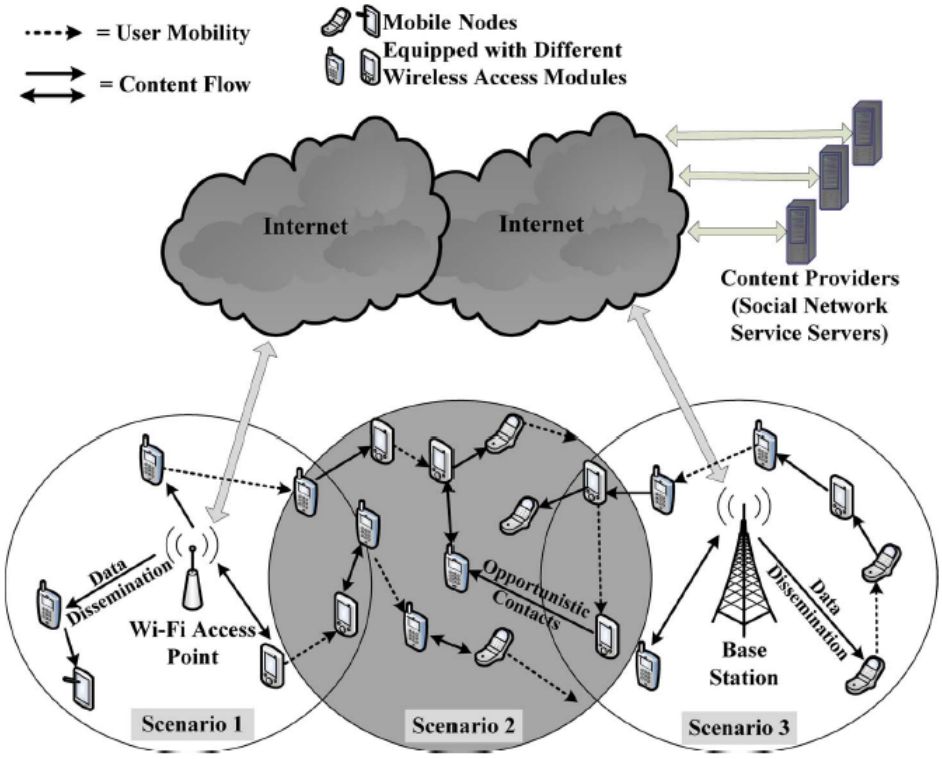
για κεντροποιημένα MSNs δίκτυα οφείλεται στην απλότητα των πρωτοκόλλων που χρησιμοποιούν.

Το δεύτερο είδος του παραπάνω δικτύου είναι τα αποκεντρωμένα MSN, όπου οι χρήστες μεταδίδουν τα δεδομένα μεταξύ των κόμβων του δικτύου δίχως τη μεσολάβηση ενός κεντρικού εξυπηρετητή. Οι χρήστες ανταλλάζουν δεδομένα κοινού ενδιαφέροντος όταν δημιουργούν συνδέσεις μεταξύ τους χρησιμοποιώντας τεχνολογίες, όπως Wi-Fi, Bluetooth. Αποτέλεσμα αυτής της αλληλεπίδρασης είναι ο σχηματισμός ενός κοινωνικού δικτύου.

Τα MSNs συντίθενται από τρείς κύριες συνιστώσες. Το πρώτο συστατικό είναι οι παροχείς περιεχομένου (content providers), όπως για παράδειγμα κάποιος εξυπηρετητής συνδεδεμένος στο διαδίκτυο. Η δεύτερη κύρια συνιστώσα είναι οι κινητοί χρήστες (mobile users/devices), οι οποίοι μπορούν να δέχονται πληροφορίες από τους εξυπηρετητές μέσω κατάλληλων συσκευών. Τέλος, για τη μετάδοση των δεδομένων είναι απαραίτητο το τρίτο μέρος των συγκεκριμένων δικτύων, οι δικτυακές υποδομές (network infrastructures). Στο Σχήμα 3.4 φαίνονται οι τρεις βασικές κατηγορίες των MSNs καθώς και η συνεργασία μεταξύ τους για να τα διαμορφώσουν.

3.5 Δίκτυα Μικρού-Κόσμου (Small-world)

Ως Small-world χαρακτηρίζονται τα δίκτυα στα οποία οι περισσότεροι κόμβοι δεν είναι συνδεδεμένοι ανά ζεύγη μεταξύ τους, αλλά μπορούν να επικοινωνήσουν μέσω ενός μικρού σε αριθμό βημάτων μονοπατιού. Τα δύο βασικά χαρακτηριστικά των συγκεκριμένων δικτύων είναι ότι περιέχει ¨κλίκες¨ (clustering), δηλαδή υποδίκτυα που κάθε κόμβος είναι συνδεδεμένος σχεδόν με κάθε άλλον γεγονός που μετράται από τη Σχέση (6). Η ιδιότητα αυτή είναι γενικότερα ένα από τα χαρακτηριστικά γνώρισματα των complex networks. Δεύτερο χαρακτηριστικό είναι το μικρό μέσο μήκος μονοπατιού. Ειδικότερα για ένα μονοπάτι με μήκος L σε δίκτυο με αριθμό κόμβων N ισχύει ότι $L \propto \log N$. Αποτέλεσμα του παραπάνω χαρακτηριστικού είναι ότι δύο τυχαίοι κόμβοι θα ενώνονται μέσω ενός μικρού μήκους μονοπατιού και επομένως δύο οποιοιδήποτε κόμβοι μπορούν εντέλει να επικοινωνήσουν γρήγορα μέσα σε



Σχήμα 3.4: Τα συστατικά στοιχεία των MSN [13].

ένα small-world δίκτυο. Τα μέλη του δικτύου είναι ικανά να εντοπίσουν τα υπάρχοντα συντομότερα μονοπάτια μεταξύ των κόμβων χωρίς να έχουν πλήρη γνώση του δικτύου, των συνδέσεων και των μοτίβων που επικρατούν χρησιμοποιώντας μόνο τοπική πληροφορία. Τα περισσότερα κοινωνικά δίκτυα, καθώς και ορισμένα τηλεπικοινωνιακά, όπως το Internet μπορούν να χαρακτηρισθούν ως small-world δίκτυα [28], [22].

Λόγω της ευρείας χρήσης του small-world μοντέλου για την περιγραφή πολλών κοινωνικών δικτύων και την ανάγκη για ένα σαφέστερο προσδιορισμό των πολυβηματικών συνδέσεων του δικτύου και των ιδιοτήτων διάδοσης σε αυτό, πραγματοποιήθηκε το πείραμα των Travers και Milgram το 1969. Το γνωστό πείραμα '6 βαθμοί διαχωρισμού (Six Degrees of Separation)' είχε

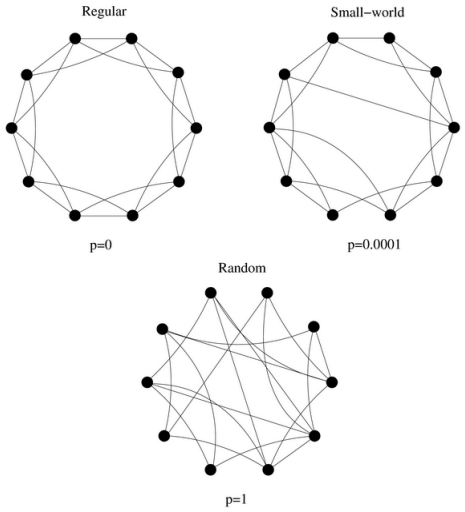
σκοπό να ελέγξει την πιθανότητα δύο άγνωστοι στις ΗΠΑ να συνδέονται μεταξύ τους με ένα μικρό σε μήκος μονοπάτι και ταυτόχρονα να μελετήσει τον τρόπο με τον οποίο τα διάφορα χαρακτηριστικά, όπως η γεωγραφική τοποθεσία και η απασχόληση των δύο συμμετεχόντων, επηρεάζουν τη σύνθεση των ατόμων που συμμετέχουν σε αυτό το μονοπάτι. Επέλεξαν τρεις αρχικές ομάδες ανθρώπων και έναν τελικό στόχο στον οποίο έπρεπε να φτάσει το γράμμα. Σε κάθε αρχικό αποστολέα δινόταν ένα γράμμα με σκοπό να το προωθήσει προς τον τελικό παραλήπτη. Η διάδοση του γράμματος είχε μια συγκεκριμένη μορφή. Αν ο αποστολέας και ο τελικός χρήστης γνωρίζονται μεταξύ τους τότε ο πρώτος δίνει το γράμμα απευθείας στο δεύτερο. Διαφορετικά ο αποστολέας προωθεί το γράμμα σε κάποιον γνωστό του ικανοποιώντας δύο περιορισμούς. Θα πρέπει να γνωρίζει τον ενδιάμεσο παραλήπτη και να συνδέονται άμεσα. Ταυτόχρονα θα πρέπει η επιλογή του παραλήπτη να γίνεται με κριτήριο την πιθανότητα να γνωρίζει τον τελικό παραλήπτη ή την ικανότητα του να προωθήσει το γράμμα προς την κατεύθυνση του τελικού παραλήπτη. Για όλα τα γράμματα που παρέλαβε ο τελικός χρήστης, οι ερευνητές μέτρησαν τον αριθμό των προωθήσεων που δέχτηκε καθένα, καθώς και τα χαρακτηριστικά των ενδιάμεσων χρηστών ανακατασκευάζοντας το ανθρώπινο μονοπάτι που είχε σχηματισθεί. Από τα 296 γράμματα που στάλθηκαν τα 64 έφτασαν στον προορισμό τους. Το μέσο μονοπάτι είχε μήκος περίπου 6 ατόμων το οποίο μπορεί να θεωρηθεί μικρό σε σχέση με τον πληθυσμό των ΗΠΑ. Οι ερευνητές συμπέραναν ότι οι άνθρωποι στις ΗΠΑ χωρίζονται από μικρά μονοπάτια που παρεμβάλλονται πέντε ενδιάμεσοι χρήστες (απόσταση έξι βημάτων). Η απόρριψη των μηνυμάτων που δεν φτάνουν στον παραλήπτη είναι ‘τυχαία’ και τα αποτελέσματα του πειράματος υποδεικνύουν το πόσο τα κοινωνικά και γεωγραφικά χαρακτηριστικά μπορούν να επηρεάσουν το μήκος του μονοπατιού. Επίσης παρατηρήθηκε ότι από τους χρήστες που συμμετείχαν στα έγκυρα μονοπάτια ένα μεγάλο ποσοστό ταυτίζονταν. Δικαιολογείται με αυτόν τον τρόπο η ύπαρξη λίγων κόμβων με μεγάλο βαθμό (node-hubs), οι οποίοι συμμετέχουν σε πολλά μονοπάτια και μεγάλο αριθμό κόμβων με μικρό βαθμό. Επιβεβαιώνεται τελικά η κατανομή με μακρυά (βαριά) ουρά (heavy-tailed) κατανομή των σύνθετων δικτύων και κατ’ επέκταση των κοινωνικών [1].

Χαρακτηριστικά κατασκευαστικά μοντέλα των small-world δικτύων είναι το μοντέλο των Watts and Strogatz (WS Model) και το μοντέλο του Kleinberg's. Το WS Model περιγράφει μια διαδικασία κατασκευής small-world δικτύων ξεκινώντας από ένα δίκτυο πλέγματος κανονικής δομής. Συγκεκριμένα ξεκινώντας από μια δομή πλέγματος προσθέτοντας τυχαίες ακμές (shortcuts) και ενώνοντας απομακρυσμένους κόμβους επιτυγχάνεται μεγάλο ποσοστό ομαδοποίησης μεταξύ των κόμβων (clustering) και μείωση του μέσου μήκους μονοπατιού, δηλαδή τα κύρια χαρακτηριστικά των small-world δικτύων. Με p συμβολίζεται η πιθανότητα αλλαγής του ενός άκρου μίας ακμής που είναι η βασική παράμετρος του μοντέλου WS. Για $p = 0$ το γράφημα παραμένει στην μορφή κανονικού πλέγματος, ενώ για $p = 1$ μετατρέπεται σε τυχαίος γράφος. Για ενδιάμεσες τιμές της πιθανότητας p σχηματίζεται το small-world δίκτυο.

Το μοντέλο Kleinberg από την άλλη αποτελείται από δύο μέρη, το κατασκευαστικό που είναι όμοιο με το WS Model και το αλγοριθμικό. Η κύρια διαφορά μεταξύ των δύο μοντέλων είναι ότι το δεύτερο έχει ως αφετηρία ένα δισδιάστατο κανονικό πλέγμα, το οποίο ενισχύεται με πρόσθετες ακμές που διασυνδέουν απομακρυσμένους κόμβους. Κάθε κόμβος είναι συνδεδεμένος με όλους τους γειτονικούς του σε απόσταση p, εξασφαλίζοντας την ομαδοποίηση του αρχικού γραφήματος. Η πιθανότητα προσθήκης μιας ακμής εξαρτάται από την παράμετρο r, η οποία καθορίζει τι απόσταση μπορεί να διανυθεί μέσω μιας συντόμευσης (stortcut) στο δίκτυο και δίνεται από τη σχέση $d(i,j)^{-r}$, όπου $d(i,j)$ η απόσταση των κόμβων (i,j). Το μοντέλο του Kleinberg θεωρείται ότι εξαρτάται από την παράμετρο r και είναι εφικτό να παραχθούν οικογένειες small-world γράφων για κάθε διαφορετική τιμή του r, αλλά μόνο για $r = 2$ οι κόμβοι μπορούν να εντοπίσουν τα συντομότερα μονοπάτια με χρήση τοπικής μόνο πληροφορίας για το δίκτυο [25].

