ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΜΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ
ΚΑΤΕΥΘΥΝΣΗ "ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΙΣΤΟΥ"

ΠΡΟΒΛΕΨΗ ΑΚΜΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΜΕΤΑΒΑΛΛΟΜΕΝΑ ΣΤΟ ΧΡΟΝΟ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ελένη Παπαδοπούλου

ΦΕΒΡΟΥΑΡΙΟΣ 2016

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΜΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ
ΚΑΤΕΥΘΥΝΣΗ "ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΙΣΤΟΥ"

ΠΡΟΒΛΕΨΗ ΑΚΜΩΝ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΜΕΤΑΒΑΛΛΟΜΕΝΑ ΣΤΟ ΧΡΟΝΟ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

Ελένη Παπαδοπούλου

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26η Φεβρουαρίου 2016.

Επιβλέπων : Τσερπές Κωνσταντίνος,
Λέκτορας

.....	.....	.....
Τσερπές Κωνσταντίνος	Βαμβακάρη Μαλβίνα	Βαρλάμης Ηρακλής
Λέκτορας	Αναπληρώτρια Καθηγήτρια	Επίκουρος Καθηγητής

ΦΕΒΡΟΥΑΡΙΟΣ 2016

Περίληψη

Τα κοινωνικά δίκτυα χαρακτηρίζονται από γρήγορες αλλαγές στην τοπολογία και γενικότερα στη δυναμική αλληλεπίδραση μεταξύ των χρηστών. Η μελέτη της χρονικής εξέλιξης του γράφου ενός κοινωνικού δικτύου μπορεί να δώσει σημαντικές πληροφορίες για μελλοντικές προβλέψεις των σχέσεων μεταξύ χρηστών. Σκοπός της παρούσας μεταπτυχιακής διατριβής είναι η αναγνώριση παραμέτρων που συνεισφέρουν στη δημιουργία ακμών επιρροής και η μετέπειτα αξιοποίησή τους για την παραγωγή ενός μοντέλου πρόβλεψης ακμών. Οι βαθμοί εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών αποτελούν τις βασικές παραμέτρους με τις οποίες μπορούμε να χαρακτηρίσουμε ένα γράφο και να ερευνήσουμε την περίπτωση που συμβάλλουν στην πρόβλεψη ακμών. Κατά τη διαδικασία της έρευνας, που χωρίστηκε σε δύο μέρη, θεωρήσαμε ότι κάθε ακμή έχει και κάποια δευτερόλεπτα που παραμένει ενεργή μετά την εμφάνισή της, δηλαδή έχει μια διάρκεια ζωής. Στο πρώτο μέρος παρατηρήσαμε τους βαθμούς που επιλέξαμε με βάση επόμενες χρονικές στιγμές, αμέσως μετά την εμφάνιση κάθε ακμής, της τάξης μερικών δευτερολέπτων. Στο δεύτερο μέρος των πειραμάτων ερευνήσαμε την περίπτωση που προηγούμενες χρονικές στιγμές, πριν την εμφάνιση κάθε ακμής, συνεισφέρουν στην πρόβλεψη ακμών. Τα αποτελέσματα της κάθε ξεχωριστής διαδικασίας έδειξαν ότι η θεώρηση της διάρκειας ζωής των ακμών δεν αποτελεί έναν αποδοτικό τρόπο παρακολούθησης της εξέλιξης του γράφου. Παράλληλα, οι παράμετροι που χρησιμοποιήθηκαν δε συμβάλλουν στην πρόβλεψη ακμών με μεγάλη ακρίβεια, Οι βαθμοί εισερχόμενων και εξερχόμενων ακμών μπορούν σε πολύ λίγες περιπτώσεις και με μικρή ακρίβεια να προβλέψουν την ύπαρξη ή μη ακμής στο μέλλον. Αντίθετα, ο βαθμός των αμφίδρομων ακμών ενός κόμβου με σιγουριά δεν καταφέρνει να προβλέψει την ύπαρξη ακμής με αξιοπιστία αλλά τυχαία.

Λέξεις κλειδιά : κοινωνικά δίκτυα, πρόβλεψη ακμών, βαθμοί κόμβων, εξόρυξη δεδομένων, διάρκεια ζωής ακμών

Abstract

Social networks are characterized by rapid changes in the topology and, more generally, by changes in the dynamic interactions between users. Studying the temporal evolution of a social network graph can provide important information about predicting future relationships between different users. This thesis aims in identifying those parameters that contribute to the creation of influential linkages and their subsequent utilization in the production of a model for predicting edges. During the investigation that was divided in two parts we considered that each edge remains active for some seconds after its appearance; in other words it has a lifetime. In the first part, we observed the chosen degrees of nodes based on the following few seconds after an edge has made its appearance. In the second part of the experiment we investigated the case if earlier seconds, before the appearance of an edge, contribute to the edge prediction. The results of each individual process show that the consideration of the lifetime of an edge is not an efficient way of monitoring the evolution of a graph. Furthermore, the degrees that were used do not contribute to the prediction of future edges with great precision. Indegree and outdegree may predict, in very few cases and with low accuracy, the absence of an edge in the future. In contrast, the bidegree of nodes does not succeed in the prediction of an edge in any reliable way, but rather randomly.

Keywords : social networks, edge prediction, node degrees, data mining, edge lifetime

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

1 Εισαγωγή

1.1 Η επιρροή στα κοινωνικά δίκτυα

Η εκρηκτική ανάπτυξη των κοινωνικών μέσων παρέχει τη δυνατότητα σε εκατομμύρια ανθρώπους να παράξουν και να κοινοποιήσουν περιεχόμενο σε βαθμό που τα παλαιότερα χρόνια φάνταζε αδύνατο (D.M. Romero et al, 2011). Ως απόρροια αυτού του γεγονότος είναι η δημιουργία ποικίλων πλατφορμών υποστήριξης κοινωνικών δικτύων, από την άμεση ανταλλαγή μηνυμάτων (π.χ. Skype[1]), τα ιστολόγια (π.χ. WordPress[2], Blogger[3]) και τα μικρο-ιστολόγια (π.χ. Twitter[4]) μέχρι τις ιστοσελίδες κοινοποίησης περιεχομένου (π.χ. Flickr[5] , Youtube[6]) και τις κοινωνικές σελίδες δικτύωσης (π.χ. Facebook[7]). Τα δίκτυα αυτά μπορούν να αναπαρασταθούν ως γράφοι, όπου οι κόμβοι αντιπροσωπεύουν ανθρώπους ή οντότητες που μετέχουν στη κοινωνική αλληλεπίδραση και οι ακμές μεταξύ τους δηλώνουν την ύπαρξη μιας αλληλεπίδρασης (T. Tylenda et al, 2009). Οι χρήστες, μέσω των κοινωνικών δικτύων, έχουν τη δυνατότητα να δημιουργήσουν το δικό τους προφίλ, να αποφασίσουν με ποιους επιθυμούν να συνδεθούν και να εξερευνήσουν τον γράφο που προκύπτει.(D.M. Boyd, N.B. Ellison, 2007).

Τα κοινωνικά δίκτυα εκτός από τα αποδοτικά εργαλεία που παρέχουν για τη σύνδεση των χρηστών επιτρέπουν στις πληροφορίες και στις ιδέες να επηρεάσουν ένα μεγάλο κομμάτι του πληθυσμού για ένα μικρό χρονικό διάστημα (Chen,Wang & Yang, 2009). Είναι ευρέως γνωστό ότι η επιρροή είναι μια πολύπλοκη και διακριτική δύναμη που διέπει τη διαμόρφωση συμπεριφορών και σχέσεων στα κοινωνικά δίκτυα (Liu, Tang & Han, 2012). Ενδιαφέρουσα είναι η μελέτη των Fowler και Christakis(2008) σχετικά με την εξάπλωση της ευτυχίας σε μεγάλα κοινωνικά δίκτυα που αποδεικνύει ότι η ευτυχία ενός χρήστη μπορεί να επηρεαστεί σε μεγάλο βαθμό από τη θέση του στο δίκτυο -κεντρικότητα χρήστη- καθώς και από την ευτυχία των φίλων των φίλων του. Σε σελίδες όπως το Facebook και το Twitter, οι χρήστες

[1] http://www.skype.com/en/
[2] https://wordpress.org/
[3] https://www.blogger.com
[4] https://twitter.com/
[5] https://www.flickr.com
[6] https://www.youtube.com
[7] https://www.facebook.com/

είναι πολύ πιθανό να ακολουθήσουν φίλους που ασκούν επιρροή στον κοινωνικό τους κύκλο, να προωθήσουν ένα μικρο-ιστολόγιο ή να δηλώσουν ότι τους αρέσει μια φωτογραφία (Liu, Tang & Han, 2012). Είναι σχεδόν βέβαιο ότι η κοινωνική επιρροή εξελίσσεται σε μια διαδεδομένη, περίπλοκη και διακριτική δύναμη που ρυθμίζει τη δυναμική όλων των κοινωνικών δικτύων, τονίζοντας την αναγκαιότητα ανάλυσης και ποσοτικοποίησής της (Tang et al., 2009).

Στο σημείο αυτό, είναι σημαντικό να ορίσουμε τους μηχανισμούς που υποδεικνύουν την ύπαρξη επιρροής μεταξύ των χρηστών στις ευρέως χρησιμοποιούμενες πλατφόρμες κοινωνικής δικτύωσης του Facebook και του Twitter. Στο Facebook, ο εκάστοτε χρήστης μπορεί να αναρτήσει οποιοδήποτε περιεχόμενο επιθυμεί (φωτογραφίες, σχόλια, βίντεο κ.α.) και να το μοιραστεί με τους φίλους του. Οι μηχανισμοί "αρέσκειας" και κοινοποίησης μιας ανάρτησης δίνει τη δυνατότητα αναπαραγωγής του περιεχομένου από τους φίλους του και μέσω μιας "αλυσιδωτής" κοινοποίησης να γίνει ορατό σε χρήστες που δεν είναι άμεσα συνδεδεμένοι με αυτόν, επηρεάζοντας έτσι από μια μικρή μέχρι μια παγκόσμια κοινότητα χρηστών. Αντίστοιχα, στο Twitter οι χρήστες μπορούν μέσω ενός κειμένου (tweet) 140 χαρακτήρων να δηλώσουν την άποψή τους για οποιοδήποτε θέμα τους απασχολεί, μέσω μηχανισμών χρήσιμων για τη διάδοση ενδιαφέρουσας πληροφορίας εντός της κοινότητας του Twitter. Κάθε ανάρτηση μπορεί να αναπαραχθεί από άλλους χρήστες (retweet), να απαντηθεί (reply), ακόμη και να προστεθεί προσωπική αναφορά του ονόματος οποιουδήποτε χρήστη (mention), προσφέροντας έναν άμεσο δίαυλο επικοινωνίας μεταξύ χρηστών, χωρίς να συνδέονται απαραίτητα με κάποιο δεσμό ακολουθίας-φιλίας.

1.2 Κοινωνικά Δίκτυα : ο παράγοντας χρόνος

Τα κοινωνικά δίκτυα χαρακτηρίζονται από γρήγορες αλλαγές στην τοπολογία και γενικότερα στη δυναμική αλληλεπίδραση μεταξύ των χρηστών. Καθημερινά τα κοινωνικά δίκτυα συλλέγουν ένα μεγάλο όγκο δεδομένων που παράγεται μέσω της χρήσης τους. Ένα αξιοσημείωτο χαρακτηριστικό αυτών των δραστήριων δικτύων είναι ότι η δομή τους μεταβάλλεται στο χρόνο, όσο οι χρήστες για παράδειγμα επικοινωνούν με διαφορετικούς

φίλους (R. Rossi, B. Gallagher, 2013). Για παράδειγμα, στο Twitter πραγματοποιούνται 6.000 αναρτήσεις ανά δευτερόλεπτο, ενώ σε ώρες αιχμής ο αριθμός αυτός μπορεί να φτάσει τις 143.200 αναρτήσεις στον ίδιο χρόνο. Παράλληλα, στο Facebook, έχουν παρατηρηθεί περίπου 41.000 αναρτήσεις ανά δευτερόλεπτο ή περίπου 2.4Mb δεδομένων ανά δευτερόλεπτο. Καταλαβαίνουμε, λοιπόν, ότι η επεξεργασία ενός τέτοιου μεγέθους γρήγορης ροής δεδομένων που εξάγει ενδιαφέρουσα γνώση σε πραγματικό χρόνο αποτελεί μεγάλη πρόκληση (Wickramaarachchi et al., 2015). Η μελέτη της χρονικής εξέλιξης ενός γράφου μπορεί να δώσει σημαντικές πληροφορίες για μελλοντικές προβλέψεις των σχέσεων μεταξύ χρηστών. Η δυναμική του χρόνου είναι το κλειδί για την κατανόηση της συμπεριφοράς του συστήματος και αποτελεί παράγοντα ζωτικής σημασίας για τη μοντελοποίηση και την πρόβλεψη των αλλαγών στην πάροδο του χρόνου (R. Rossi, B. Gallagher, 2013).

1.3 Περιγραφή Προβλήματος

Σκοπός της παρούσας μεταπτυχιακής διατριβής είναι η αναγνώριση παραμέτρων που συνεισφέρουν στη δημιουργία ακμών επιρροής και η μετέπειτα αξιοποίησή τους για την παραγωγή ενός μοντέλου πρόβλεψης ακμών. Η διαδικασία περιλαμβάνει την εύρεση των μετρικών που με βάση το πρόσφατο ιστορικό συνδεσμολογίας και ανάλογα με την επιρροή που ασκούν οι χρήστες μεταξύ τους μπορούν να αποτελέσουν σημαντικό παράγοντα δημιουργίας νέων ή υπάρχουσων ακμών στο μέλλον. Η προσέγγισή μας περιλαμβάνει την παρατήρηση της επιρροής που ασκείται μεταξύ των διαφόρων κόμβων του δικτύου με βάση προηγούμενες αλλά και επόμενες χρονικές στιγμές, μετά την εμφάνιση μιας ακμής. Μετά από έναν αριθμό πειραμάτων εξόρυξης γνώσης μπορούμε να εξάγουμε ένα συμπέρασμα σχετικά με την ικανότητα των επιλεγμένων παραμέτρων να προβλέπουν την ύπαρξη ή μη ακμών μεταξύ των διαφορετικών κόμβων του εξελισσόμενου στο χρόνο γράφου.

1.4 Δομή Μεταπτυχιακής Διατριβής

Στα κεφάλαια που ακολουθούν θα γίνει αναφορά στις διαφορετικές οπτικές προσέγγισης παρόμοιων προβλημάτων, καθώς και η πραγματοποίηση και η συμβολή της παρούσας

έρευνας. Πιο συγκεκριμένα, στο κεφάλαιο 2 θα παρουσιαστούν εκτενώς παλαιότερες έρευνες που σχετίζονται με την πρόβλεψη ακμών σε δυναμικούς γράφους και πως η καθεμία έχει συμβάλλει στην επίλυση του γενικότερου προβλήματος. Στο κεφάλαιο 3, θα γίνει περιγραφή της μεθοδολογίας στην οποία βασίστηκε η τρέχουσα έρευνα και τα πειράματα που πραγματοποιήθηκαν, ενώ στο κεφάλαιο 4 θα παρουσιαστούν τα αποτελέσματα αυτής της υλοποίησης. Τέλος, το κεφάλαιο 5 περιλαμβάνει τα συμπεράσματα και τις μελλοντικές επεκτάσεις που μπορούν να γίνουν για τη βελτίωση της προσέγγισης του όλου ζητήματος.

2. Ανασκόπηση Βιβλιογραφίας

2.1 Εισαγωγή

Η πρόβλεψη ακμών σε κοινωνικούς γράφους που εξελίσσονται στο χρόνο αποτελεί ένα ζήτημα για το οποίο πολλές έρευνες έχουν πραγματοποιηθεί. Κάθε προσέγγιση διαφέρει ανάλογα με τη θεώρηση της ακμής ή του κόμβου ως κεντρικό στοιχείο της έρευνας, το είδος του δικτύου που μελετά (ομογενές ή ετερογενές), καθώς και τον τρόπο διαχείρισης του γράφου στο χρόνο. Οι μελέτες αυτές αποτέλεσαν το πρώτο βήμα για να κατανοήσουμε τις ήδη υπάρχουσες έρευνες σχετικά με την πρόβλεψη ακμών σε χρονικά εξελισσόμενους γράφους, αλλά και για να προσδιορίσουμε τη δική μας προσέγγιση που θα ακολουθήσουμε στα πλαίσια της παρούσας διπλωματικής διατριβής.

2.2 Ακμές και κόμβοι ως κεντρικά στοιχεία της έρευνας

Ανάλογα με την περίπτωση που η ακμή ή ο κόμβος λαμβάνεται ως κεντρικό στοιχείο της έρευνας η προσέγγιση μπορεί να θεωρηθεί ακμοκεντρική (edge-centric) ή κομβοκεντρική (node-centric). Στην πρώτη περίπτωση, σκοπός της μελέτης είναι η πρόβλεψη ακμής μεταξύ

δύο κόμβων i και j, ανεξάρτητα από τις συνδέσεις των i και j με άλλους κόμβους. Το μειονέκτημα αυτής της θεώρησης είναι ότι τελικά στοχεύει σε νέες ακμές χωρίς να λαμβάνει υπόψη την υπόλοιπη συνδεσμολογία του δικτύου, καθώς και τις αποστάσεις μεταξύ των κόμβων. Αντίθετα στην κομβοκεντρική προσέγγιση η μελέτη επικεντρώνεται στους κόμβους, στην απόσταση που υπάρχει μεταξύ τους, καθώς και τις ομοιότητες που παρουσιάζουν (Leskovec et al. (2008), Leskovec (2008), Newman M. (2003)).