3.6 Δίκτυα Χωρίς-Κλίμακα (Scale-free)

Τα Scale-free δίκτυα είναι μια άλλη κατηγορία των σύνθετων δικτύων με βασικό χαρακτηριστικό την έλλειψη κλιμάκωσης του βαθμού του κόμβου. Το τελευταίο σημαίνει ότι δεν υπάρχει ομοιομορφία στην κατανομή των βαθμών

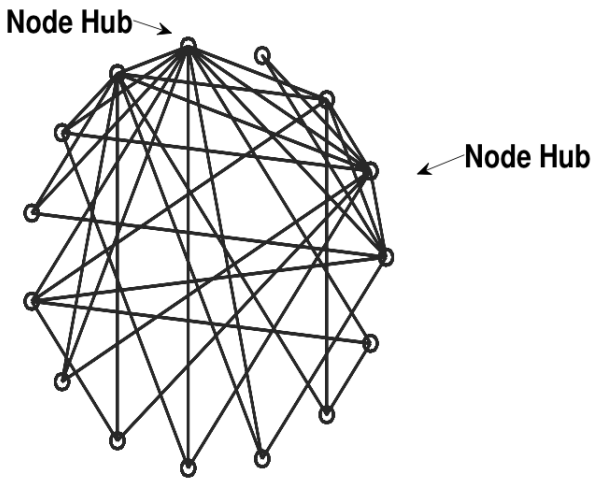


Σχήμα 3.5: Ένα δίκτυο κανονικού πλέγματος με βαθμό κόμβου 4 μετατρέπεται σε ένα small world δίκτυο μετά από ένα αριθμό τυχαίων αλλαγών των ακμών με πιθανότητα p να συμβεί η κάθε αλλαγή. Αυξάνοντας την πιθανότητα p των αλλαγών στις ακμές, η δομή του τελικού δικτύου καταλήγει σε τυχαίο γράφο [25].

στους κόμβους του δικτύου, δηλαδή παρατηρούνται σχηματικές διαφορές στη συνδεσιμότητα γειτονικών κόμβων. Η ανάπτυξη των συγκεκριμένων δικτύων και εν γένει των σύνθετων δικτύων βασίζονται σε δύο κανόνες, οι οποίοι δεν συναντώνται στο σχηματισμό των τυχαίων δικτύων [25]. Ο πρώτος κανόνας αναφέρεται στο ότι οι νέοι κόμβοι τείνουν να συνδέονται με ήδη υπάρχοντες. Ο δεύτερος κανόνας υπαγορεύει ότι όταν οι κόμβοι δημιουργούν νέες συνδέσεις, τείνουν να συνδέονται με κόμβους μεγαλύτερου βαθμού με μεγαλύτερη πιθανότητα. Ο κανόνας αυτός λέγεται 'σύνδεση κατά προτίμηση' ή (preferential attachment). Συνήθως ο κανόνας του preferential attachment θεωρεί γραμμική σχέση σύνθεσης ως προς τον βαθμό των κόμβων. Καταλήγει στη δημιουργία δύο ομάδων κόμβων στο τελικό δίκτυο. Ο τύπος του κανόνα preferential attachment δίνεται από τη σχέση (7) όπου x_i είναι μια παράμετρος η οποία καθορίζει την πιθανότητα επιλογής του κόμβου. Για παράδειγμα, στο μοντέλο Barabási-Albert (BA Model) η παράμετρος x_i ταυτίζεται με το βαθμό κόμβου

k_i του κόμβου i. Το μοντέλο Barabási-Albert (BA Model) είναι από τα πρώτα που χρησιμοποιούν τη 'σύνδεση κατά προτίμηση' για το σχηματισμό του δικτύου του. Το ΒΑ μοντέλο δεν προσπαθεί να μεταμορφώσει τη τοπολογική διάταξη του δικτύου με χρήση διαφόρων χαρακτηριστικών, αλλά προσπαθεί να μοντελοποιήσει την εξέλιξη της δυναμικής του δικτύου και πώς η τελική μορφή του δικτύου είναι αποτέλεσμα αυτής της εξέλιξης.

$$P(x_i) = \frac{x_i}{\sum_{\forall j} x_j} \tag{7}$$



Σχήμα 3.6: Ένα Scale-free δίκτυο που έχει δημιουργηθεί από το Barabási-Albert (BA Model) μοντέλο με 15 κόμβους και 3 συνδέσεις για κάθε νέο κόμβο. Στο σχήμα φαίνονται οι κόμβοι με τις περισσότερες συνδέσεις με ένδειξη 'hub' [25].

Οι κόμβοι του δικτύου μπορούν να διαχωριστούν σε δύο κατηγορίες με βάση το βαθμό τους. Η πρώτη αποτελείται από κόμβους με μεγάλο βαθμό που έχουν πολλές συνδέσεις και η δεύτερη, στην οποία ανήκει η πλειοψηφία των κόμβων, αποτελείται από κόμβους με μικρό βαθμό. Τα μεγάλα δίκτυα λόγω των παραπάνω ιδιοτήτων αυτό-οργανώνονται σε Scale-free δίκτυα και επιδεικνύουν δύο σημαντικές ιδιότητες. Αντιστέκονται σε τυχαία λάθη, αλλά είναι

ιδιαίτερα ευάλωτα σε συντονισμένες και στοχευμένες επιθέσεις. Ένα τυχαίο λάθος πιθανότατα θα πλήξει έναν κόμβο με μικρό βαθμό, καθώς οι τελευταίοι αποτελούν την πλειοψηφία, προκαλώντας περιορισμένη καταστροφή. Αντίθετα σε οργανωμένες επιθέσεις πλήττονται κατά κύριο λόγο οι πιο κεντρικοί κόμβοι καταστρέφοντας σε μεγαλύτερο βαθμό τη συνδεσιμότητα του δικτύου.

Για την κατασκευή του δικτύου Scale-free χρησιμοποιείται το μοντέλο Barabási-Albert (BA Model), το οποίο είναι το πρώτο που βασίζεται στους δύο παραπάνω κανόνες.

4 Κεφάλαιο
Διάχυση Πληροφορίας σε Δίκτυα

Η διάχυση της πληροφορίας είναι μία βασική διαδικασία κατά την οποία οι πληροφορίες, ασθένειες και άλλες συμπεριφορές διαχέονται στα δίκτυα. Η βασική προσέγγιση είναι η απλοποιημένη υπόθεση ότι οι συμπεριφορές (πληροφορία, ασθένειες) διαδίδονται στο περιβάλλον, το οποίο μοντελοποιείται χρησιμοποιώντας απλές δομές των κανονικών γραφημάτων και σπανιότερα των τυχαίων γράφων. Οι βασικές προσεγγίσεις των δικτύων έχουν ορισμένα μειονεκτήματα ως προς την ακρίβεια με την οποία μπορούν να μελετήσουν τη διάχυση της πληροφορίας σε πραγματικού-κόσμου (real) δίκτυα, καθώς δεν λαμβάνουν υπόψη τη τοπολογία ενός τέτοιου δικτύου, η οποία αποτελεί σημαντική γνώση για τη μοντελοποίηση και τη μελέτη της διάδοσης της πληροφορίας. Τα δίκτυα που σχηματίζονται από πραγματικά δεδομένα δεν αποτελούν ούτε κανονικά ούτε τυχαία γραφήματα, αλλά παρουσιάζουν μερικά ενδιαφέροντα χαρακτηριστικά. Τα χαρακτηριστικά αυτά επηρεάζουν σε μεγάλο βαθμό τη διάχυση της πληροφορίας μέσα στα πραγματικά δίκτυα.

4.1 Επιδημιολογικό Μοντέλο

Η χρήση του επιδημιολογικού μοντέλου συνεισφέρει στην εξαγωγή γνώσης σχετική με τη διάδοση και τον έλεγχο των 'ασθενειών' πάνω σε πληθυσμούς. Τα δίκτυα αποτελούν βασικές πλατφόρμες για την μελέτη της διάδοσης των 'ασθενειών'. Παρομοιάζοντας τη διάδοση των 'ασθενειών' με τη διάχυση της πληροφορίας στα δίκτυα, η χρήση του επιδημιολογικού μοντέλου είναι απαραίτητη για τη παρούσα εργασία. Κάνοντας μία εισαγωγή στις βασικές γνώσεις και στα επιμέρους μοντέλα του επιδημιολογικού φαινομένου μπορεί να γίνει καλύτερη κατανόηση της διάδοσης της πληροφορίας στο δίκτυο του Twitter και να μελετηθεί πιο αποτελεσματικά το μοντέλο που έχει θεωρηθεί ότι ακολουθεί αυτή.

Για τη χρήση του επιδημιολογικού μοντέλου σε δίκτυα υπολογιστών/τηλεπικοινωνιών γίνεται η θεώρηση ότι αναφέρεται σε ένα κλειστό ομογενή πληθυσμό ατόμων

που έχουν κοινά χαρακτηριστικά και εμφανίζουν παρόμοιες συμπεριφορές. Οι σύνδεσμοι μεταξύ των ατόμων παρουσιάζονται ως ένας μη κατευθυνόμενος γράφος G ο οποίος θεωρείται σταθερός κατά τη διάρκεια του φαινομένου της διάδοσης της 'ασθένειας'. Τη στιγμή $t = 0$ θεωρείται ότι υπάρχει ένας συγκεκριμένος αριθμός 'μολυσμένων' χρηστών στο δίκτυο, το αρχικό σύνολο I_0), βάση των οποίων γίνεται και η εξέλιξη του φαινομένου. Συγκεκριμένα για τα κοινωνικά δίκτυα κάθε 'μολυσμένος' χρήστης μπορεί να μεταδώσει την 'ασθένεια' μόνο σε όσους είναι συνδεδεμένοι άμεσα με αυτόν (έχουν απόσταση 1 hop), δηλαδή είναι συσχετισμένοι μέσα στο δίκτυο, και όχι σε όλους τους χρήστες του δικτύου, όπως θεωρείται στο απλό επιδημικό μοντέλο.

4.2 Το Υγιής-Μολυσμένος (Susceptible-Infected) μοντέλο

Το απλούστερο επιδημιολογικό μοντέλο είναι το δυναμικό μοντέλο Υγιής-Μολυσμένος (Susceptible- Infected). Στη συγκεκριμένη περίπτωση υπάρχουν δύο κατηγορίες χρηστών, οι 'υγιείς' (susceptibles) που συμβολίζονται με το σύνολο S και οι 'μολυσμένοι' (infectives) που αντιστοιχούν στο σύνολο I. Κάθε χρονική στιγμή t υπάρχει συγκεκριμένο πλήθος χρηστών στα δύο σύνολα και ισχύει :

$$N(t) = S(t) + I(t)$$

Όπου $S(t)$ αποτελεί το πλήθος των 'υγιών' χρηστών τη χρονική στιγμή t, $I(t)$ το πλήθος των 'μολυσμένων' χρηστών στο δίκτυο τη στιγμή t και $N(t)$ το σύνολο των χρηστών στο δίκτυο τη συγκεκριμένη χρονική στιγμή. Οι χρήστες μπορούν να μεταβούν από το S σύνολο στο I, όπου και παραμένουν μόνιμα σε αυτό, καθώς δεν υπάρχει η δυνατότητα ανάρρωσης. Εφόσον υπάρχει μία μοναδική και μονόδρομη μετάβαση μεταξύ των διαφορετικών καταστάσεων, το μοντέλο ονομάζεται S-I [21]. Θεωρώντας τη σταθερά γ ως την πιθανότητα με την οποία οι 'μολυσμένοι' χρήστες (infective) μπορούν να μολύνουν τους 'υγιείς', η διαφορική εξίσωση που περιγράφει την εξέλιξη της 'ασθένειας' διαμορφώνεται ως εξής:

$$\frac{dS}{dt} = -\gamma \cdot S(t) \cdot I(t)$$

Πιο συγκεκριμένα ο ρυθμός με τον οποίο μειώνονται οι 'υγιείς' χρήστες του δικτύου στη μονάδα του χρόνου, $\frac{dS}{dt}$ εξαρτάται από τον αριθμό των 'υγιών' και των 'μολυσμένων' χρηστών του δικτύου καθώς και από την πιθανότητα ένας 'μολυσμένος' χρήστης να μολύνει με τη σειρά του έναν 'υγιή'. Ο αριθμός των 'υγιών' χρηστών μπορεί μόνο να μειώνεται στο πέρασμα του χρόνου, καθώς δεν υπάρχει μηχανισμός ανάρρωσης. Στην αρχή της μελέτης του φαινομένου υπάρχει ένας συγκεκριμένος αριθμός 'μολυσμένων' χρηστών στον πληθυσμό, ο οποίος συμβολίζεται με I_0. Στο συγκεκριμένο μοντέλο το επιδημιολογικό φαινόμενο θα μεγαλώνει συνεχώς μέχρι ολόκληρος ο πληθυσμός να μολυνθεί εφόσον $\gamma > 0$. Στο SI μοντέλο το μακροπρόθεσμο αποτέλεσμα είναι τελικά να μολυνθούν όλοι οι χρήστες του πληθυσμού.

4.3 Το Υγιής-Μολυσμένος-Απομακρυσμένος (Susceptible-Infected-Removed) μοντέλο

Το Susceptible- Infected-Removed (SIR) είναι ένα απλό μοντέλο της διάδοσης μίας επιδημιολογικής 'ασθένειας' σε ένα μεγάλο πληθυσμό, το οποίο δημιουργήθηκε από τους Kermack και McKendrick [27]. Στο SIR μοντέλο γίνεται η θεώρηση ότι ο πληθυσμός αποτελείται από τρεις τύπους χρηστών, τους 'υγιείς' susceptibles (S)), 'μολυσμένους' infected (I) και 'απομακρυσμένους' removed (R). Πιο συγκεκριμένα :

- S(t) αντιπροσωπεύει τον αριθμό των χρηστών που δεν έχουν μολυνθεί ακόμα από την 'ασθένεια' τη χρονική στιγμή t.
- I(t) υποδηλώνει τον αριθμό των χρηστών τη χρονική στιγμή t , οι οποίοι έχουν μολυνθεί και είναι ικανοί να τη διαδώσουν και στο σύνολο των 'υγιών'.
- R(t) χρησιμοποιείται, για να απεικονίσει τους χρήστες οι οποίοι μολύνθηκαν και στη συνέχεια απομακρύνθηκαν από το σύνολο των μολυσμένων είτε λόγω 'ανάρρωσης' είτε λόγω 'θανάτου', τη χρονική στιγμή t. Οι χρήστες που ανήκουν σε αυτή την κατηγορία δεν γίνεται να μολυνθούν ξανά ή να διαδώσουν την ασθένεια σε άλλους χρήστες.

- $N(t) = S(t) + I(t) + R(t)$, ο συνολικός πληθυσμός του δικτύου.