Κομβοκεντρική προσέγγιση ακολουθούν και οι Tylenda et. al (2009) που για κάθε κόμβο v υπολογίζουν τους γειτονικούς κόμβους N(v) και στοχεύουν στην πρόβλεψη νέων ή ήδη υπάρχουσων ακμών. Η δόμηση της γειτονιάς N(v) πραγματοποιείται με κόμβους που δεν είναι απαραίτητα άμεσα συνδεδεμένοι με τον κόμβο v, αλλά απέχουν ελάχιστη ή απόσταση ίση με δύο από τον κεντρικό κόμβο. Παράλληλα, οι Rossi και Gallagher (2013) δίνουν έμφαση στη συμπεριφορά των κόμβων που εκφράζεται μέσω διαφόρων χαρακτηριστικών, όπως οι βαθμοί τους (εισερχόμενοι/εξερχόμενοι κόμβοι, με βάρη κλπ) και δημιουργούν πρότυπα -που ονομάζουν ρόλους- για την παρατήρησή τους στο χρόνο. Στόχος αυτής της έρευνας είναι η πρόβλεψη της δομής του δικτύου με την πάροδο του χρόνου, καθώς και η πρόβλεψη αλλαγής του ρόλου ενός κόμβου (για παράδειγμα εάν ο ένας κόμβος με υψηλό βαθμό εισερχόμενων ακμών μετατρέπεται σε κόμβο με υψηλό βαθμό ενδιάμεσων ακμών).

2.3 Είδη Δικτύων : Ομογενή και Ετερογενή

Ένα δίκτυο ανάλογα με το διαφορετικό είδος οντοτήτων από τις οποίες αποτελείται μπορεί να θεωρηθεί ομογενές ή ετερογενές. Ένα ομοιογενές δίκτυο αποτελείται από οντότητες ίδιου είδους, όπως για παράδειγμα τα δίκτυα φιλίας αλλά και τα δίκτυα που σχετίζονται με τη συνεργασία μεταξύ συγγραφέων (co-author). Στην πραγματικότητα όμως τα περισσότερα δίκτυα είναι ετερογενή, δηλαδή αποτελούνται από περισσότερα από ένα διαφορετικά είδη οντοτήτων και ακμών. Χαρακτηριστικό παράδειγμα αποτελεί ένα δίκτυο με ταινίες, που περιλαμβάνει πληροφορίες σχετικά με τα είδη των ταινιών, των ηθοποιών, τους χρήστες και τα σχόλια, με τις ακμές που τα συνδέουν να εκφράζουν διαφορετικές ενέργειες.

Μια ενδιαφέρουσα και αξιόλογη προσέγγιση πρόβλεψης ακμών σε ετερογενή δίκτυα πραγματοποίησαν οι Sun et al (2012). Τα δεδομένα που χρησιμοποίησαν προήλθαν από το DBLP[8] βιβλιογραφικό δίκτυο, με τη συμμετοχή 4 διαφορετικού τύπου οντοτήτων (συγγραφείς, άρθρα, όροι και συνέδριο) με διαφορετικές σχέσεις να τις συνδέουν. Η ετερογένεια των κόμβων δημιουργεί δυσκολίες στην υιοθέτηση και χρήση όλων των ήδη γνωστών τοπολογικών προσεγγίσεων (Hasan et al. (2006), Leroy et al. (2010), Liben-Nowell et al. (2003), Lichtenwalter et al. (2010), Wang et al. (2007)) που αφορούν τα ομογενή δίκτυα. Οι συγγραφείς επικεντρώθηκαν όχι μόνο στον εντοπισμό της περίπτωσης που θα εμφανισθεί μια ακμή, αλλά και στον υπολογισμό του ακριβή χρόνου που μια ακμή θα δημιουργηθεί βασιζόμενοι στην τοπολογία του δικτύου. Με τον προσδιορισμό γενικευμένων σχέσεων και ακολουθώντας μετα-μονοπάτια (meta-paths), στοιχεία που προκύπτουν μέσα από το δίκτυο, οι συγγραφείς μπορούν να προβλέψουν το χρόνο που χρειάζεται να δημιουργηθεί μια σχέση-ακμή μεταξύ δύο οντοτήτων. Η πρόβλεψη ακμής δηλαδή μετατράπηκε σε πρόβλεψη σχέσης μεταξύ των διαφορετικών αντικειμένων του δικτύου.

Την ίδια περίοδο οι Yang et al.(2012) ερευνώντας την πρόβλεψη ακμών σε ετερογενή δίκτυα πρότειναν ένα νέο τοπολογικό χαρακτηριστικό που μπορεί να αναγνωρίσει τη συσχέτιση μεταξύ των διαφόρων τύπων συνδέσεων του δικτύου για το ζήτημα της πρόβλεψης ακμών. Υιοθετώντας παλαιότερες και ήδη γνώριμες τεχνικές (Adamic and Adar (2001), Mitzenmacher (2001)) προσθέτουν επιπλέον χαρακτηριστικά που σχετίζονται με το χρόνο για τη βελτίωση της απόδοσης. Σύμφωνα με πειράματα σε πραγματικά σύνολα δεδομένων, η προσέγγισή αυτή αποτελεί μια αποτελεσματική μέθοδο σε σχέση με άλλες πρόσφατες δημοσιευμένες λύσεις.

2.4 Διαχείριση γραφου σε χρονικά παράθυρα

Κάθε χρονική στιγμή για ένα γράφο ενδέχεται να κρύβει σημαντικές πληροφορίες για τη μελλοντική συνδεσμολογία του δικτύου, όπως τη χρονική στιγμή δημιουργίας ή διαγραφής μια ακμής ή ενός κόμβου. Αρκετές έρευνες έχουν ασχοληθεί με το ζήτημα της πρόβλεψης

[8] http://dblp.uni-trier.de/

ακμών σε δίκτυο που εξελίσσεται στο χρόνο που όμως διαχειρίζονται το γράφο ως ένα στατικό στιγμιότυπο και όχι παρατηρώντας κάθε χρονική στιγμή ξεχωριστά. Σύμφωνα με τον Vu et al.(2011) τα δεδομένα είναι προτιμότερο να παρατηρούνται ως μια ακολουθία από στιγμιότυπα ενός εξελισσόμενου γράφου ή ως μια συνεχόμενη χρονική διαδικασία.

Τη λογική αυτή ακολούθησαν οι Sarkar et al. (2011), οι οποίοι παρήγαγαν ένα μη παραμετροποιήσιμο μοντέλο όπου ένας γράφος εξάγεται για κάθε διαφορετική χρονική στιγμή. Η συμπεριφορά στις συνδέσεις κάθε κόμβου i είναι ανεξάρτητη από τον υπόλοιπο γράφο, που σημαίνει ότι οι εξερχόμενες ακμές του i τη χρονική στιγμή t μπορούν να μοντελοποιηθούν ως μια συνάρτηση της "τοπικής" γειτονιάς του επί ενός κινούμενου παραθύρου. Επίσης, η αποτελεσματικότητα αυτής της προσέγγισης έγγειται στο γεγονός ότι ενσωματώνει χαρακτηριστικά τοπολογίας που δεν προέρχονται από το δίκτυο (όπως ετικέτες), ενώ παράλληλα υποστηρίζει την ύπαρξη διαφορετικών ειδών γειτονιών σε ένα γράφο με διαφορετικές δυναμικές το κάθε ένα.

Παράλληλα, οι Rossi και Gallagher (2013), επέλεξαν να μελετήσουν μεγάλους εξελισσόμενους γράφους με βάση τη συμπεριφορά του συστήματος, αναπτύσσοντας μοντέλα για την αποκάλυψη πρότυπων συμπεριφοράς των κόμβων ενός γράφου και πως αυτά αλλάζουν στο χρόνο. Οι συγγραφείς παρέχουν ένα βελτιωμένο μοντέλο σε σχέση με παλαιότερες μεθόδους (Fu et al. (2009), Xing et al. (2010)) κατά το οποίο για κάθε χρονικό παράθυρο-γράφο εξάγονται αντιπροσωπευτικά χαρακτηριστικά (όπως είπαμε εισερχόμενοι/εξερχόμενοι κόμβοι, με βάρη κλπ) και στη συνέχεια ανακαλύπτονται συμπεριφορικοί ρόλοι επαναληπτικά από τη χρονική ακολουθία των στιγμιοτύπων του δικτύου. Τελικά, αναπτύσσεται ένα μοντέλο που μπορεί να προβλέψει πως οι συμπεριφορές αυτές εξελίσσονται στο χρόνο, τις μελλοντικές δομικές αλλαγές που ενδέχεται να προκύψουν και τις ασυνήθιστες μεταβάσεις χρονικών συμπεριφορών των κόμβων του δικτύου.

3. Μεθοδολογία

Στο παρόν κεφάλαιο, θα παρουσιάσουμε τις παραμέτρους που επιλέχθηκαν για την εξέταση της συμβολής τους στην πρόβλεψη ακμών, καθώς και τον τρόπο υπολογισμού τους. Στη συνέχεια, θα αναφέρουμε τη θεώρησή μας σχετικά με το χρόνο ζωής της κάθε ακμής, ενώ στο τελευταίο μέρος του κεφαλαίου θα γίνει εκτενής περιγραφή των πειραμάτων που πραγματοποιήθηκαν για την πρόβλεψη ακμών σε γράφους που μεταβάλλονται στο χρόνο με βάση τις παραμέτρους και το χρόνο ζωής των ακμών που επιλέξαμε.

3.1 Δεδομένα

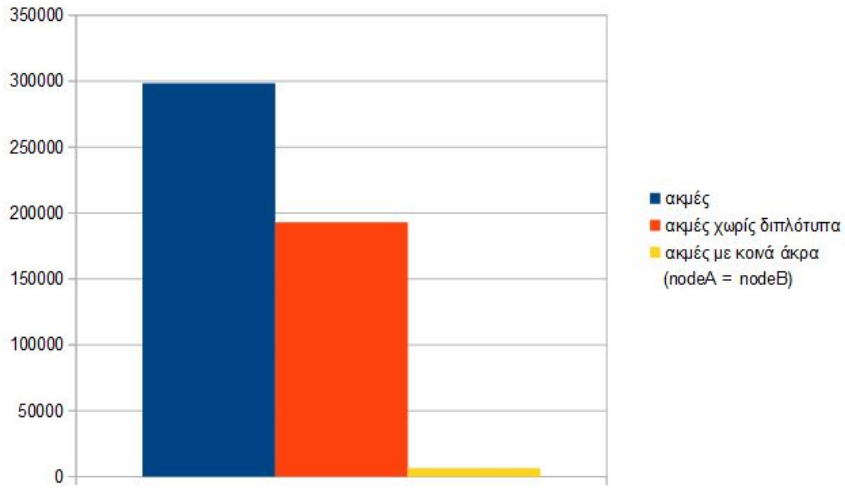
Τα δεδομένα[9] που χρησιμοποιήθηκαν προέρχονται από το κοινωνικό δίκτυο του Twitter και καταγράφουν αλληλεπιδράσεις κοινοποίησης αναρτήσεων (retweets) μεταξύ 148.918 διαφορετικών χρηστών σε διάστημα 12 ωρών. Κάθε εγγραφή αποτελείται από μια ακμή, δηλαδή ένα ζευγάρι κόμβων, και το δευτερόλεπτο που αυτή δημιουργήθηκε (πίνακας 3.1). Κατά την παρατήρηση των δεδομένων διαπιστώσαμε αρχικά ότι υπάρχουν αρκετές εγγραφές που αναφέρονται για την ίδια κοινοποίηση την ίδια χρονική στιγμή (πχ. 1,2,6 και 80,20,6), ενώ υπάρχουν και εγγραφές με κοινοποίηση περιεχομένου από τον ίδιο χρήστη από τον οποίο παράχθηκε η ανάρτηση (πχ 82,82,7). Όπως φαίνεται στο σχήμα 3.1, αρκετές ήταν οι ακμές που εμφανίζονταν δύο ή περισσότερες φορές το ίδιο δευτερόλεπτο, ενώ λιγότερες ήταν οι ακμές που αποτελούνται από τα ίδια άκρα, στις περιπτώσεις δηλαδή όπου η ανάρτηση ενός χρήστη κοινοποιήθηκε από τον ίδιο (κίτρινη στήλη).

node A	node B	timestamp
78	79	2
94	95	3
1	**2**	**6**
1	**2**	**6**
3	4	6
80	**20**	**6**
80	**20**	**6**
81	20	6
5	6	7

[9] http://www.maths.manchester.ac.uk/~weijian/EvolvingGraphDatasets/#twitter

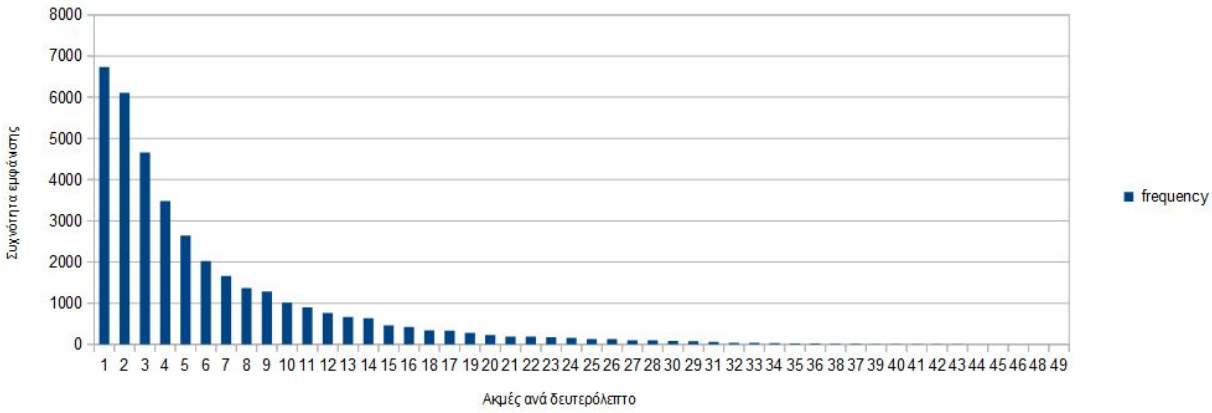
82	83	7
82	**82**	**7**

Πίνακας 3.1 Ο χρήστης Α κοινοποιεί περιεχόμενο του χρήστη Β.



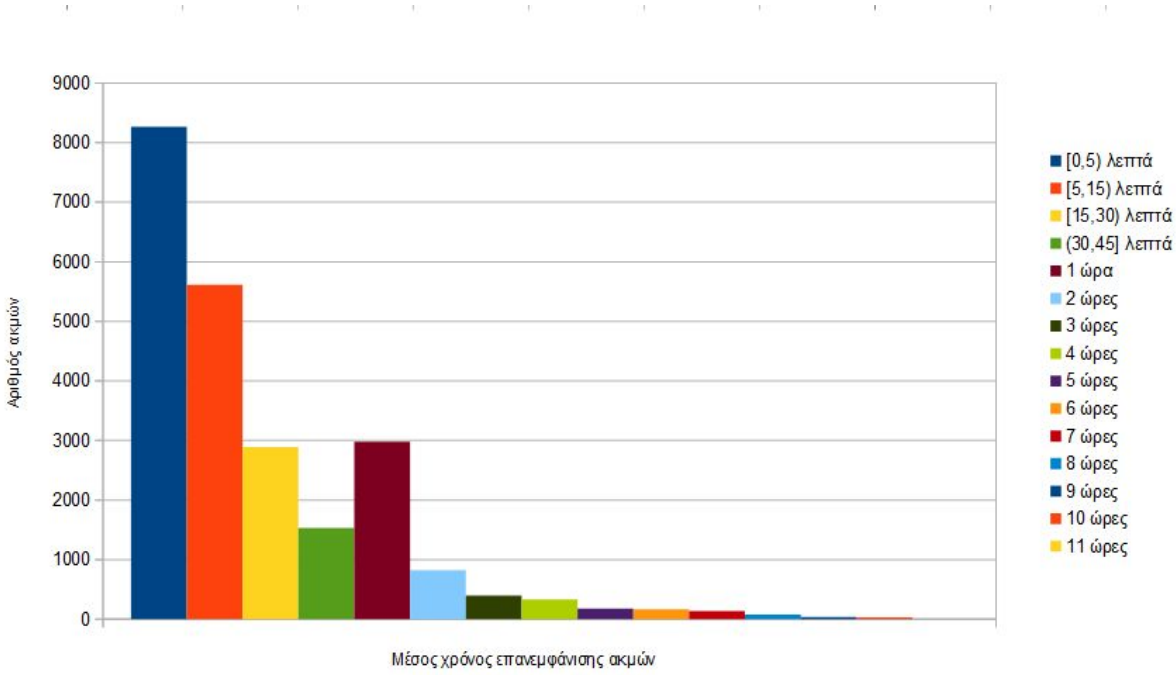
Σχήμα 3.1 Ανάλυση δεδομένων κατά αριθμό ακμών, ακμές χωρίς διπλότυπα και ακμές με ίδια άκρα

Παράλληλα, αναλύοντας τα δεδομένα με τεχνικές προγραμματισμού σε γλώσσα Java εντοπίσαμε ότι σε 6734 χρονικές στιγμές εμφανίζεται μόνο μία ακμή και όσο αυξάνονται οι ακμές ανά δευτερόλεπτο τόσο μειώνεται η συχνότητα εμφάνισής τους. Μάλιστα μόνο σε μία χρονική στιγμή έχει παρατηρηθεί η δημιουργία 49 ακμών που αποτελεί και τη μέγιστη στο διάστημα των 12 ωρών ενώ η συχνότητα εμφάνισης μόνο μίας ακμής σε ένα δευτετόλεπτο αγγίζει τις 6734 (σχήμα 3.2).



Σχήμα 3.2 Γραφική αναπαράσταση δεδομένων με τη συχνότητα που εμφανίζονται οι ακμές ανά δευτερόλεπτο.

Τέλος, στο διάστημα των 12 ωρών οι χρήστες παρουσίαζαν διαφορετική συχνότητα εμφάνισης κοινοποίησης αναρτήσεων (σχήμα 3.3). Πιο συγκεκριμένα, ακμές που δημιουργούνταν κάποια χρονική στιγμή t ενδέχεται να επανεμφανίζονταν κατά μέσο όρο μετά από μία μέχρι και 11 ώρες. Υψηλότερη δραστηριότητα επανεμφάνισης ακμών παρατηρείται κατά τη χρονική περίοδο ενός δευτερολέπτου μέχρι 5 λεπτών που μας υποδεικνύει ότι η παρατήρηση της δραστηριότητας των χρηστών στο χρονικό διάστημα αυτό μπορεί να δώσει κάποιο συμπέρασμα σχετικά με την επιρροή που ασκείται μεταξύ τους για το σύνολο των δεδομένων.