Στο συγκεκριμένο μοντέλο, όπως και στο Υγιής-Μολυσμένος (SI) μοντέλο, οι νέες μολύνσεις συμβαίνουν ως αποτέλεσμα της επαφής μεταξύ ζευγών 'υγιών' και 'μολυσμένων' κόμβων. Αγνοώντας τις γεννήσεις και τους θανάτους για αυτό το μοντέλο, ο μοναδικός τρόπος με τον οποίο αλλάζει το μέγεθος του συνόλου S είναι μέσω μολύνσεως. Ο ρυθμός που γίνεται η μόλυνση μεταξύ των χρηστών εξαρτάται από τον αριθμό των 'υγιών' χρηστών, τον αριθμό των ήδη 'μολυσμένων', καθώς και από τη μεταξύ τους επαφή. Πιο συγκεκριμένα, θεωρείται ότι κάθε μολυσμένος χρήστης μπορεί να μολύνει υγιείς χρήστες με πιθανότητα β. Κατά αυτό τον τρόπο κάθε 'μολυσμένος' χρήστης μπορεί να μολύνει ένα συγκεκριμένο αριθμό 'υγιών' χρηστών παράγοντας β*S(t) νέους 'μολυσμένους' χρήστες. Συνολικά για όλους τους 'μολυσμένους' χρήστες ο αριθμός των μολύνσεων που παράγεται είναι:

$$\beta \cdot S(t) \cdot I(t)$$

Στην απλούστερη περίπτωση, η μόλυνση των 'υγιών' χρηστών είναι ο μοναδικός τρόπος μετάβασης από το σύνολο των 'υγιών' στο σύνολο των μολυσμένων και η διαφορική εξίσωση που περιγράφει αυτή τη διαδικασία, η οποία είναι παρόμοια με αυτή που ισχύει για το μοντέλο Υγιής-Μολυσμένος (SI), δίνεται από τη Σχέση (8). Η επόμενη μετάβαση που μπορεί να συμβεί σε αυτό το μοντέλο είναι από το σύνολο των μολυσμένων (I) στο σύνολο των απομακρυσμένων από το δίκτυο χρηστών (R), η οποία γίνεται με ρυθμό γ, θετική σταθερά και δίνεται από τη Σχέση (10).

$$\frac{dS(t)}{dt} = -\beta \cdot I(t) \cdot S(t) \tag{8}$$

$$\frac{dI(t)}{dt} = \beta \cdot I(t) \cdot S(t) - \gamma \cdot I(t) \tag{9}$$

$$\frac{dR(t)}{dt} = \gamma \cdot I(t) \tag{10}$$

Από τις παραπάνω διαφορικές εξισώσεις γίνεται φανερό ότι ο ρυθμός αλλαγής του πληθυσμού των 'μολυσμένων' χρηστών εξαρτάται από το ρυθμό με

τον οποίο αυξάνονται οι νέοι 'μολυσμένοι' χρήστες του δικτύου και ταυτόχρονα από το ρυθμό που 'μολυσμένοι' χρήστες αναρρώνουν, όπως φαίνεται και από τη Σχέση (9).

Στο παραπάνω μοντέλο που εξετάζεται ντετερμινιστικά μπορεί να φανεί από τις διαφορικές εξισώσεις ότι όταν εισάγεται σε ένα πληθυσμό ένα μικρό δείγμα 'μολυσμένων' χρηστών, ο αριθμός των infected χρηστών θα αυξηθεί, εάν ισχύει ότι $\beta > \gamma$. Στην αντίθετη περίπτωση το επιδημιολογικό φαινόμενο θα εξασθενήσει πολύ γρήγορα. Εξαιτίας αυτού, το πηλίκο $\frac{\beta}{\gamma}$ το οποίο αναφέρεται ως επιδημιολογικό κατώφλι (epidemic threshold) είναι μια πολύ σημαντική ποσότητα στη μελέτη του επιδημιολογικού μοντέλου [2]. Έχει αποδειχτεί ότι ισχύει η σχέση $\log(s(\infty)) = \frac{\beta}{\gamma} \cdot (s(\infty) - 1)$ [14]. Αυτό δεν οδηγεί σε μία κλειστή έκφραση για το $s(\infty)$, αλλά μπορεί να υπολογιστεί αριθμητικά. Η παραπάνω τιμή αντιστοιχεί στον αριθμό των χρηστών που δεν μολύνθηκαν τελικά από την ασθένεια. Επίσης πρέπει να σημειωθεί ότι η διάρκεια ενός επιδημικού φαινομένου δεν μπορεί να μελετηθεί με ρεαλιστικό τρόπο κάνοντας χρήση ενός ντετερμινιστικού μοντέλου, καθώς στην αρχή και στο τέλος αυτού στοχαστικές διακυμάνσεις κατά τη διάρκεια του κατέχουν σημαντικό ρόλο. [15]

Μία επέκταση του παραπάνω μοντέλου είναι να ληφθούν υπόψιν και οι ρυθμοί γεννήσεως και θανάτου στους πληθυσμούς. Θεωρείται ότι γεννήσεις σημειώνονται μόνο στο σύνολο των 'υγιών' χρηστών, ενώ θάνατοι μπορούν να εμφανιστούν και στις τρεις κατηγορίες πληθυσμού. Ο ρυθμός των θανάτων θεωρείται ίσος και για τα τρία διαφορετικά σύνολα, ενώ ταυτόχρονα οι δύο ρυθμοί λαμβάνονται ίσοι, ώστε ο συνολικός πληθυσμός να παραμένει σταθερός. Οι διαφορικές εξισώσεις που περιγράφουν τις μεταβάσεις μεταξύ των σχετικών συνόλων διαμορφώνονται ως εξής:

$$\frac{dS(t)}{dt} = -\beta \cdot I(t) \cdot S(t) - \mu \cdot S(t) + \mu \cdot (S(t) + I(t) + R(t))$$

$$\frac{dI(t)}{dt} = \beta \cdot I(t) \cdot S(t) - \gamma \cdot I(t) - \mu \cdot I(t)$$

$$\frac{dR(t)}{dt} = \gamma \cdot I(t) - \mu \cdot R(t)$$

Όπου μ αντιπροσωπεύει το ρυθμό γεννήσεων και θανάτων. Από τις παραπάνω διαφορικές εξισώσεις είναι φανερό ότι ο ρυθμός με τον οποίο αλλάζει ο αριθμός των 'υγιών' χρηστών στο δίκτυο εξαρτάται από την πιθανότητα οι 'μολυσμένοι' χρήστες να μολύνουν 'υγιείς', από το ρυθμό γεννήσεων του συνολικού πληθυσμού και το ρυθμό θανάτων των υγιών. Ο ρυθμός με τον οποίο αλλάζει ο πληθυσμός των 'μολυσμένων' εξαρτάται από την πιθανότητα μόλυνσης των 'υγιών', το ρυθμό ανάρρωσης των 'μολυσμένων' καθώς και το ρυθμό θανάτου των 'μολυσμένων'. Τέλος ο ρυθμός μεταβολής των 'απομακρυσμένων' χρηστών εξαρτάται από το ρυθμό ανάρρωσης των 'μολυσμένων', καθώς και από το ρυθμό θανάτου των χρηστών που έχουν ήδη αναρρώσει.

4.4 Το Υγιής-Μολυσμένος-Υγιής (SIS) μοντέλο

Στο μοντέλο SIS οι 'μολυσμένοι' χρήστες που ανήκουν στο σύνολο I επιστρέφουν στο σύνολο S των 'υγιών' χρηστών μετά την ανάρρωση τους, καθώς θεωρείται ότι μπορούν να μολυνθούν ξανά από την ίδια 'αρρώστια'. Στο απλούστερο SIS model οι διαφορικές εξισώσεις που περιγράφουν τη δυναμική της εξέλιξης είναι :

$$\frac{dS(t)}{dt} = -\beta \cdot S(t) \cdot I(t) + a \cdot I(t) \tag{11}$$

$$\frac{dI(t)}{dt} = \beta \cdot S(t) \cdot I(t) - a \cdot I(t) \tag{12}$$

Ο όρος $\beta \cdot S(t) \cdot I(t)$ εκφράζει τον αριθμό των μολύνσεων που προκαλούνται κάθε χρονική στιγμή ως συνάρτηση των 'υγιών' και των 'μολυσμένων' χρηστών, καθώς και μιας σταθεράς β, η οποία εκφράζει τον ρυθμό μολύνσεως των χρηστών. Ο όρος a αντιστοιχεί στο μέρος των μολυσμένων χρηστών οι οποίοι ανάρρωσαν και επανεντάχθηκαν στο σύνολο των υγιών. Αντίστοιχα, όπως και στα άλλα μοντέλα που αναφέρθηκαν παραπάνω, ισχύει ότι $S(t) + I(t) = N$ με N, δηλαδή το σύνολο του πληθυσμού σταθερό. [3]

5 Κεφάλαιο
Μελέτη Διάχυσης Πληροφορίας στο Twitter

5.1 Ορισμός του Δικτύου και Βασικές Έννοιες

Τα κοινωνικά δίκτυα αποτελούν έναν μηχανισμό αλληλεπίδρασης μεταξύ ανθρώπων, ο οποίος έχει εξελιχθεί ραγδαία στην εποχή του Internet. Καθώς οι χρήστες μπορούν να έρχονται σε επαφή με τους φίλους τους, τους συγγενείς τους και τις οικογένειες τους, ο αριθμός των ανθρώπων που χρησιμοποιούν τα κοινωνικά δίκτυα αυξάνεται εκθετικά. Για παράδειγμα, κοινωνικά δίκτυα όπως το Facebook και LinkedIn περιέχουν εκατομμύρια μέλη που κάνουν χρήση των υπηρεσιών τους, ώστε να διατηρούν επαφή με άλλους χρήστες, να βρίσκουν θέσεις εργασίας και να κάνουν ακόμα και εμπορικές συναλλαγές. Ταυτόχρονα, γίνεται εκμετάλλευση των δικτύων από εταιρίες και οργανισμούς και για διαφημιστικούς σκοπούς.

Παρατηρώντας τα κοινωνικά δίκτυα από ακαδημαϊκή άποψη, περιέχουν ένα μεγάλο όγκο πληροφορίας σχετικά με τη διαμόρφωση και τη δυναμική εξέλιξη αυτών των δικτύων. Την ίδια στιγμή η πρόσβαση σε αυτό τον όγκο πληροφορίας είναι εν μέρει δυνατή επιτρέποντας τη συλλογή δεδομένων για περαιτέρω ανάλυση και εξαγωγή συμπερασμάτων.

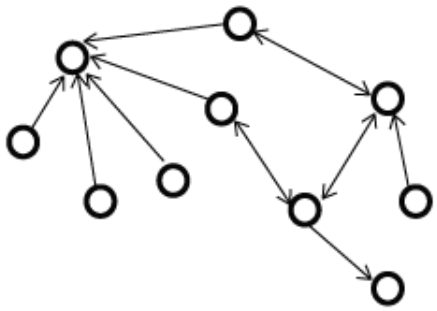
Σύμφωνα με το βασικό ορισμό ενός κοινωνικού δικτύου, κάθε χρήστης αλληλεπιδρά με όλο το σύνολο των επαφών που έχει και στην πραγματικότητα αυτό δεν συμβαίνει, καθώς κάθε άνθρωπος αλληλεπιδρά με ένα μικρό ποσοστό των επαφών του. Το γεγονός αυτό οφείλεται στον όγκο των δεδομένων που δέχονται σε συνδυασμό με τον περιορισμένο χρόνο να τα επεξεργαστούν, με αποτέλεσμα να επιλέγουν ορισμένες επαφές με τις οποίες αλληλεπιδρούν.

5.2 Το Δίκτυο Twitter

Το Twitter είναι ένα κοινωνικό δίκτυο το οποίο επιτρέπει στους χρήστες να παραθέτουν και να διαβάζουν σύντομα μηνύματα έως 140 χαρακτήρες τα οποία ονομάζονται tweets. Δημιουργήθηκε το Μάρτιο του 2006 από τους Jack Dorsey, Evan Williams, Biz Stone και Noah Glass και βασίστηκε στην ιδέα του Dorsey, όπου κάθε άτομο θα μπορεί να επικοινωνεί με ένα σύνολο ανθρώπων μέσω μιας υπηρεσίας μηνυμάτων [4] . Η υπηρεσία έγινε ταχύτατα δημοφιλής σε παγκόσμιο επίπεδο μετρώντας 100 εκατομμύρια χρήστες, οι οποίοι το 2012 δημοσίευαν περίπου 340 εκατομμύρια μηνύματα την ημέρα. Η υπηρεσία επίσης χειριζόταν 1.6 δισεκατομμύρια ερωτήματα ιστού (web search queries) κάθε μέρα το 2012. Το Δεκέμβριο του 2014 οι χρήστες του δικτύου ήταν πάνω από 500 εκατομμύρια, από τους οποίους περισσότεροι από 284 εκατομμύρια είναι ενεργοί χρήστες. Αυξάνεται με ταχείς ρυθμούς λόγω της λειτουργίας του σε πραγματικό χρόνο και της ποικιλίας των χρηστών του.

Το συγκεκριμένο κοινωνικό δίκτυο αποτελείται από χρήστες οι οποίοι εκμεταλλεύονται την υπηρεσία, για να δημοσιεύουν σύντομα μηνύματα, τα tweets, καθώς και να παρακολουθούν tweets άλλων χρηστών τους οποίους ακολουθούν. Κάθε χρήστης διαμορφώνει ένα κύκλο επαφών, ο οποίος αποτελείται από άλλους εγγεγραμμένους χρήστες του δικτύου. Υπάρχουν δύο ειδών συνδέσμων μεταξύ των ατόμων. Στο Twitter κάθε χρήστης μπορεί να 'ακολουθεί' (following) κάποιον άλλο χρήστη, έτσι ώστε να μπορεί να ενημερώνεται για τα μηνύματα που δημοσιεύει στην δική του προσωπική σελίδα. Αυτή είναι η βασική κοινωνική σχέση που υποστηρίζει το δίκτυο. Η σχέση δεν είναι συμμετρική, όπως σε άλλα κοινωνικά δίκτυα, καθώς αν κάποιος χρήστης ακολουθεί έναν άλλον, μπορεί ο δεύτερος να επιλέξει να μην ακολουθεί τον πρώτο και να μην ενημερώνεται για τα μηνύματά του. Ο δεύτερος βασικός σύνδεσμος μεταξύ χρηστών του δικτύου είναι ο χρήστης να έχει ακολούθους (followers), οι οποίοι ενημερώνονται και λαμβάνουν όλα τα μηνύματα που δημοσιεύει αυτός.