Σχήμα 3.3 Μέσος χρόνος επανεμφάνισης ακμών στο διάστημα 12 ωρών καταγραφής

3.2 Επιλογή Παραμέτρων

Όπως έχει ήδη αναφερθεί, ένα δίκτυο μπορεί να αναπαρασταθεί από έναν κατευθυνόμενο γράφο, ο οποίος αποτελείται από κόμβους, που αντιπροσωπεύουν οντότητες, και από κατευθυνόμενες ακμές, που αντιστοιχούν σε αλληλεπιδράσεις μεταξύ των διαφόρων οντοτήτων. Η κατεύθυνση κάθε ακμής αντιπροσωπεύει μια δραστηριότητα που πραγματοποιείται μεταξύ των κόμβων που μετέχουν σε αυτή, όπως για παράδειγμα μια

αναφορά (mention), μια απάντηση (reply), ένα δεσμό φιλίας-ακολουθίας (follow) ή την κοινοποίηση μιας ανάρτησης (retweet). Ένας τρόπος προσέγγισης ενός γράφου είναι να αγνοήσουμε τα πρότυπα που δημιουργούνται μεταξύ των διαφορετικών κόμβων του και να επικεντρωθούμε στην παρατήρηση κάθε κόμβου ξεχωριστά.

Ένας κόμβος αποτελεί έναν κόμβο προορισμό, εάν συνδέεται με κόμβους που δείχνουν σε αυτόν, ή/και έναν κόμβο αφετηρία, εάν ο κόμβος αυτός δείχνει σε άλλους κόμβους. Κάθε κόμβος μπορεί να χαρακτηριστεί από δύο μέτρα ή αλλιώς βαθμούς (degrees), δηλαδή τις εισερχόμενες και τις εξερχόμενες ακμές του. Ο εισερχόμενος βαθμός (indegree) ενός κόμβου i μπορεί να υπολογιστεί από το άθροισμα των εισερχόμενων ακμών του από οποιοδήποτε κόμβο j, ενώ ο εξερχόμενος βαθμός ενός κόμβου i (outdegree) υπολογίζεται από το άθροισμα των εξερχόμενων ακμών του προς οποιοδήποτε κόμβο j. Η μέτρηση των εισερχόμενων και εξερχόμενων ακμών ενός κόμβου ενδέχεται να εκφράζει και σε ένα βαθμό την επιρροή που ασκείται μεταξύ των κόμβων ενός γράφου καθώς αυτός εξελίσσεται στο χρόνο. Ένας κόμβος με υψηλό βαθμό εισερχόμενων ακμών μπορεί να αποτελεί έναν κόμβο με ενδιαφέρον περιεχόμενο και κατά συνέπεια να επηρεάζει ένα μεγάλο πλήθος χρηστών. Σε κάθε χρονική στιγμή t οι βαθμοί αυτοί μεταβάλλονται κατά αντιστοιχία με τη δομή του γράφου, όπως φαίνεται στα σχήματα 1 και 2.

$$\text{indegree}_i(t) = \sum_j a_{ji}(t)$$

Σχήμα 3.4. Υπολογισμός εισερχόμενου βαθμού κόμβου i τη χρονική στιγμή t

$$\text{outdegree}_i(t) = \sum_j a_{ij}(t)$$

Σχήμα 3.5. Υπολογισμός εξερχόμενου βαθμού κόμβου i τη χρονική στιγμή t

Εκτός από τα μέτρα που μόλις περιγράψαμε υπάρχει και η περίπτωση όπου ένας κόμβος συνδέεται ταυτόχρονα με εισερχόμενη και εξερχόμενη ακμή με έναν άλλον κόμβο, φαινόμενο που εμφανίζεται σε ένα μικρό αλλά υπαρκτό πλήθος των δεδομένων μας. Οι περιπτώσεις αυτές ενδέχεται να αποτελούν ένα σημαντικό παράγοντα επιρροής μεταξύ των διαφορετικών κόμβων του δικτύου. Λαμβάνοντας υπόψιν αυτή την παρατήρηση θεωρήσαμε έναν ακόμη παράγοντα ως παράμετρο στα πειράματα που πραγματοποιήσαμε, το βαθμό

αμφίδρομης ακμής (bidegree). Οι βαθμοί εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών κάθε κόμβου μιας ακμής αποτελούν τις τρεις τιμές με τις οποίες θα εισάγουμε στα πειράματά μας, η περιγραφή των οποίων θα ακολουθήσει προς το τέλος του κεφαλαίου.

$$\text{bidegree}_i(t) = \sum_j (x_{ij} \wedge y_{ji}),$$

$$x, y = \begin{cases} 1, & \text{υπάρχει ακμή} \\ 0, & \text{δεν υπάρχει} \end{cases}$$ όπου μεταξύ των κόμβων i και j και ^ λογικό AND

Σχήμα 3.6. Υπολογισμός αμφίδρομου βαθμού κόμβου i τη χρονική στιγμή t

3.3 Διάρκεια ζωής Ακμών

Καθώς τα δεδομένα μεταβάλλονται κάθε δευτερόλεπτο είναι φυσικό να αλλάζει και η συνδεσμολογία του δικτύου. Ακμές που δημιουργούνται σε ένα δευτερόλεπτο ενδέχεται να μην εμφανισθούν ξανά το επόμενο ή και ποτέ ξανά στο μέλλον. Κρίνεται, λοιπόν, αναγκαίος ο ορισμός της διάρκειας ζωής (lifetime) των ακμών, δηλαδή ο χρόνος που θα συνεχίσουμε να θεωρούμε την ακμή ενεργή μετά τη δημιουργία της. Όπως αναφέραμε και προηγουμένως, κατά τη χρονική περίοδο 1 δευτερολέπτου μέχρι 5 λεπτών εμφανίζεται υψηλή δραστηριότητα επανεμφάνισης ακμών. Για το σκοπό αυτό θεωρήσαμε τέσσερις διαφορετικές διάρκειες ζωής που αντιστοιχούν σε 3, 5, 10 και 272 δευτερόλεπτα ζωής, με τα 272 δευτερόλεπτα να αντιπροσωπεύουν το μέσο όρο όλων των χρόνων επανεμφάνισης όλων των ακμών. Πρακτικά αυτό σημαίνει ότι στην περίπτωση ζωής ίση με 3 δευτερόλεπτα, ο χρόνος που μια ακμή θα θεωρείται ενεργή είναι από την αρχή που θα εμφανισθεί μέχρι και 2 δευτερόλεπτα μετά. Μετά το πέρας της διάρκειας αυτής η ακμή θα θεωρείται ανενεργή οπότε και οι βαθμοί εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών των κόμβοι που μετέχουν σε αυτή θα ανανεώνονται σύμφωνα με τη δομή του δικτύου εκείνη τη χρονική στιγμή. Στον πίνακα 3.1 παραθέτουμε ένα παράδειγμα αυτής της θεώρησης για διάρκεια ζωής ίσης με 3 δευτερόλεπτα.

timestamp	Edges <nodeA,nodeB,timestamp>		node = 78		
			indegree	outdegree	bidegree
2		78,79,2	0	1	0
3	78,76,3	78,79,3	0	2	0
4	78,76,4	78,79,4	0	**2**	0
5	78,76,5		0	**1**	0
6			0	**0**	0

Πίνακας 3.2 Παράδειγμα υπολογισμού βαθμών για τον κόμβο 78 με διάρκεια ακμής 3 δευτερολέπτων.

3.4 Πειράματα

3.4.1 Πρόβλεψη ακμών με βάση τη διάρκεια ζωής τους

Στο πρώτο μέρος των πειραμάτων, επιλέξαμε να παρατηρήσουμε τα δεδομένα εξετάζοντας τις ακμές που προκύπτουν ανά δευτερόλεπτο, λαμβάνοντας υπόψη τη διάρκεια ζωής τους. Πιο συγκεκριμένα, η διαδικασία περιλαμβάνει την εξαγωγή των βαθμών εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών για κάθε κόμβο που μετέχει σε μια ακμή τη χρονική στιγμή t, με τη θεώρηση ότι η ακμή αυτή παραμένει ενεργή για t+2 χρονικές στιγμές (στην περίπτωση διάρκειας ζωής ίσης με 3 δευτερόλεπτα). Στη συνέχεια, εκπαιδεύσαμε το σύστημά μας με τις τιμές αυτές, χρησιμοποιώντας ήδη υπάρχοντες αλγορίθμους και καταλήξαμε σε συμπεράσματα που θα αναλυθούν στο επόμενο κεφάλαιο. Ακολουθεί η περιγραφή της εκτέλεσης του πρώτου πειράματος.

3.4.1.1 Διαδικασία εξαγωγής βαθμών κόμβων ανά ακμή

Το αρχείο με τα δεδομένα των 12 ωρών καταγραφής δραστηριοτήτων στο κοινωνικό δίκτυο του Twitter περιλαμβάνει ακμές που προκύπτουν σε κάθε δευτερόλεπτο. Δεδομένου ότι το

πρόβλημά μας είναι τελικά η πρόβλεψη ύπαρξης ή μη ακμής a_{ij} με βάση τις παραμέτρους που υπολογίζουμε για κάθε κόμβο i,j, χρειάζεται με κάποιο τρόπο όταν εκπαιδεύουμε το μοντέλο μας να ορίζουμε πότε η εγγραφή τη χρονική στιγμή t δηλώνει ύπαρξη ακμής και πότε όχι. Το πρόβλημα επιλύεται με την προσθήκη μιας δυαδικής παραμέτρου "isEgde" με την οποία μπορούμε να δηλώσουμε την ύπαρξη ή μη ακμής a_{ij} τη χρονική στιγμή t, με τιμές 1 και 0 αντίστοιχα.

Η διαδικασία εξαγωγής των βαθμών των κόμβων i,j μιας ακμής a_{ij} μπορεί να χωριστεί σε δύο μέρη. Αρχικά, υπολογίζουμε τους βαθμούς για κάθε κόμβο i,j μιας ακμής a_{ij} τη χρονική στιγμή που θα εμφανισθεί ("isEgde"=1) και στη συνέχεια υπολογίζουμε τους βαθμούς για κάθε κόμβο i,j μιας ακμής a_{ij} την αμέσως προηγούμενη χρονική στιγμή, ακριβώς πριν δημιουργηθεί ("isEgde"=0). Και στις δύο περιπτώσεις, ακολουθώντας τη χρονική εξέλιξη των δεδομένων, για κάθε ακμή που δημιουργείται τη χρονική στιγμή t θεωρούμε ότι παραμένει ενεργή για χρονική διάρκεια t+2, t+4, t+9 και t+271, ανάλογα με αυτή που έχουμε ορίσει κατά την εκτέλεση του προγράμματος. Κάθε εγγραφή που εξάγεται και δηλώνει την ύπαρξη ακμής είναι της μορφής <$node_i$, $node_j$, $indegree_i(t)$, $outdegree_i(t)$, $bidegree_i(t)$, $indegree_j(t)$, $outdegree_j(t)$, $bidegree_j(t)$, 1>, ενώ όταν δηλώνει τη μη ύπαρξη της ακμής που δημιουργείται τη χρονική στιγμή t είναι της μορφής <$node_i$, $node_j$, $indegree_i(t-1)$, $outdegree_i(t-1)$, $bidegree_i(t-1)$, $indegree_j(t-1)$, $outdegree_j(t-1)$, $bidegree_j(t-1)$, 0>

timestamp	**node A**	**node B**	**node A**	**node B**	**isEdge**
			indegree,outdegree,bidegree	**indegree,outdegree,bidegree**	
6	80	20	0,1,0	**1,0,0**	1
6	80	20	0,0,0	**0,0,0**	0
6	81	20	0,1,0	**2,0,0**	1
6	81	20	0,0,0	**0,0,0**	0
8	84	20	0,1,0	3,0,0	1
8	84	20	0,0,0	2,0,0	0

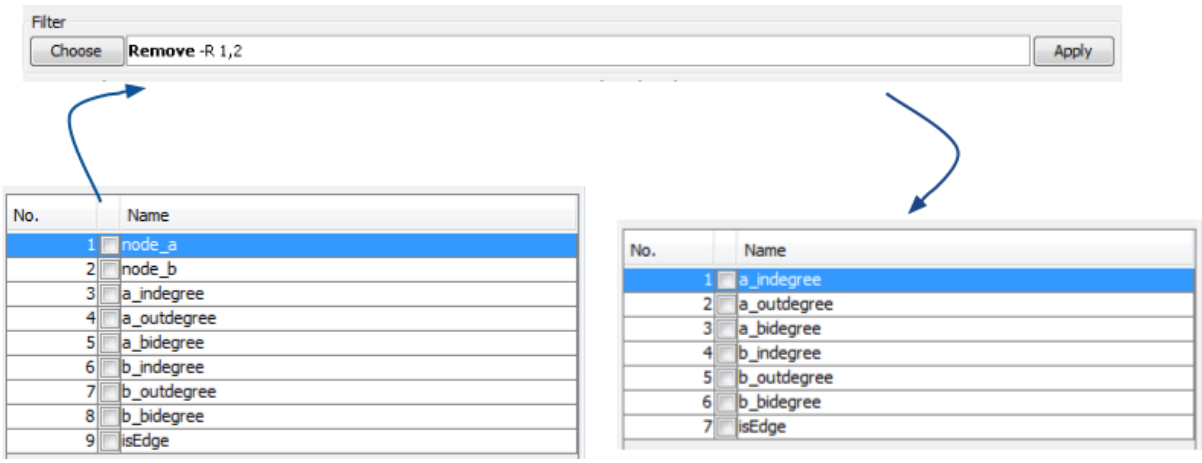
Πίνακας 3.3 Δεδομένα μετά τον υπολογισμό των εισερχόμενων, εξερχόμενων και αμφίδρομων βαθμών για κάθε κόμβο για διάρκεια ζωής ίση με 3 δευτερόλεπτα.

Στις περιπτώσεις όπου οι ακμές εμφανίζονται το ίδιο ακριβώς δευτερόλεπτο δεν τις λαμβάνουμε υπόψην καθώς θεωρούμε ότι θα προσθέσουν επιπλέον τιμές στις μετρήσεις μας οι οποίες θα είναι πλασματικές αφού δε μας ενδιαφέρει στο στάδιο αυτό το πλήθος των προβλέψεων αλλά η ύπαρξη ή μη ακμής σε γενικότερο επίπεδο. Δημιουργείται με τον τρόπο αυτό ένα αρχείο για κάθε διαφορετική διάρκεια ζωής, με ίσο αριθμό εγγραφών που περιγράφουν την ύπαρξη ακμής και τη μη-ύπαρξη ακμής.

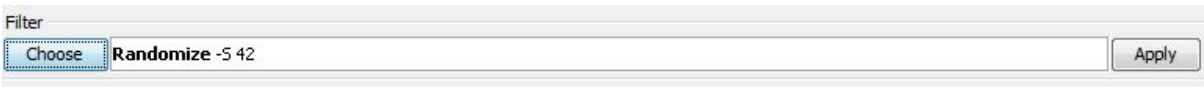
3.4.1.2 Εκπαίδευση και παραγωγή μοντέλου πρόβλεψης ακμών

Τα δεδομένα που εξήχθησαν με τον τρόπο που μόλις περιγράψαμε χρησιμοποιήθηκαν για την εκπαίδευση και την παραγωγή ενός μοντέλου πρόβλεψης ακμών με βάση τα έξι χαρακτηριστικά που ορίσαμε, τρία για κάθε κόμβο. Η διαδικασία που ακολουθήσαμε ήταν αρχικά να φορτώσουμε με τη σειρά τα 4 διαφορετικά αρχεία -ένα για κάθε όρισμα διάρκειας ζωής ίσης με 3,5,10 και 272 αντίστοιχα- στο εργαλείο Weka[10]. Στη συνέχεια αφαιρέσαμε τις τιμές των κόμβων που δεν επηρεάζουν την πρόβλεψη, μέσω του φίλτρου weka.filters.unsupervised.attribute.Remove-R1,2 (εικ. 3.1). Για την αποφυγή υπερεκπαίδευσης (overfitting) του μοντέλου εκτελέσαμε την εντολή weka.filters.unsupervised.instance.Randomize-S42 (εικ. 3.2), έτσι ώστε οι εγγραφές εντός του αρχείου να ανακατευτούν. Το επόμενο βήμα ήταν η εκτέλεση πλήθους αλγορίθμων που παρέχονται μέσω του Weka για δυαδική κατηγοριοποίηση, με βάση τις τιμές του πεδίου isEdge, με την επιλογή 10 fold cross validation.

[10] http://www.cs.waikato.ac.nz/ml/weka/



Σχήμα 3.7 Αποκοπή των ονομάτων των κόμβων μέσου του φίλτρου Remove.



Σχήμα 3.8 Τυχαία ταξινόμηση των εγγραφών με τη χρήση του φίλτρου Randomize.

3.4.2 Πρόβλεψη ακμών με βάση προηγούμενες χρονικές στιγμές

Στο δεύτερο μέρος των πειραμάτων επιλέξαμε να παρατηρήσουμε τη συμβολή των προηγούμενων στιγμών στην πρόβλεψη ακμών. Θέλήσαμε να ερευνήσουμε κατά πόσο οι χρονικές στιγμές πριν δημιουργηθεί μια ακμή την χρονική στιγμή t, επηρεάζουν την εξέλιξη του γράφου και τη συνδεσμολογία μεταξύ των χρηστών.