Οι βασικοί όροι που χρησιμοποιούνται στο δίκτυο είναι ο όρος follower, δηλαδή το άτομο που ακολουθεί έναν χρήστη καθώς και ο όρος following



Σχήμα 5.1: Απλό γράφημα που παρουσιάζει τη κατευθυνόμενη μη συμμετρική σύνδεση μεταξύ των χρηστών του Twitter.

που δηλώνει τους χρήστες που ακολουθεί κάποιος. Το tweet είναι ο βασικός τρόπος επικοινωνίας στο δίκτυο, καθώς αποτελεί ένα μήνυμα που στέλνει κάθε χρήστης στους followers του. Καθώς το δίκτυο βασίζεται στην ικανότητα των χρηστών να 'μοιράζονται' μηνύματα και πληροφορίες που βρίσκουν ενδιαφέρουσες, το Twitter υποστηρίζει έναν μηχανισμό αναδημοσίευσης μηνυμάτων, γνωστό ως retweet. Υπάρχουν δύο τρόποι αναδημοσίευσης ενός μηνύματος, μέσω της διεπαφής που προσφέρει η υπηρεσία είτε χειροκίνητα από τους χρήστες προσθέτοντας στην αρχή του κειμένου το "RT @username" όπου username είναι ο χρήστης που δημοσίευσε το αρχικό μήνυμα.

Η αλληλεπίδραση μεταξύ των χρηστών γίνεται και μέσω άμεσων δημόσιων μηνυμάτων χρησιμοποιώντας το σύμβολο @ και το όνομα του χρήστη στον οποίο απευθύνονται. Η διαδικασία αυτή είναι γνωστή ως mention. Τοποθετώντας το στην αρχή του μηνύματος περιορίζει το πλήθος των χρηστών που θα ενημερωθούν για τον συγκεκριμένο συμβάν, καθώς θα εμφανιστεί σε αυτούς που ακολουθούν και τους δύο χρήστες. Αν τοποθετηθεί οπουδήποτε αλλού μέσα στο μήνυμα τότε εμφανίζεται σε όλους που ακολουθούν τον χρήστη που το δημιουργεί. Μια ακόμα υπηρεσία που υποστηρίζει το δίκτυο είναι εκείνη των άμεσων ιδιωτικών μηνυμάτων DM, τα οποία, ενώ δεν διαφέρουν σε μέγεθος από τα απλά tweets, είναι προσβάσιμα μόνο μεταξύ του αποστολέα και

του παραλήπτη, δεδομένου ότι ο παραλήπτης ακολουθεί τον αποστολέα. Ένας πολύ δημοφιλής όρος που έχει αναδυθεί μέσω από το παραπάνω δίκτυο είναι ο όρος hashtag. Οι άνθρωποι στο Twitter εισάγουν τα "hashtags" μέσα στα μηνύματα τους για να ενισχύσουν το περιεχόμενο του μηνύματος τους και να κάνουν πιο εύκολη την κατηγοριοποίηση του. Το hashtag είναι απλά μια λέξη κλειδί στο οποίο προηγείται το σύμβολο της δίεσης, π.χ. #ekloges2015 που αναφέρεται στις εθνικές εκλογές του Ιανουαρίου του 2015. Με την τοποθέτηση μίας τέτοιας λέξης στο μήνυμα, οποιοσδήποτε αναζητήσει μηνύματα σχετικά με αυτό το περιεχόμενο θα μπορεί να δει και το συγκεκριμένο tweet και επομένως διευκολύνεται η εύρεση του συγκεκριμένου tweet και άλλων με το ίδιο hashtag. Τέλος, όταν στο δίκτυο υπάρχει ένα πολύ δημοφιλές θέμα για το οποίο δημοσιεύονται μεγάλος αριθμός μηνυμάτων, τότε το θέμα αυτό ονομάζεται "trending". Η κεντρική σελίδα του Twitter προσφέρει μια λίστα με τα πιο δημοφιλή θέματα τα οποία μπορούν να κατηγοριοποιούνται και αναλόγως την περιοχή στην οποία εμφανίζονται. Τα περισσότερα από αυτά τα θέματα είναι αρκετά εφήμερα και μπορούν να εξαφανιστούν από το προσκήνιο μέσα σε λίγες ώρες.

5.3 Μοντέλο Διάχυσης της Πληροφορίας στο Δίκτυο του Twitter

Η μελέτη της διάχυσης της πληροφορίας στο Twitter αποτελεί θέμα με το οποίο έχουν ασχοληθεί ήδη αρκετοί ερευνητές. Σε ορισμένες μελέτες υπάρχουν κάποιες βασικές διαφοροποιήσεις σε σχέση με το μοντέλο που ακολουθείται στην παρούσα διπλωματική [19], [20]. Ακολουθώντας τις βασικές έννοιες του επιδημιολογικού μοντέλου, θεωρούνται ως 'μολυσμένοι' (Infectious) εκείνοι που έχουν κάνει κάποια δημοσίευση ενώ ως 'υγιείς' (Susceptibles) θεωρούνται εκείνοι που ακολουθούν κάποιον 'μολυσμένο' χρήστη, αλλά δεν έχουν δημοσιεύσει ακόμα κάποια πληροφορία. Τέλος υπάρχουν και οι χρήστες που σταματούν να μπορούν εν δυνάμει να μολύνουν μετά από κάποιο χρονικό διάστημα (Recovered). Υπάρχουν οι κατάλληλα διαμορφωμένες διαφορικές εξισώσεις για να περιγράφουν τις μεταβάσεις από τη μία κατάσταση στην άλλη και να περιγράψουν τη διάδοση της πληροφορίας[4]. Η βασική διαφορά αυτής

[4]Οι διαφορικές εξισώσεις αναφέρονται στο παρών κεφάλαιο 13, 14

της θεώρησης με αυτή που ακολουθείται στην συγκεκριμένη εργασία είναι ο ορισμός των 'μολυσμένων' χρηστών. Στη σχετική εργασία, ως 'μολυσμένοι' θεωρούνται μόνο εκείνοι που κάνουν κάποια αναδημοσίευση της πληροφορίας ενώ στην παρούσα διπλωματική θεωρούνται όσοι έλαβαν γνώση της πληροφορίας ανεξάρτητα αν τη δημοσίευσαν ή όχι. Χρησιμοποιώντας τον τελευταίο ορισμό στην παρούσα διπλωματική, υπάρχει πιο πλήρης και ρεαλιστική περιγραφή της ροής της πληροφορίας στο δίκτυο, καθώς η διάδοση της πληροφορίας αφορά την ενημέρωση των χρηστών και όχι μόνο την ενημέρωση και αναδημοσίευση από αυτούς, δηλαδή δεν περιορίζεται μόνο στη διάδοση της πληροφορίας μέσω των retweets. Αντιθέτως, οι προηγούμενες προσεγγίσεις υποεκτιμούν τον αριθμό των ενημερωμένων κόμβων στο δίκτυο.

Το συγκεκριμένο κοινωνικό δίκτυο μπορεί να μοντελοποιηθεί μέσω ενός κατευθυνόμενου γράφου. Εάν ένας χρήστης Α ακολουθεί κάποιον χρήστη Β, τότε οι δύο χρήστες συνδέονται με ένα βέλος (connection) με κατεύθυνση από τον Β προς τον Α. Αντίστοιχα εάν ο Α ακολουθείται από τον Β τότε συνδέονται με βέλος από τον Α στον Β. Η κατεύθυνση του βέλους αντιπροσωπεύει την ροή της πληροφορίας [17]. Στο συγκεκριμένο κοινωνικό δίκτυο κάθε χρήστης έχει έναν αριθμό χρηστών που ακολουθεί. Αυτοί οι χρήστες ονομάζονται ως friends-following και ο χρήστης που τους ακολουθεί λαμβάνει οποιαδήποτε μήνυμα δημοσιεύουν. Ταυτόχρονα ορίζεται και ένα σύνολο χρηστών που ακολουθούν έναν χρήστη, οι οποίοι ονομάζονται ως followers, και ενημερώνονται για τα μηνύματα που ο ίδιος δημοσιεύει. Τα δύο αυτά σύνολα χρηστών δεν είναι απαραιτήτως ξένα μεταξύ τους. Κάθε χρήστης έχει την δυνατότητα να συνδεθεί με κάθε άλλο χρήστη. Καθότι υπάρχει κατεύθυνση στις ακμές, ορίζεται για κάθε κόμβο ο μέσα-βαθμός (in-degree) και ο έξω-βαθμός (out-degree). Ο μέσα-βαθμός αντιπροσωπεύει τον αριθμό των εισερχόμενων ακμών, δηλαδή αυτούς που ακολουθεί ο χρήστης-κόμβος, ενώ ο έξω-βαθμός αντιπροσωπεύει το πλήθος των εξερχόμενων ακμών και ταυτίζεται με τον αριθμό των χρηστών-κόμβων που ακολουθούν τον συγκεκριμένο χρήστη. Η ροή της πληροφορίας ταυτίζεται με την κατεύθυνση της σύνδεσης μεταξύ των χρηστών, καθώς η πληροφορία εξέρχεται από κάθε κόμβο και κατευθύνεται στους followers του κόμβου, ακολουθώντας τις συνδέσεις του γραφήματος.

Το Twitter μπορεί να χαρακτηρισθεί ως scale-free δίκτυο, δηλαδή έχει κατανομή power-law ως προς τον έξω-βαθμό των κόμβων του. Οι κόμβοι με μεγάλο έξω-βαθμό (out-degree) στο δίκτυο είναι σπανιότεροι, ενώ οι κόμβοι με μικρότερο είναι οι περισσότεροι, καθώς λίγοι κόμβοι έχουν πάρα πολλούς followers. Η κατανομή του in-degree των κόμβων ακολουθεί στην πλειοψηφία των χρηστών την κατανομή του out-degree. Οι περισσότεροι χρήστες έχουν κατά μέσο όρο τον ίδιο αριθμό ατόμων που ακολουθούν και τους ακολουθούν. Ένα μικρό ποσοστό κυρίως αυτό των νέων χρηστών έχουν περίπου τριπλάσιο αριθμό friends από ότι followers [10]. Τέλος, υπάρχει και ένα ποσοστό, περίπου $3-4\%$ των πιο δημοφιλών χρηστών οι οποίοι έχουν ποσοστό friends-followers 1 προς 3, τα στοιχεία αυτά το καθιστούν Power law δίκτυο [11], [26]. Θα μπορούσε επίσης να χαρακτηρισθεί ως ένα scale-free network, καθώς ακολουθεί τους δύο βασικούς μηχανισμούς ανάπτυξης των συγκεκριμένων δικτύων, δηλαδή πρώτον κάθε νέος χρήστης έχει την τάση να συνδέεται με ήδη υπάρχοντες χρήστες του δικτύου. Δεύτερον οι χρήστες δημιουργούν καινούριες συνδέσεις με υπάρχοντες χρήστες, με βάση τη δημοτικότητα τους, δηλαδή χρησιμοποιώντας τον κανόνα preferential attachment. Όσο πιο δημοφιλής είναι κάθε χρήστης τόσο μεγαλύτερη είναι η πιθανότητα να επιλεχθεί από νέους χρήστες του δικτύου, για να σχηματίσουν συνδέσμους μεταξύ τους.

Η μελέτη της διάδοσης πληροφορίας στο Twitter δύναται να γίνει επικεντρώνοντας σε ένα συγκεκριμένο θέμα, δηλαδή μηνύματα που περιέχουν ένα συγκεκριμένο κοινό στοιχείο. Εδώ θα θεωρηθεί ως κοινό στοιχείο το hashtag. Μελετώντας αυτή την περίπτωση, σε αντιστοιχία με το epidemic model, το δίκτυο μπορεί να χαρακτηρισθεί ως SI, καθώς θεωρείται ότι η ύπαρξη 'ανάρρωσης' των κόμβων δεν είναι εφικτή, εφόσον ένας κόμβος μπορεί πάντα να κάνει κάποια δημοσίευση με το συγκεκριμένο hashtag. Ως *Susceptibles* θεωρούνται οι κόμβοι-χρήστες οι οποίοι δεν έχουν δεχτεί την πληροφορία από κάποιο χρήστη, οπότε θεωρούνται ότι δεν έχουν λάβει γνώση για το συγκεκριμένο θέμα μηνυμάτων. Αντίστοιχα *Infected* ονομάζονται οι κόμβοι-χρήστες οι οποίοι είναι ενήμεροι για το θέμα από κάποιο friend τους είτε οι ίδιοι έχουν δημοσιεύσει κάποιο αντίστοιχο μήνυμα, όχι απαραιτήτως κάποιο retweet. Ορίζονται τα παρακάτω:

- S(t) = ο αριθμός των susceptible τη χρονική στιγμή t.
- I(t) = ο αριθμός των infected τη χρονική στιγμή t.
- N(t) = ο συνολικός αριθμός των κόμβων στο δίκτυο τη χρονική στιγμή t για τον οποίο ισχύει $N(t) = S(t) + I(t)$.
- λ_1 = η πιθανότητα αναδημοσίευσης ενός hashtag από έναν infected χρήστη.
- λ_2 = η πιθανότητα δημοσίευσης ενός μηνύματος, όπου περιέχει κάποιο hashtag.
- $\lambda_2^{'}$ = η πιθανότητα δημιουργίας ενός μηνύματος με συγκεκριμένο hashtag καθώς και κάποιο mention σε χρήστη.

Ο αριθμός των συνολικών κόμβων εξαρτάται από το χρόνο, καθώς το μέγεθος του δικτύου αλλάζει με ταχείς ρυθμούς. Τα ντετερμινιστικά μοντέλα διάδοσης πληροφορίας περιγράφονται με διαφορικές εξισώσεις, οι οποίες αναλύουν τις μεταβάσεις μεταξύ των διαφορετικών καταστάσεων, όπως παρουσιάστηκε στο επιδημιολογικό μοντέλο. Οι διαφορικές εξισώσεις δείχνουν τον ρυθμό με τον οποίο μεταβάλλονται τα σύνολα των υγιών και μολυσμένων με βάση το χρόνο. Η διάδοση της πληροφορίας βασίζεται στην ικανότητα των Infected κόμβων να μεταδίδουν την πληροφορία σε Susceptibles χρήστες. Οι διαφορικές εξισώσεις για την περιγραφή της δυναμικής διάδοσης της πληροφορίας στο Twitter διαμορφώνονται, με βάση τα παραπάνω μεγέθη, ως εξής :

$$\begin{aligned}\frac{dS(t)}{dt} = &-I(t) \cdot \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t) \cdot \lambda 1 - S(t) \cdot \lambda 2 - S(t) \cdot \lambda 2 \\ &\cdot \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t) - \lambda 2^{'} \cdot N(t) \cdot (S(t) - \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t))\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{dI(t)}{dt} = &I(t) \cdot \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t) \cdot \lambda 1 + S(t) \cdot \lambda 2 + S(t) \cdot \lambda 2 \\ &\cdot \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t) + \lambda 2^{'} \cdot N(t) \cdot (S(t) - \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t))\end{aligned} \tag{14}$$

Ως $\frac{dS}{dt}$ ορίζεται η μοναδιαία μεταβολή του αριθμού των υγιών κόμβων του δικτύου στη μονάδα του χρόνου. Το δεύτερο κομμάτι της εξίσωσης δηλώνει ότι ο ρυθμός που μειώνονται οι υγιείς στο δίκτυο είναι αποτέλεσμα αρκετών συνιστωσών. Ο πρώτος παράγοντας αποτελεί το γινόμενο των ήδη μολυσμένων χρηστών του δικτύου $I(t)$, το ποσοστό των υγιών χρηστών $\frac{S(t)}{N(t)}$, το μέσο αριθμό ακολούθων των χρηστών $K_{avg}^{out}(t)$ καθώς και την πιθανότητα αναδημοσίευσης ενός μηνύματος με συγκεκριμένο περιεχόμενο από ένα ήδη μολυσμένο χρήστη λ_1. Στη συνέχεια αφαιρείται το γινόμενο των υγιών χρηστών που κάνουν μια καινούρια δημοσίευση σχετική με το συγκεκριμένο θέμα με πιθανότητα λ_2 καθώς και οι followers αυτού του χρήστη, υπολογισμένοι σε με βάση το ποσοστό των υγιών και το μέσο αριθμό out-degree των κόμβων. Η παράμετρο $K_{avg}^{out}(t)$ αντιπροσωπεύει τον μέσο αριθμό εξερχόμενων ακμών για τους κόμβους, δηλαδή ένα μέσο αριθμό χρηστών που μπορούν να μολύνουν ανά μονάδα χρόνου. Τέλος με πιθανότητα $\lambda_2^{'}$ αφαιρούνται οι υγιείς χρήστες που γίνονται mention σε ένα 'μολυσμένο' μήνυμα και δεν είχαν ενημερωθεί προηγουμένως. Για το $\frac{dI}{dt}$ ισχύουν ακριβώς οι ίδιες επεξηγήσεις με τη διαφορά ότι εδώ αυξάνουν τον ενημερωμένο-μολυσμένο πληθυσμό.

5.4 Μηχανισμός για τη Συλλογή των Δεδομένων

Το Twitter παρέχει δεδομένα σε εξωτερικές υπηρεσίες μέσω ενός Application Programming Interface (API). Το API χαρακτηρίζεται ως RESTful, καθώς η πρόσβαση σε αυτό απαιτεί το χειρισμό κατάλληλων URLs με τη χρήση GET, POST requests, ώστε να διαμορφώσουν, να ζητήσουν και να χειριστούν δεδομένα από το API. Το Representational State Transfer (REST) αποτελεί τον κορμό του API. Επιτρέπει σε άλλους developers να έχουν πρόσβαση και να χειρίζονται διάφορα είδη δεδομένων, όπως δημοσιευμένα μηνύματα και κοινωνικές σχέσεις μεταξύ κόμβων. Το Streaming API επιτρέπει στους χρήστες να δημιουργούν συνδέσεις και να έχουν πραγματικού χρόνου πρόσβαση σε διάφορα υποσύνολα δημοσιευμένων μηνυμάτων στο δίκτυο. Επιπλέον δημιουργήθηκε το Search API που προσφέρει την ευκαιρία να ψάξει ο ενδιαφερόμενος χρήστης για πιο σύνθετες πληροφορίες, όπως αναδυόμενα trends. Οι αιτήσεις στο REST API είναι ανάλογες απλών ερωτημάτων (queries), οι συνδέσεις στο

Streaming API είναι αρκετά πιο μεγάλα σε διάρκεια ζωής ερωτήματα.

Το σύστημα ανακτά τα μηνύματα που δημοσιεύονται στις διάφορες ροές από το Twitter μέσω των αιτήσεων στο REST API. Σε περίπτωση που μία καινούρια ροή προστεθεί από ένα χρήστη, γίνεται η ανάκτηση του ιστορικού της. Το ιστορικό της χρησιμοποιείται με σκοπό την κατανόηση της φύσης της ροής με σκοπό το φιλτράρισμά της. Το API δέχεται τις απαντήσεις σε μορφή JavaScript Object Notation (JSON) και αναλύονται με JSON parser.

Αφού ληφθούν τα τελευταία μηνύματα, επεξεργάζονται, πριν αποθηκευτούν. Τα μηνύματα επεξεργάζονται με σκοπό να εξαχθούν διάφορα χαρακτηριστικά τα οποία μετά θα χρησιμοποιηθούν, για να αποφασιστεί, εάν είναι σχετικά με τη ροή. Αρχικά, τα links κάθε μηνύματος αναλύονται. Καθώς κάθε μήνυμα έχει περιορισμένο αριθμό χαρακτήρων (140), είναι πολύ συνηθισμένο να χρησιμοποιούνται συντομότερα URL. Αυτά κατά τη διαδικασία της ανάλυσης αντικαθίστανται με τα αρχικά URLs, των οποίων είναι συντομεύσεις. Αφού εντοπίσει το αρχικό URL μία υπηρεσία του API για εξαγωγή HTML κειμένου, επιστρέφει το κυρίως κείμενο της σελίδας απαλλαγμένο από διαφημίσεις και άλλα links. Αυτές οι δύο ενέργειες γίνονται παράλληλα και πολλές διαδικασίες περιμένουν, καθώς το αποτέλεσμα των HTTP αιτήσεων απαιτεί χρόνο.

Για να αποκτηθεί η δυνατότητα χειρισμού του API Twitter απαιτούνται πιστοποιητικά, με σκοπό τη σύνδεση στο δίκτυο και την εξουσιοδοτημένη αποστολή αιτήσεων σε αυτό. Για την απόκτηση αυτών, πρέπει ο χρήστης να δημιουργήσει μία εφαρμογή και στη συνέχεια να παράγει τα παρακάτω κλειδιά (keys-tokens). Μπορεί η εφαρμογή να έχει οποιοδήποτε όνομα και η δημιουργία της είναι απλή διαδικασία, η οποία ολοκληρώνεται ακολουθώντας μια σειρά βημάτων.

- *Consumer Key*
- *Consumer Secret*
- *Access Token*
- *Access Token Secret*

Το Twitter περιορίζει το σύνολο των αιτημάτων που γίνονται από κάθε χρήστη σε συγκεκριμένο αριθμό. Το όριο του REST API διαφέρει ανάλογα με το ερώτημα σε συγκεκριμένες κλήσεις την ώρα. Κατά συνέπεια αυτός ο περιορισμός λαμβάνεται υπόψιν κατά τη διαδικασία συλλογής των δεδομένων. [5]

Για τη συλλογή δεδομένων χρησιμοποιήθηκε μια βιβλιοθήκη της Java η Twitter4j που απευθύνεται 'ανεπισήμως' στο Twitter. Με τη συγκεκριμένη βιβλιοθήκη μπορεί ο χρήστης πολύ εύκολα να ενσωματώσει στη Java εφαρμογή του την υπηρεσία του Twiter. Η βιβλιοθήκη είναι συμβατή με τη νέα έκδοση ασφάλειας του δικτύου, Twitter API 1.1. Η ένταξη των έτοιμων εργαλείων στα προγράμματα γίνεται μέσω jar αρχείων. Στη συγκεκριμένη εργασία χρησιμοποιήθηκε το twitter4j-core-4.0.2 και το twitter4j-stream-4.0.2.Στη συνέχεια μέσω διαφόρων παρεχόμενων συναρτήσεων μπορεί ο χρήστης να συλλέξει δεδομένα που τον ενδιαφέρουν από το δίκτυο. Συγκεκριμένα χρησιμοποιήθηκε το REST APIs, για να γίνει στοχευμένη συλλογή δεδομένων αλλά και το Streaming API για την συλλογή μηνυμάτων σε πραγματικό χρόνο. [6]

Η μελέτη της διάδοσης της πληροφορίας στο Twitter διαφοροποιήθηκε σε δύο συγκεκριμένα πεδία, τη διάδοση συγκεκριμένου θέματος στο δίκτυο και τη δημιουργία μοντέλου για αυτή τη διαδικασία, καθώς και το ρυθμό με τον οποίο ένας χρήστης λαμβάνει νέες ειδοποιήσεις και ενημερώνεται η συλλογή των αποτελεσμάτων.

Για τη μελέτη του ρυθμού των εισερχόμενων πληροφοριών στην 'κεντρική-προσωπική σελίδα' κάθε χρήστη, δεν υπάρχει καθορισμένη παρεχόμενη εντολή για την άμεση συλλογή της συγκεκριμένης πληροφορία. Εξαιτίας αυτού χρησιμοποιήθηκε ο συνδυασμός διαφορετικών requests για τη συλλογή των δεδομένων. Για κάθε χρήστη που μελετήθηκε βρέθηκε το σύνολο των χρηστών που ακολουθεί. Με βάση το δίκτυο και τη ροή της πληροφορίας, όπως έχει θεωρηθεί, οι χρήστες που ακολουθεί ένας κόμβος είναι η πηγή της πληροφορίας που δέχεται. Στη συνέχεια για ένα συγκεκριμένο χρονικό διάστημα, συλλέχθηκαν μέχρι και 100 μηνύματα που παρέθεσαν οι χρήστες φίλοι του

εξεταζόμενου κόμβου. Να σημειωθεί ότι από το σύνολο των tweets έχουν αφαιρεθεί τα μηνύματα των χρηστών, οι οποίοι κατακλύζουν το αρχικό κόμβο με πάρα πολλά μηνύματα, πολύ περισσότερα από ότι οι υπόλοιποι friends. Η αφαίρεση αυτή έγινε με σκοπό την αποφυγή επικράτησης μικρού αριθμού χρηστών έναντι του συνόλου. Μετά από την επεξεργασία τόσο του χρόνου άφιξης των μηνυμάτων, όσο και του διαστήματος μεταξύ των αφίξεων, μπορούν να εξαχθούν ορισμένα συμπεράσματα για την κατανομή που ακολουθούν αυτές οι μεταβλητές, όπως φαίνεται αναλυτικότερα στο κεφάλαιο 5.5.

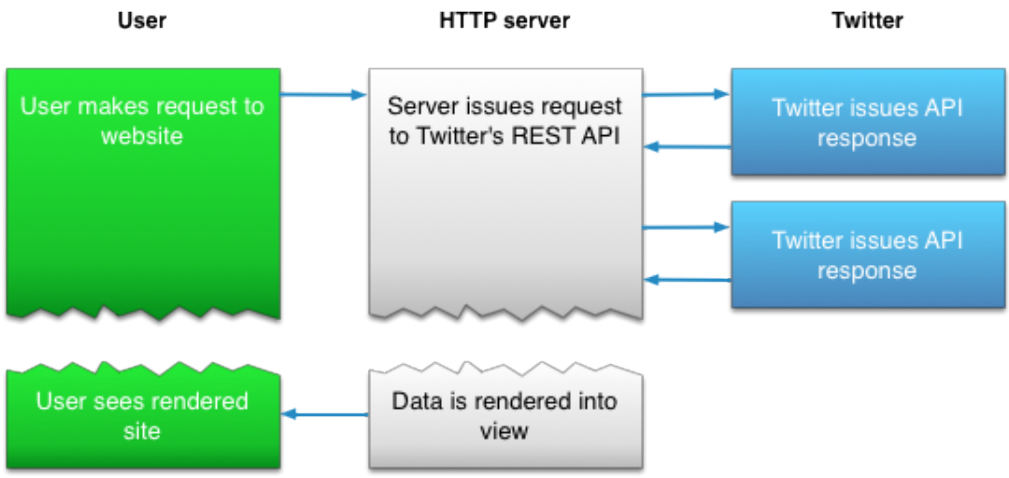
Πέρα από τη στοχευμένη εύρεση πληροφοριών σχετικών με χρήστες του δικτύου απαιτήθηκε και η συλλογή μηνυμάτων σε πραγματικό χρόνο. Για τη συνεχή συλλογή των tweets χρησιμοποιήθηκε το Streaming API. Για την αποθήκευση των δεδομένων, ώστε να γίνει offline επεξεργασία τους χρησιμοποιήθηκε η non-sql βάση δεδομένων MongoDB. Επιλέχθηκε ένα hashtag ως κριτήριο για την επιλογή των tweets, ώστε να μελετηθεί η διάδοση της πληροφορίας για ένα θέμα ευρύτερα και όχι για ένα μοναδικό μήνυμα. Το φιλτράρισμα των tweets από τη ροή γίνεται δίνοντας ένα σύνολο λέξεων τα οποία πρέπει να περιέχονται στα μηνύματα που επιστρέφονται. Η συλλογή των μηνυμάτων επιτυγχάνεται, καθώς μέσω της εφαρμογής τοποθετείται ένας StatusListener, ο οποίος καταγράφει κάθε νέο tweet που έρχεται στη συγκεκριμένη ροή. Να σημειωθεί ότι ο αριθμός των μηνυμάτων που επιστρέφονται από την υπηρεσία δεν είναι απεριόριστος, καθώς κάθε ώρα υπάρχει ένας συγκεκριμένο όριο δεδομένων που επιστρέφεται. Στη συνέχεια αποθηκεύεται στη βάση ο κωδικός του χρήστη που δημοσιεύει το μήνυμα, η χρονική στιγμή που γίνεται η δημοσίευση και ο κωδικός του μηνύματος. Η επιλογή του θέματος, σχετικά με το οποίο θα γίνει η συλλογή των μηνυμάτων και θα μελετηθεί η μόλυνση των χρηστών, έγινε με κάποιο δημοφιλές hashtag εκείνης της περιόδου.