3.4.2.1 Διαδικασία εξαγωγής βαθμών κόμβων ανά ακμή

Στο πείραμα αυτό θεωρήσαμε και πάλι, όπως και στο πρώτο μέρος, ότι οι ακμές έχουν διάρκεια ζωής 3,5 και 10 δευτερολέπτων. Υπολογίσαμε για τα χρονικά διαστήματα 3,5 και 10 δευτερολέπτων από τη χρονική στιγμή που θα εμφανισθεί η ακμή και πριν τους εισερχόμενους, εξερχόμενους και αμφίδρομους βαθμούς των κόμβων που μετέχουν στην ακμή. Με τον τρόπο αυτό, παράξαμε εγγραφές που περιέχουν τρεις τιμές(βαθμούς) για κάθε κόμβο, μια τριάδα για κάθε δευτερόλεπτο παρατήρησης. Για παράδειγμα, θέλωντας να

υπολογίσουμε τις τιμές για διάρκεια 3 δευτερολέπτων από τη στιγμή που εμφανίσθηκε η ακμή και πριν θα παράξουμε μια εγγραφή της μορφής <$node_i$, $node_j$, $indegree_i$(t-2), $outdegree_i$(t-2), $bidegree_i$(t-2), $indegree_i$(t-1), $outdegree_i$(t-1), $bidegree_i$(t-1), $indegree_i$(t), $outdegree_i$(t), $bidegree_i$(t), $indegree_j$(t-2), $outdegree_j$(t-2), $bidegree_j$(t-2), $indegree_j$(t-1), $outdegree_j$(t-1), $bidegree_j$(t-1), $indegree_j$(t), $outdegree_j$(t), $bidegree_j$(t), 1>, δηλαδή συνολικά 6 τριάδες, όπως φαίνεται στον πίνακα 3.3. Αντιστοιχα για τα 10 δευτερόλεπτα θα παράξουμε 20 τριάδες κ.ο.κ. Όπως και στο πρώτο πείραμα, με τον ίδιο τρόπο υπολογίζουμε αντίστοιχα και τις εγγραφές για τις περιπτώσεις που δηλώνουν τη μη ύπαρξη ακμής, δηλαδή για κάθε δευτερόλεπτο κοιτάζουμε τον αντίστοιχο βαθμό του κόμβου την προηγούμενη χρονική στιγμή από αυτή που λαμβάνουμε υπόψην.

node A	**node B**	**tmp**	**node A**			**node B**			**is Edge**
			(t-2) in,out,bi	**(t-1) in,out,bi**	**(t) in,out,bi**	**(t-2) in,out,bi**	**(t-1) in,out,bi**	**(t) in,out,bi**	
80	20	6	0,0,0	0,0,0	0,1,0	0,0,0	0,0,0	1,0,0	1
80	20	6	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0
81	20	6	0,0,0,	0,0,0	0,1,0	0,0,0	0,0,0	2,0,0	1
81	20	6	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0
84	20	8	0,0,0	0,0,0	0,1,0	2,0,0	2,0,0	3,0,0	1
84	20	8	0,0,0	0,0,0	0,0,0	0,0,0	2,0,0	2,0,0	0

Πίνακας 3.4 Εγγραφές κατά την παρατήρηση των 3 δευτερολέπτων από τη στιγμή που εμφανίσθηκε η ακμή και πριν.

3.4.2.2 Εκπαίδευση και παραγωγή μοντέλου πρόβλεψης ακμών με βάση προηγούμενες χρονικές στιγμές

Κατά αντιστοιχία με το πρώτο πείραμα κατά τη διαδικασία της εκπαίδευσης του μοντέλου, εισάγαμε τα αρχεία μας στο weka και αφαιρέσαμε τα ονόματα των κόμβων που δεν

επηρεάζουν το αποτέλεσμα. Στη συνέχεια, εκτελέσαμε έναν αριθμό από αλγορίθμους που παρέχονται μέσω του προγράμματος για δυαδική κατηγοριοποίηση, με την επιλογή 10 fold cross validation και καταλήξαμε στα αποτελέσματα που παρουσιάζονται στο επόμενο κεφάλαιο,

4. Αποτελέσματα

Στο κεφάλαιο αυτό παραθέτουμε τα αποτελέσματα από την εκτέλεση ενός συνόλου από αλγόριθμους και τεχνικές κατηγοριοποίησης που παρέχει το εργαλείο Weka για κάθε πείραμα ξεχωριστά. Η ερμηνεία της απόδοσής τους θα συνεισφέρει στην εκτίμηση της επιρροής που ενδέχεται να ασκούν τα χαρακτηριστικά που επιλέξαμε να μελετήσουμε για την πρόβλεψη ακμών σε δίκτυο που μεταβάλλεται στο χρόνο.

4.1 Πρόβλεψη ακμών με βάση τη διάρκεια ζωής τους

Όπως ήδη αναφέραμε στο προηγούμενο κεφάλαιο, στο πρώτο μέρος των πειραμάτων μας μελετήσαμε τους βαθμούς εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών των κόμβων που μετέχουν σε κάθε ακμή με βάση τη διάρκεια ζωής τους. Στόχος μας ήταν να εξετάσουμε αν τα χαρακτηριστικά αυτά συνεισφέρουν στην πρόβλεψη ακμών. Τα αποτελέσματα της εκτέλεσης αλγορίθμων και τεχνικών που χρησιμοποιήθηκαν για αυτό το σκοπό παραθέτονται παρακάτω (πίνακας 4.1).

Διάρκεια Ζωής	**bayes**					
	Bayesian Logistic Regression	**BayesNet**	**Complement Naive Bayes**	**Naive Bayes**	**NaiveBayes Simple**	**NaiveBayes Updateable**
3’’	95.93%	99.10%	83.38%	93.97%	93.81%	93.97%

5''	94.66%	98.90%	82.64%	63.44%	63.14%	63.44%
10''	92.22%	98.47%	81.56%	59.04%	59.01%	59.04%
272''	79.91%	91.78%	71.92%	51.94%	51.92%	51.94%
12 ώρες	52.60%	86.38%	64.81%	49.89%	49.86%	49.89%
Διάρκεια Ζωής	**functions**					
	Logistic	**RBF Network**	**SPegasos**	**SMO**	**VotedPerceptron**	
3''	96.02%	91.85%	96.89%	95.55%	96.09%	
5''	94.68%	86.57%	94.54%	94.88%	95,26%	
10''	92.23%	90.20%	90.19%	92.87%	92.74%	
272''	81.07%	54.45%	50,01%	54.63%	77.77%	
12 ώρες	51.98%	49.90%	49.57%	50.34%	54.73%	
Διάρκεια Ζωής	**trees**					
	ADTree	**J48**	**Random Forest**		**RandomTree**	
3''	99.29%	99.46%	99.45%		99.45%	
5''	99.14%	99.37%	99.34%		99.33%	
10''	98.63%	99.13%	99.04%		99.06%	
272''	93.18%	94.32%	93.21%		93.22%	
12 ώρες	86.61%	87.62%	83.65%		83.68%	
Διάρκεια Ζωής	**meta**					
	AdaBoost M1 (Decision Stump)	**AttributeSelectedClassifier (J48)**	**ClassificationVia Regression (MSP)**		**Dagging (SMO)**	
3''	99.10%	99.10%	94.91%		95.61%	
5''	98.90%	99.10%	94.29%		94.91%	
10''	98.47%	98.47%	97.29%		93.86%	
272''	91.78%	91.78%	92.75%		50.08%	
12 ώρες	86.46%	86.46%	86.67%		49.98%	
Διάρκεια Ζωής	**rules**					
	OneR		**DecisionTable**			

3’’	94.91%	99.35%
5’’	94.24%	99.20%
10’’	93.06%	98.84%
272’’	82.06%	93.03%
12 ώρες	72.82%	87.03%

Πίνακας 4.1 Αποτελέσματα εκτέλεσης αλγορίθμων για το πρώτο μέρος των πειραμάτων.

Όπως παρατηρούμε όλοι οι αλγόριθμοι εμφανίζουν πολύ υψηλή απόδοση (πάνω από 90%) για διάρκεια ζωής ίσης με 3, 5 και 10 δευτερόλεπτα, γεγονός που μπορεί να δικαιολογηθεί αφού όπως ήδη αναφέραμε το αρχείο καταγραφής των ακμών περιέχει πολύ αραιά δεδομένα. Ακμές που εμφανίζονται ένα δευτερόλεπτο μπορεί να επανεμφανισθούν μετά από κάποια λεπτά, κάποιες ώρες ή και ποτέ ξανά στις 12 ώρες καταγραφής. Έτσι, οι βαθμοί των εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών των κόμβων που μετέχουν σε κάθε ακμή μηδενίζονται αρκετά συχνά αφού και η διάρκεια ζωής τους είναι μικρή (τάξης των 3, 5 και 10 δευτερολέπτων).

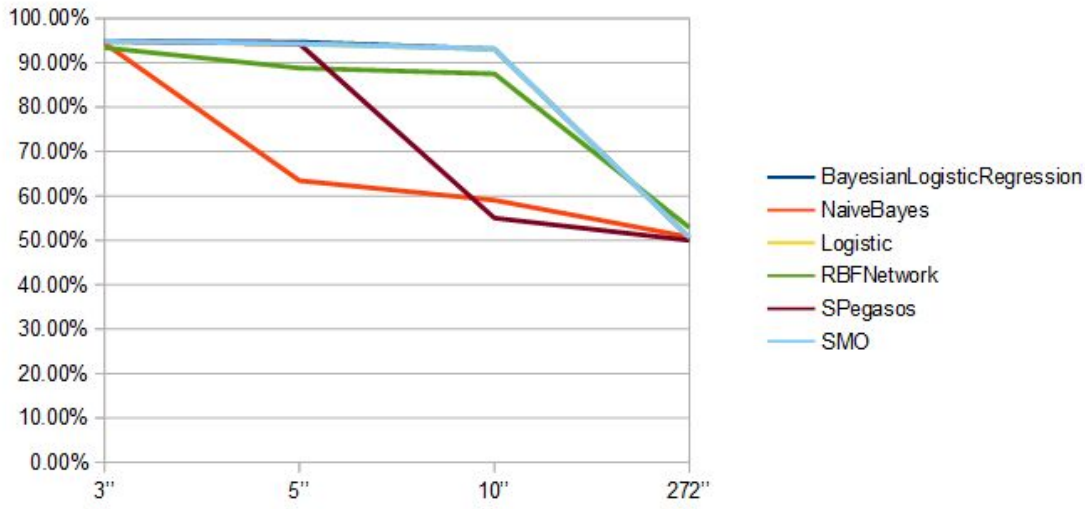
Λαμβάνοντας υπόψην ότι μια ακμή μπορεί να διαρκεί 272 δευτερόλεπτα, δηλαδή το μέσο χρόνο επανεμφάνισης όλων των ακμών του αρχείου, παρατηρούμε ότι η απόδοση μειώνεται αισθητά, με πιο χαρακτηριστικούς τους αλγορίθμους Bayesian Logistic Regression, Bayes Net, Complement Naive Bayes, RBFNetwork, SPegasos και SMO που δεν ξεπερνούν το 54%. Ο λόγος που συμβαίνει αυτό είναι γιατί τα 272 δευτερόλεπτα είναι ένας χρόνος που σχετίζεται μόνο με περίπου 8000 ακμές που χρειάζονται μέχρι 5 λεπτά για να επανεμφανισθούν, σύμφωνα με το σχήμα 3.3.