Μετά την αποθήκευση όλων των δεδομένων στη βάση έγινε η επεξεργασία αυτών, ώστε να σχηματισθεί μία εικόνα της διάδοσης της πληροφορίας. Όλα τα tweets έχουν αποθηκευτεί για συγκεκριμένο χρονικό διάστημα. Το παραπάνω διάστημα χωρίζεται σε υποδιαστήματα μέσα στα οποία εξετάζεται η αύξηση του συνολικού αριθμού μολυσμένων χρηστών στο δίκτυο. Αυτό ε-

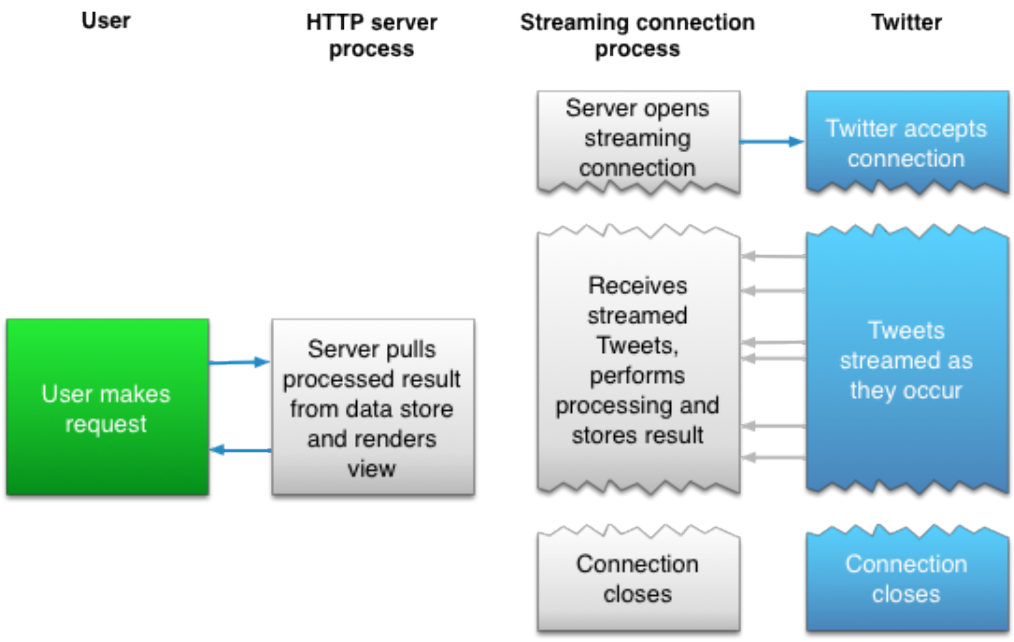
Hashtag	Αριθμός tweets	Αριθμός βημάτων	Αριθμός χρηστών
ekloges2015	8095	48	3132
2014moments	7111	30	4231

Πίνακας 5.1: Στοιχεία για τα δύο θέματα που μελετήθηκαν.

πιτυγχάνεται καθώς για κάθε χρήστη που κάνει σχετική με το συγκεκριμένο hashtag δημοσίευση, προστίθεται στο σύνολο των ήδη μολυσμένων. Ταυτόχρονα προστίθενται και οι χρήστες που τον ακολουθούν, καθώς σύμφωνα με το μοντέλο, ανήκουν στο σύνολο των μολυσμένων. Βέβαια κάθε φορά που γίνεται μια προσθήκη στο σύνολο αυτό, έχει προηγηθεί έλεγχος, ώστε κάθε χρήστης να προστίθεται μόνο μία φορά και να είναι ακριβές το σύνολο των μολυσμένων. Στο τελευταίο στάδιο έχουν διαμορφωθεί τα σύνολα των μολυσμένων χρηστών για κάθε υποδιάστημα.



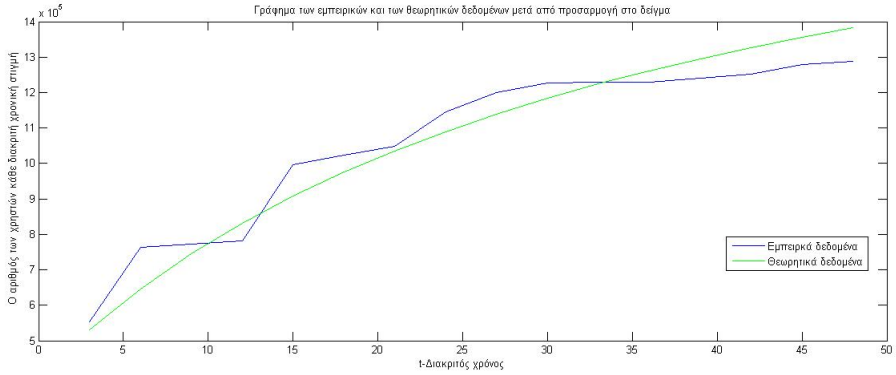
Σχήμα 5.2: Παράδειγμα λειτουργίας του REST API του Twitter.



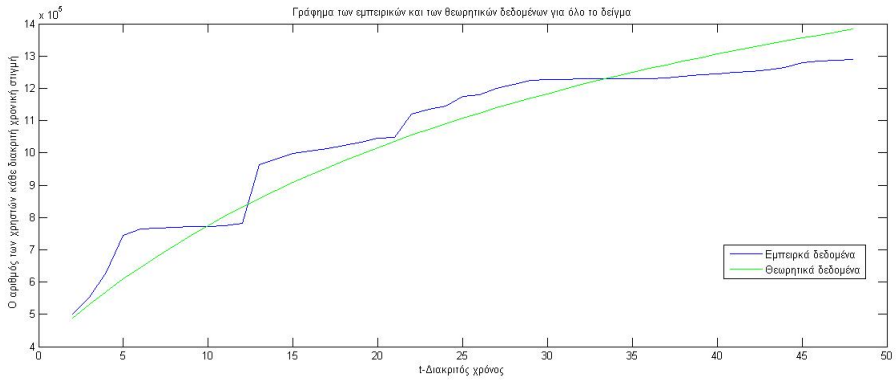
Σχήμα 5.3: Παράδειγμα λειτουργίας του Streaming API του Twitter.

5.5 Επαλήθευση Μοντέλου για τη Διάδοση Πληροφορίας στο Twitter

Παραπάνω έγινε η εισαγωγή και περιγραφή του μηχανισμού συλλογής και επεξεργασίας των δεδομένων που χρησιμοποιήθηκε για την μελέτη της δυναμικής της διάδοσης της πληροφορίας στο κοινωνικό δίκτυο του Twitter και συγκεκριμένα με σκοπό την επαλήθευση του μοντέλου που κατασκευάστηκε. Όπως αναφέρθηκε, έγινε συλλογή tweets με συγκεκριμένο hashtag για συγκεκριμένη χρονική περίοδο και στη συνέχεια έγινε offline επεξεργασία αυτών των μηνυμάτων. Τα αποτελέσματα που παρήχθησαν αφορούν το σύνολο των μολυσμένων χρηστών κάθε χρονική στιγμή t όπου $t = 1, 2, 3.., k$, k το συνολικό διάστημα χρόνου κατά το οποίο έγινε η καταγραφή. Χρησιμοποιήθηκαν δύο ξεχωριστά θέματα (hashtags) το ekloges2015 και το 2014moments. Όπως φαίνεται και στον πίνακα 5.1, για το πρώτο θέμα συγκεντρώθηκαν 8095



(α') Γράφημα εμπειρικών και θεωρητικών δεδομένων για μέρος του δείγματος και για πληθυσμό $21 \cdot 10^5$

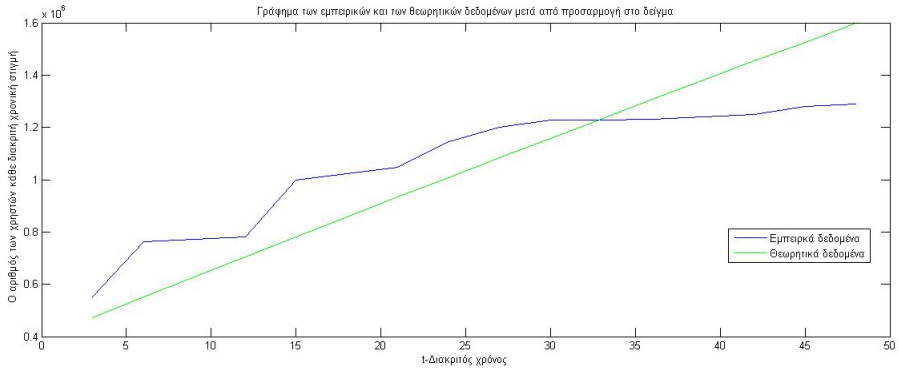


(β') Γράφημα εμπειρικών και θεωρητικών δεδομένων για όλο το δείγμα και για πληθυσμό $21 \cdot 10^5$

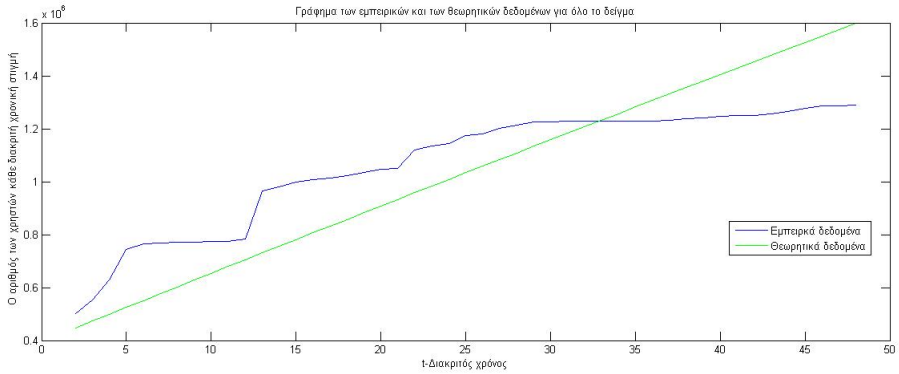
Σχήμα 5.4: Μετρήσεις για το hashtag "ekloges2015" για πληθυσμό $21 \cdot 10^5$

μηνύματα σε διάστημα 48 ωρών. Στον πίνακα ο αριθμός των βημάτων ταυτίζεται με τον αριθμό των ωρών που διήρκεσε η συλλογή των δεδομένων. Στην τρίτη στήλη δηλώνεται ο αριθμός των χρηστών που δημοσίευσαν αυτά τα μηνύματα. Στη συνέχεια, με χρήση της συνάρτησης της Matlab lsqcurvefit έγινε προσαρμογή της εξίσωσης του μοντέλου, Σχέση (14), για να υπολογιστούν οι παράμετροι της.

Πιο συγκεκριμένα η συνάρτηση lsqcurvefit έχει τη δυνατότητα να λύσει μη-γραμμικά προβλήματα ελαχίστων τετραγώνων, καθώς της παρέχονται πλη-



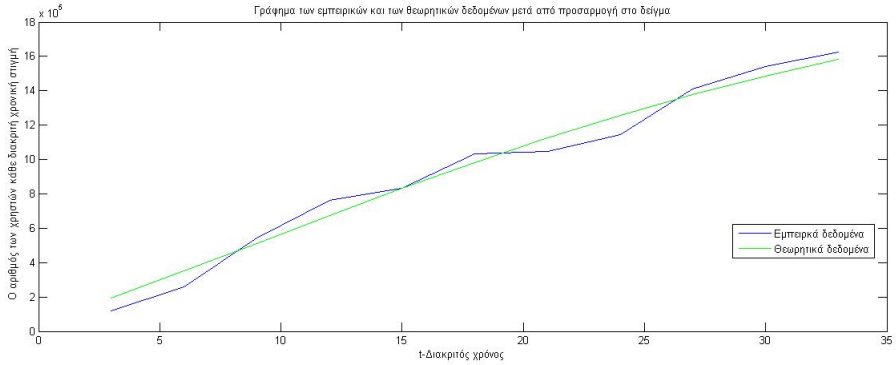
(α') Γράφημα εμπειρικών και θεωρητικών δεδομένων για μέρος του δείγματος και για πληθυσμό $21 \cdot 10^6$



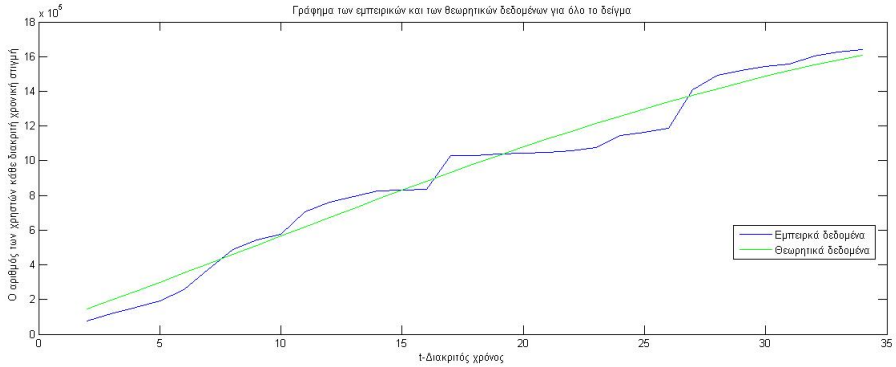
(β') Γράφημα εμπειρικών και θεωρητικών δεδομένων για όλο το δείγμα και για πληθυσμό $21 \cdot 10^6$

Σχήμα 5.5: Μετρήσεις για το hashtag "ekloges2015" για πληθυσμό $21 \cdot 10^6$

ροφορίες για τη μη γραμμική συνάρτηση στην οποία εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων, αρχικές τιμές για τις παραμέτρους που θέλει να βελτιστοποιήσει, τα επιθυμητά αποτελέσματα της μη γραμμικής συνάρτησης, τα οποία θέλει να προσεγγίσει βέλτιστα και οποιαδήποτε άλλα δεδομένα είναι απαραίτητα. Το αποτέλεσμα που επιστρέφει είναι οι βέλτιστες τιμές των παραμέτρων προς υπολογισμό, ώστε τα εμπειρικά δεδομένα να 'πλησιάζουν' όσο το δυνατόν καλύτερα τα θεωρητικά δεδομένα. Η lsqcurvefit ξεκινά από τις αρχικές τιμές των παραμέτρων και βρίσκει το διάνυσμα αυτών που προσαρμόζουν καλύτερα τη μη γραμμική συνάρτηση στα δεδομένα που δόθηκαν.



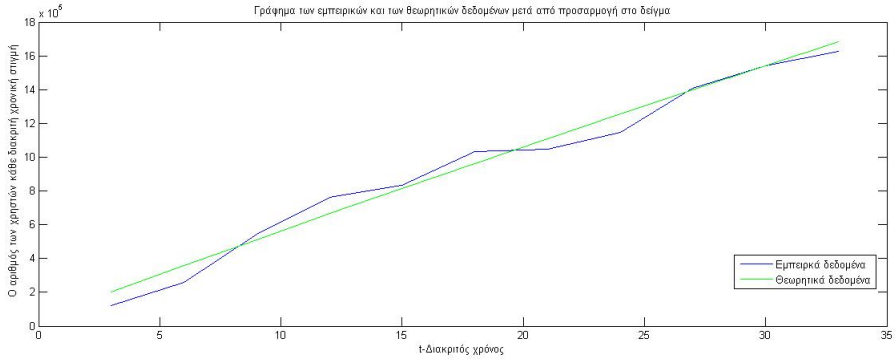
(α') Γράφημα εμπειρικών και θεωρητικών δεδομένων για μέρος του δείγματος και για πληθυσμό $21 \cdot 10^5$



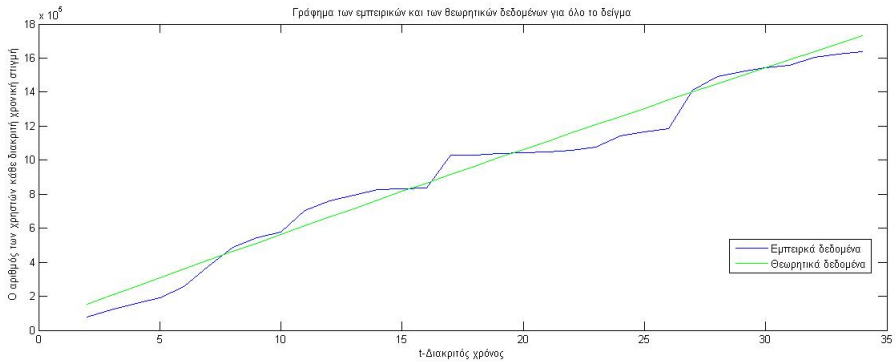
(β') Γράφημα εμπειρικών και θεωρητικών δεδομένων για όλο το δείγμα και για πληθυσμό $21 \cdot 10^5$

Σχήμα 5.6: Μετρήσεις για το hashtag "2014moments" για πληθυσμό $21 \cdot 10^5$

Η συνάρτηση στην οποία γίνεται η προσαρμογή μέσω του εργαλείου Matlab αποτελεί τη λύση της διαφορικής εξίσωσης για τον αριθμό των μολυσμένων χρηστών $I(t)$ που παρουσιάστηκε παραπάνω στην εξίσωση (14). Από τη διαφορική εξίσωση αγνοήθηκε ο όρος $\beta' \cdot N(t) \cdot (S(t) - \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t))$, καθώς η συμβολή του όρου αυτού είναι αρκετά μικρή, ώστε να μπορεί να αγνοηθεί. Θεωρήθηκε πολύ μικρή η πιθανότητα να γίνει κάποια δημοσίευση με το συγκεκριμένο θέμα και να γίνεται mention κάποιος συγκεκριμένος χρήστης, οπότε η συμβολή στις αλλαγές των συνόλων των υγιών και των μολυσμένων είναι αρκετά μικρή που μπορεί να αγνοηθεί. Τελικά, η διαφορική



(α') Γράφημα εμπειρικών και θεωρητικών δεδομένων για μέρος του δείγματος και για πληθυσμό $21 \cdot 10^5$



(β') Γράφημα εμπειρικών και θεωρητικών δεδομένων για όλο το δείγμα και για πληθυσμό $21 \cdot 10^6$

Σχήμα 5.7: Μετρήσεις για το πρώτο hashtag "2014moments" για πληθυσμό $21 \cdot 10^6$

εξίσωση περιορίζεται στη μορφή :

$$\frac{dI}{dt} = I(t) \cdot \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t) \cdot \beta 1 + S(t) \cdot \beta 2 + S(t) \cdot \beta 2 \cdot \frac{S(t)}{N(t)} \cdot K_{avg}^{out}(t) \quad (15)$$

Αφού υπολογίστηκε μέσω κατάλληλου εργαλείου του Mathematica η λύση της εξίσωσης (15) έχει τη μορφή :

$$\frac{e^{C \cdot t + A \cdot N \cdot t} \cdot N + C \cdot \left(\frac{D-N}{C+A \cdot D - B \cdot D + B \cdot N}\right) + B \cdot N \cdot \left(\frac{D-N}{C+A \cdot D - B \cdot D + B \cdot N}\right)}{e^{C \cdot t + A \cdot N \cdot t} - A \cdot \left(\frac{D-N}{C+A \cdot D - B \cdot D + B \cdot N}\right) + B \cdot \left(\frac{D-N}{C+A \cdot D - B \cdot D + B \cdot N}\right)} \quad (16)$$

όπου :

- $A = \frac{\lambda 1 \cdot K_{avg}}{N}$
- $A = \frac{\lambda 2 \cdot K_{avg}}{N}$
- $C = \lambda 2$
- $D =$ αρχικός αριθμός μολυσμένων χρηστών στο δείγμα

Από το σύνολο των εμπειρικών δεδομένων που είχαν συλλεχθεί αρχικά χρησιμοποιήθηκε ένα μέρος αυτών, ώστε να γίνει η προσαρμογή και εύρεση των βέλτιστων παραμέτρων. Αυτό αποτυπώνεται στα σχήματα 5.6α΄,5.4α΄, 5.5α΄, 5.7α΄. Στη συνέχεια έγινε η χρήση του συνόλου των δεδομένων για επαλήθευση της ορθότητας των υπολογισμένων παραμέτρων, καθώς και του θεωρημένου μοντέλου. Η αξιολόγηση του μοντέλου γίνεται στα σχήματα 5.6β΄, 5.4β΄, 5.5β΄, 5.7β΄.

Αφού γίνει ο υπολογισμός των παραμέτρων, πραγματοποιείται υπολογισμός των θεωρητικών αποτελεσμάτων από τη λύση της εξίσωσης του μοντέλου διάχυσης πληροφορίας εξίσωση (16), για να γίνει η σύγκριση με τα εμπειρικά αποτελέσματα που έχουν συλλεχθεί και να γίνει αξιολόγηση της ακρίβειας του μοντέλου. Θα πρέπει να σημειωθεί ότι έχουν γίνει κάποιες παραδοχές σε σχέση με τον αριθμό των χρηστών που θεωρούνται ως το συνολικό πλήθος κόμβων στο δίκτυο. Όπως αναφέρθηκε παραπάνω στη συλλογή των δεδομένων, η υπηρεσία του Twitter επιστρέφει περιορισμένο αριθμό tweets ανά ώρα, οπότε θεωρήθηκε σωστό να μην λαμβάνονται ως πληθυσμός το σύνολο των ενεργών χρηστών του Twitter, καθώς από το σύνολο των δεδομένων έχουν αγνοηθεί πιθανώς ένα μεγάλο ποσοστό μηνυμάτων. Ταυτόχρονα επιλέχθηκαν hashtags, τα οποία ήταν αρκετά δημοφιλή σε συγκεκριμένες περιοχές (Αυστραλία, Ελλάδα), ώστε να είναι πιο βατή η διαδικασία της επεξεργασίας των μηνυμάτων και εύλογη από άποψη χρόνου. Δεδομένου αυτού ο πληθυσμός που λαμβάνεται υπόψιν είναι μικρότερος σε σχέση με τους καταγεγραμμένους ενεργούς χρήστες του δικτύου. Παρουσιάζεται η προσαρμογή στο μοντέλο λαμβάνοντας υπόψιν το συνολικό πληθυσμό, όπου θα φανεί ότι τα αποτελέσματα δεν είναι τόσο ικανοποιητικά σε σχέση με τα αποτελέσματα που εξάγονται όταν γίνουν οι παραπάνω θεωρήσεις, σχήματα 5.5β΄ και 5.7β΄.

User	Followers	Following
1	367	186
2	662	277
3	1.87M	263

Πίνακας 5.2: Πίνακας με στοιχεία για τις συνδέσεις τριών χρηστών

5.6 Ανάλυση Ρυθμού Ενημέρωσης Χρηστών

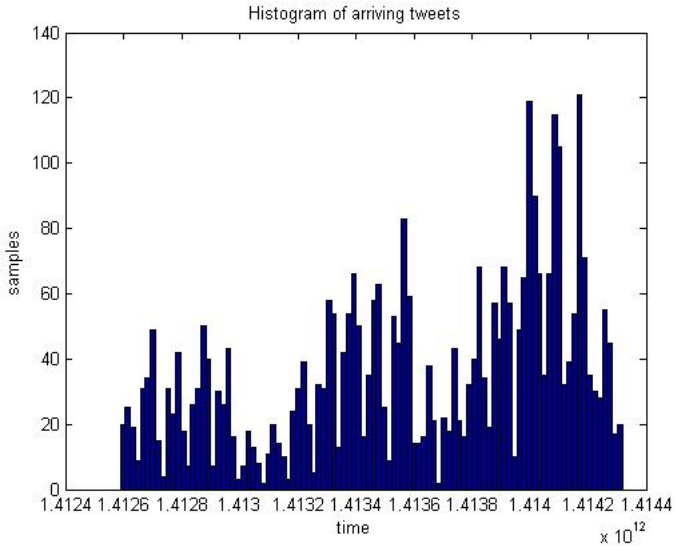
Στην 'κεντρική-προσωπική σελίδα' κάθε χρήστη φτάνουν μηνύματα που έχουν δημοσιεύσει οι χρήστες που ακολουθεί ο συγκεκριμένος χρήστης. Αναλόγως με τον αριθμό των friends του διαμορφώνεται και ο όγκος της πληροφορίας που δέχεται. Για κάθε επιμέρους χρήστη συλλέχθηκαν τα 100 τελευταία tweets των friends του, από τα οποία διατηρήθηκαν εκείνα που είχαν γίνει μέσα σε ένα συγκεκριμένο χρονικό διάστημα. Ταυτόχρονα αφαιρέθηκαν όλα τα μηνύματα των φίλων που είχαν πραγματοποιήσει και τις 100 δημοσιεύσεις σε λιγότερο από 3 ημέρες. Ο συνολικός χρόνος εξέτασης είναι περίπου 20 ημέρες και προέκυψε εμπειρικά μελετώντας την κατανομή των χρόνων της εισερχόμενης πληροφορίας για τους συγκεκριμένους χρήστες, καθώς κατά τη συλλογή των χρόνων παρατηρήθηκαν αρκετά μηνύματα που είχαν αυτό το χρόνο δημοσίευσης και δεν θα μπορούσαν να αγνοηθούν. Αυτή η παραδοχή έγινε, ώστε να αποφευχθεί η ύπαρξη ελάχιστων χρηστών που θα επιβάλλουν το δικό τους ρυθμό στη δημοσίευση των μηνυμάτων. Ταυτόχρονα για κάποιους χρήστες που έχουν 100 μηνύματα μέσα σε τρεις ημέρες σίγουρα θα έχουν αρκετά και τις προηγούμενες 17 ημέρες, τα οποία δεν συλλέχθηκαν και τα οποία δεν πρέπει να αγνοηθούν. Για αυτό το λόγο επιλέγεται να μην προσμετρηθούν οι χρήστες αυτοί.

Για κάθε χρήστη που μελετήθηκε δημιουργήθηκε ένα ιστόγραμμα για τον όγκο των μηνυμάτων που δέχεται, καθώς και τους αντίστοιχους χρόνους αυτών. Ταυτόχρονα μελετήθηκε και το διάστημα που μεσολαβεί μεταξύ συνεχόμενων εισερχόμενων μηνυμάτων (interarrival time). Με χρήση του εργαλείου dfittool Matlab έγινε η προσέγγιση των δεδομένων μέσω δύο διαφορετικών κατανομών, της εκθετικής (Exponential) και της Generalized Pareto. Το δεύτε-

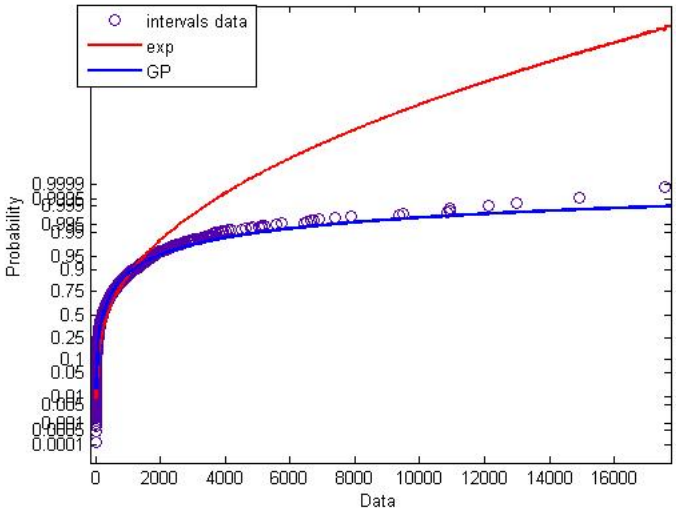
ρο σχήμα αποτελεί ένα probability plot των δύο κατανομών και των δεδομένων. Στον οριζόντιο άξονα αναφέρονται ο αριθμός των εισερχομένων μηνυμάτων, ενώ στον κάθετο η αθροιστική πιθανότητα να παρατηρηθεί τέτοιο μεσοδιάστημα. Το σχήμα αυτό αποτελεί στην ουσία ένα εμπειρικό σχήμα της αθροιστικής συνάρτησης πιθανότητας. Ο άξονας y είναι αριθμημένος από το 0 έως το 1, καθώς αντιπροσωπεύει πιθανότητα, αλλά η κλίμακα δεν είναι γραμμική. Όσο πιο κοντά πέφτουν τα εμπειρικά δεδομένα σε κάποια κατανομή είναι πιο εύλογο να χρησιμοποιηθεί εκείνη για την μοντελοποίηση της κατανομής των χρόνων.

Από τα σχήματα και συγκεκριμένα από το probability plot παρατηρείται ότι η generalized pareto κατανομή προσομοιώνει τα εμπειρικά δεδομένα σε καλύτερο βαθμό από την εκθετική κατανομή. Η generalized pareto ταιριάζει στα δεδομένα περίπου στο 99%, ενώ για τις ακραίες τιμές παρουσιάζει ορισμένες διακυμάνσεις. Οι ακραίες τιμές αναλόγως με την προσομοίωση και το χρήστη διαφέρουν. Παρόλα αυτά στους περισσότερους χρήστες παρατηρήθηκε ότι τα μεγαλύτερα διαστήματα μεταξύ συνεχόμενων tweets που δέχονται είναι λιγότερο πιθανό να συμβούν και ανήκουν στο 1% των δεδομένων.

Παρατηρήθηκε ότι η χρήση της εκθετικής κατανομής για την μοντελοποίηση του ρυθμού της εισερχόμενης πληροφορίας δεν είναι η καταλληλότερη, καθώς τα ενδιάμεσα διαστήματα των αφίξεων δεν ακολουθούν αυτή την κατανομή. Βάση αυτού, η μοντελοποίηση για τη διάδοση της πληροφορίας στο δίκτυο δεν μπορεί να βασιστεί σε ένα μοντέλο ουρών, όπου ο ρυθμός μετάβασης από τη μία κατάσταση στην επόμενη χαρακτηρίζεται από την εκθετική κατανομή. Εξαιτίας αυτού μία διαφορετική περιγραφή του μοντέλου που ακολουθήθηκε στη συγκεκριμένη διπλωματική βασίζεται στο epidemic model SI.

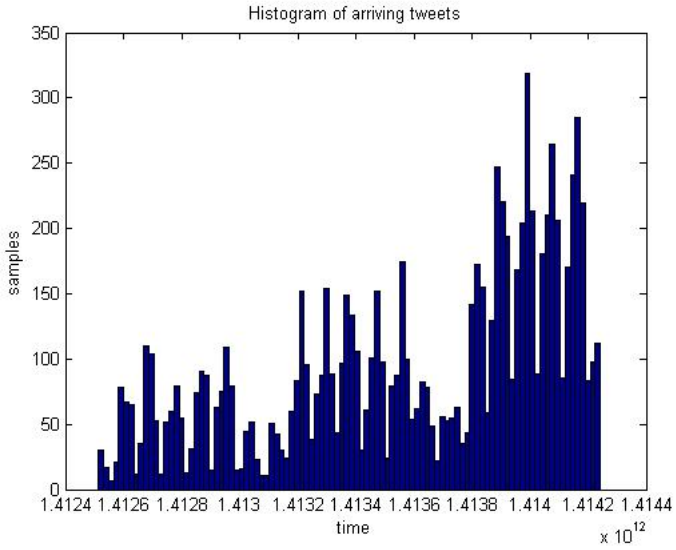


(α') Ιστόγραμμα των αφιχθέντων tweets

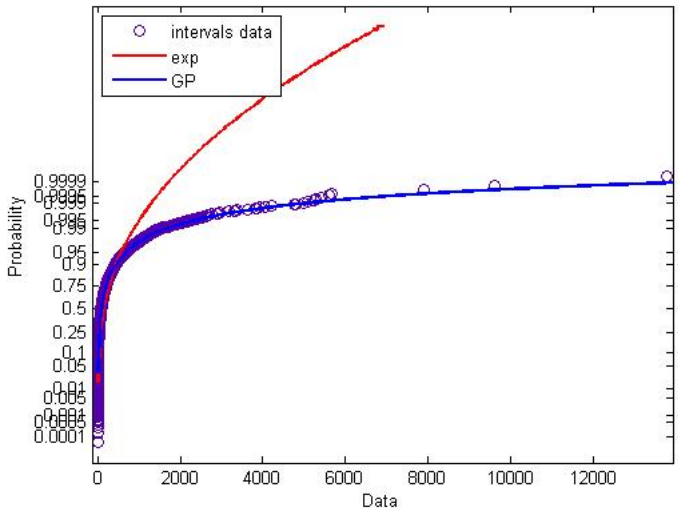


(β') Συνάρτηση Πιθανότητας και σύγκριση με εκθετική και generalized pareto κατανομή

Σχήμα 5.8: Μετρήσεις για τον 1ο χρήστη

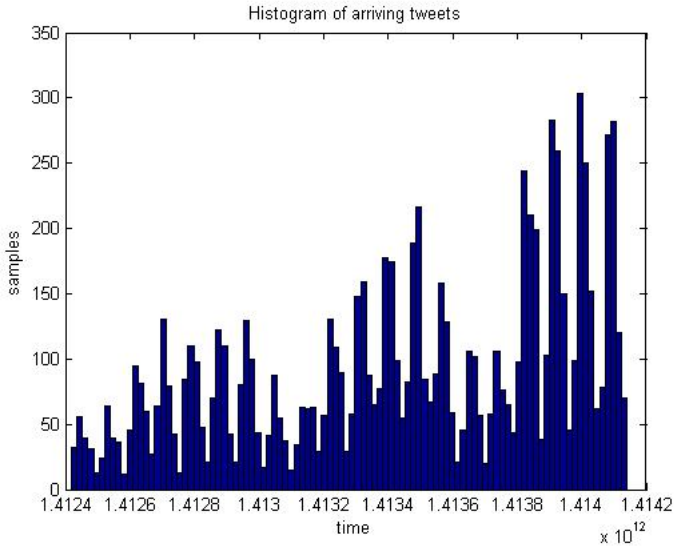


(α') Ιστόγραμμα των αφιχθέντων tweets

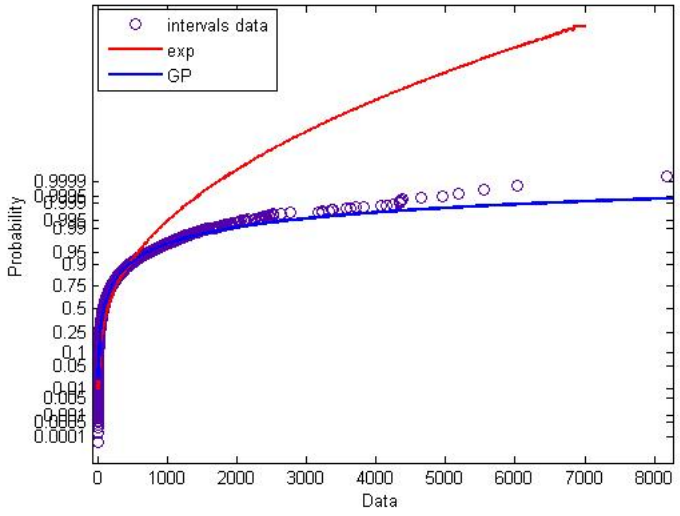


(β') Συνάρτηση Πιθανότητας και σύγκριση με εκθετική και generalized pareto κατανομή

Σχήμα 5.9: Μετρήσεις για τον 2ο χρήστη



(α') Ιστόγραμμα των αφιχθέντων tweets



(β') Συνάρτηση Πιθανότητας και σύγκριση με εκθετική και generalized pareto κατανομή

Σχήμα 5.10: Μετρήσεις για τον 3ο χρήστη

6 Κεφάλαιο
Επίλογος

6.1 Σύνοψη

Η παρούσα διπλωματική εργασία αφορά τη μελέτη της διάδοσης της πληροφορίας σε κοινωνικά δίκτυα όπως το Twitter. Δημιουργήθηκε ένα μοντέλο για την περιγραφή της διάδοσης και τον υπολογισμό της έκτασης στην οποία μπορεί να διαδοθεί ένα θέμα. Ταυτόχρονα ασχολείται και με το ρυθμό με τον οποίο οι χρήστες του δικτύου ενημερώνονται για τα θέματα που προκύπτουν μέσα στο δίκτυο και καταλήγει σε μία κατανομή που περιγράφει τους χρόνους μεταξύ των νέων μηνυμάτων. Τέλος έγινε επαλήθευση του μοντέλου που είχε σχηματιστεί μέσω δεδομένων που συλλέχθηκαν από την υπηρεσία του Twitter. Για την συλλογή των δεδομένων χρησιμοποιήθηκε η γλώσσα προγραμματισμού Java, καθώς και μια έτοιμη βιβλιοθήκη που παρέχεται, για να είναι δυνατή η χρήση της υπηρεσίας του δικτύου, πιο συγκεκριμένα η twitter4j. Για την προσαρμογή του μοντέλου και την επαλήθευσή του, καθώς και για τον υπολογισμό της κατανομής που ακολουθούν τα ενδιάμεσα διαστήματα μεταξύ των εισερχομένων tweets σε κάθε χρήστη, χρησιμοποιήθηκαν συναρτήσεις της Matlab.

6.2 Μελλοντική Εργασία

Τα αποτελέσματα της διπλωματικής εργασίας αποτελούν μια διαφορετική προσέγγιση σε σχέση με υπάρχουσες εργασίες. Καθώς μελετάται η διάδοση με βάση ένα σύνολο θεμάτων, θα μπορούσε μελλοντικά να μελετηθεί και η ροή της εισερχόμενης πληροφορίας και η κατανομή που ακολουθεί με βάση ένα γενικότερο θέμα και όχι συγκεκριμένα μηνύματα, όπως έγινε σε αυτή την εργασία. Ταυτόχρονα τα αποτελέσματα για τη δυναμική της διάδοσης της πληροφορίας στο δίκτυο μπορούν να χρησιμοποιηθούν πρακτικά από διαφημιστικές και εμπορικές εταιρίες που θέλουν να χρησιμοποιήσουν το δίκτυο ως μέσω προώθησης της επιχείρησης και διαφήμισης του έργου τους. Μια επέκταση που θα μπορούσε να βελτιώσει τα αποτελέσματα του μοντέλου

είναι να γίνει η παραπάνω διαδικασία για πολύ μεγαλύτερο χρονικό διάστημα και σε πιο μεγάλη κλίμακα χρηστών, ώστε να μπορεί να μελετηθεί πιο σφαιρικά χωρίς έντονες διακυμάνσεις και καλύπτοντας την πλήρη έκταση του φαινομένου που θα μελετηθεί.

Αναφορές

[1] Béla Bollobás, Oliver Riordan, Joel Spencer, and Gábor Tusnády. "The degree sequence of a scale-free random graph process." Random Structures and Algorithms. Vol. 18, Pages 279-290, 2001.

[2] H Trottier and P Philippe . "Deterministic Modeling Of Infectious Diseases: Theory And Methods". *The Internet Journal of Infectious Diseases*, Vol. 1, 2000.

[3] Herbert W. Hethcote. "The Mathematics of Infectious Diseases". *SIAM Review*, Vol. 42 Issue 4, Pages 599-653, Dec. 2000 .

[4] https://about.twitter.com/company.

[5] https://dev.twitter.com/overview/api.

[6] https://github.com/twitter/hbc. " A Java HTTP client for consuming Twitter's Streaming Api".

[7] https://twitter.com.

[8] https://www.facebook.com.

[9] https://www.linkedin.com.

[10] http://www.beevolve.com/twitter statistics.

[11] http://www.beevolve.com/twitter statistics. " An Exhaustive Study of Twitter Users Across the World.

[12] Jon Kleinberg and Eva Tardos. *"Algorithm Design"*. Addison-Wesley Longman Publishing Co., 2005.

[13] Kayastha N., Niyato D., Ping Wang, and Hossain E.. " Applications, Architectures, and Protocol Design Issues for Mobile Social Networks: A Survey", Vol. 99, Issue 12, Pages 2130 - 2158 ,November 2011 .

[14] Keeling Matt. “the mathematics of diseases.” , Plus Magaazine: Living Mathematics. Mar. 2001. Fall 2008.

[15] Maria Deijfen. “ Epidemics on social network graphs”, January 2000.

[16] Miller McPherson, Lynn Smith-Lovin, and James M Cook . “Birds of a Feather: Homophily in Social Networks” , Vol. 27, Pages 415-444, August 2001 .

[17] Pei Li We, Li HuiWan, and Xin Zhang. “Modeling of Information Diffusion in Twitter-Like Social Networks under Information Overload”. *The Scientific World Journal, Vol. 2014, 8 Pages, 2014* .

[18] Reinhard Diestel. *“Graph Theory”*. Springer, 4th edition, , 2010.

[19] Saeed Abdullah and Xindong Wu . “An Epidemic Model for News Spreading on Twitter”, ICTAI, Pages 163-169, 2011 .

[20] Seth Myers, Chenguang Zhu, and Jure Leskovec. “Information Diffusion and External Influence in Networks”. *KDD ’12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Pages 33-41, 2012.

[21] Shin-Ming Cheng, Vasileios Karyotis, Pin-Yu Chen, Kwang-Cheng Chen, and Symeon Papavassiliou. “Diffusion Models for Information Dissemination Dynamics in Wireless Complex Communication Networks”. *Journal of Complex Systems, Vol. 2013, Pages 13* , 2013.

[22] Steven H. Strogatz . “Collective dynamics of ’small-world’ networks”. *Nature*, Vol. 393, Pages 440-442, 1998.

[23] Steven H. Strogatz . “Exploring Complex Networks”. *Nature*, Vol. 410, Pages 268-276, 2001.

[24] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *“Introduction to Algorithms”*. The MIT Press, 2009.

[25] V. Karyotis, E. Stai, and S. Papavassiliou. *"Evolutionary Dynamics of Complex Communications Networks". CRC Press* 1est edition , October 14 2013.

[26] VenkataSwamy Martha, Weizhong Zhao, and Xiaowei Xu. "A Study on Twitter User-Follower Network A network based analysis", Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[27] W. O. Kermack and A. G. McKendrick. "A Contribution to the Mathematical Theory of Epidemics" . *The Royal Society*, Vol. 115, Pages 700-722, 1927.

[28] Yini Wang, Sheng Wen, Yang Xiang, and Wanlei Zhou. " Modeling the Propagation of Worms in Networks: A Survey", Communications Surveys and Tutorials, IEEE Vol. 16, Pages 942 - 960, 2014 .