Εκτός από τα αποτελέσματα διάρκειας ζωής ίσης με 3, 5, 10 και 272 δευτερολέπτων παραθέσαμε και την απόδοση των αλγορίθμων για το σύνολο των ακμών που καταγράφηκαν για τις 12 ώρες, θέλωντας να διαπιστώσουμε πόσο επηρεάζουν την πρόβλεψη ακμών τα χαρακτηριστικά που επιλέξαμε χωρίς να λάβουμε υπόψη την εξέλιξη του γράφου στο χρόνο. Παρατηρήσαμε, λοιπόν, ότι εκτός από αλγορίθμους trees, meta και rules η πλειοψηφία των υπολοίπων εμφανίζει χαμηλά ποσοστά, χωρίς να ξεπερνούν σε κάποιες περιπτώσεις το 50%,

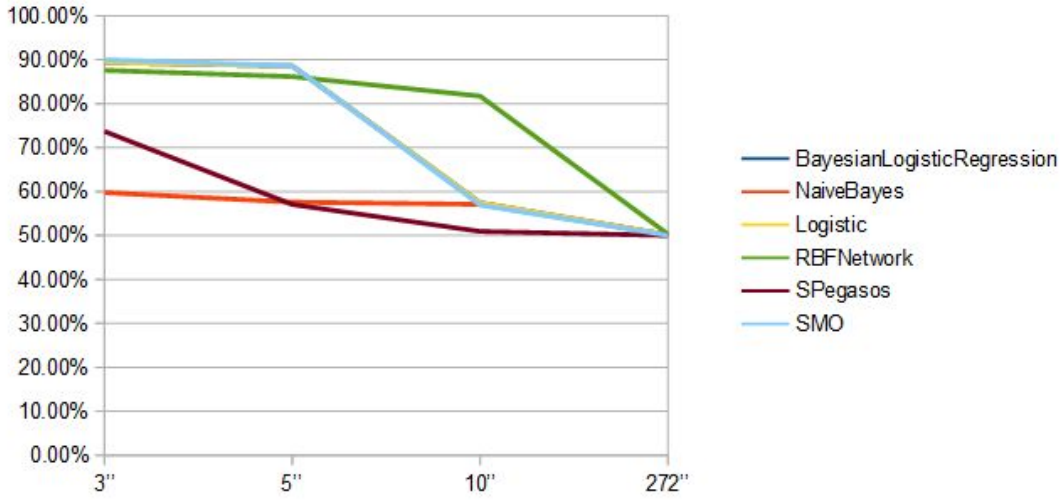
που είναι πιο λογικό αφού κατά την εμφάνιση κάθε ακμής υποθέτουμε ότι η ακμή αυτή έχει διάρκεια ζωής μέχρι και 12 ώρες (από την αρχή μέχρι και το τέλος του αρχείου).

Παράλληλα, οι αλγόριθμοι trees και κάποιοι από τους meta εμφανίζουν σχετικά μικρές μειώσεις στην απόδοσή τους (γύρω στο 12%) στα 3, 5 και 10 δευτερόλεπτα σε σχέση με τα 272 δευτερόλεπτα και τις 12 ώρες καταγραφής. Αυτό που συμβαίνει σε αυτές τις περιπτώσεις είναι ότι τα δέντρα απόφασης που δημιουργούνται κατά την εκτέλεση των αλγορίθμων trees αποτελούνται από λίγα μονοπάτια, αφού το κάθε δέντρο βασίζεται μόνο σε 6 χαρακτηριστικά, δηλαδή τους βαθμούς των εισερχόμενων, εξερχόμενων και αμφίδρομων κόμβων για κάθε ακμή. Τα μοντέλα που παράγονται από τα δέντρα απόφασης ενδέχεται, λοιπόν, να είναι υπερεκπαιδευμένα, γεγονός που δικαιολογεί τις μικρές διαφοροποιήσεις στην απόδοση των αλγορίθμων trees και meta.

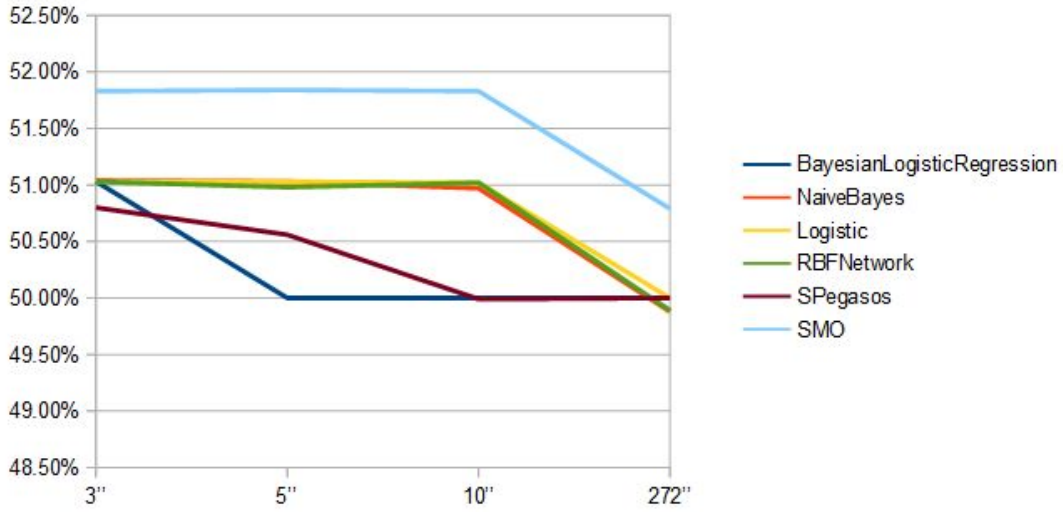
Για το λόγο αυτό επιλέξαμε να παρατηρήσουμε την απόδοση των χαρακτηριστικών που επιλέξαμε, δηλαδή τους βαθμούς των εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών, ξεχωριστά για να διαπιστώσουμε εάν συνεισφέρουν στην εξαγωγή κάποιου συμπεράσματος σχετικά με την πρόβλεψη ακμών. Στα σχήματα 4.1, 4.2 και 4.3 παραθέτονται οι αποδόσεις έξι αλγορίθμων που επιλέξαμε να εκτελέσουμε, που στον προηγούμενο πίνακα εμφάνιζαν ποσοστά γύρω στο 50% κατά τις 12 ώρες καταγραφής, για τους βαθμούς των εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών αντίστοιχα.



Σχήμα 4.1 Αξιοπιστία βαθμών εισερχόμενων κόμβων κατά την εκτέλεση αλγορίθμων για διάρκεια ζωης ίσης με 3, 5, 10 και 272 δευτερόλεπτα.



Σχήμα 4.2 Αξιοπιστία βαθμών εξερχόμενων κόμβων κατά την εκτέλεση αλγορίθμων για διάρκεια ζωης ίσης με 3, 5, 10 και 272 δευτερόλεπτα.



Σχήμα 4.3 Αξιοπιστία βαθμών αμφίδρομων κόμβων κατά την εκτέλεση αλγορίθμων για διάρκεια ζωης ίσης με 3, 5, 10 και 272 δευτερόλεπτα.

Όπως βλέπουμε, υψηλή απόδοση εμφανίζει ο βαθμός εισερχόμενων ακμών μέχρι και για διάρκεια ζωής ίσης με 10 δευτερόλεπτα, που σε αντίθεση με το βαθμό εξερχόμενων ακμών, η απόδοση μειώνεται κατακόρυφα ήδη μετά τα 5 δευτερόλεπτα διάρκειας ζωής. Απογοητευτικά ποσοστά εμφανίζει ο βαθμός αμφίδρομων ακμών αφού ακόμη από τα 3

δευτερόλεπτα η πρόβλεψη ακμών γίνεται τυχαία. Συμπεραίνουμε, λοιπόν, ότι τα χαρακτηριστικά που επιλέξαμε δε μπορούν να προβλέψουν με ασφάλεια ακμές με βάση τη διάρκεια ζωής τους.

4.2 Πρόβλεψη ακμών με βάση προηγούμενες χρονικές στιγμές

Στο τελευταίο μέρος των πειραμάτων μας επιλέξαμε να παρατηρήσουμε εάν οι βαθμοί εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών στις προηγούμενες χρονικές στιγμές, μόλις πριν εμφανισθεί δηλαδή μια ακμή, μπορεί να συνεισφέρει στην πρόβλεψη ακμών. Τα αποτελέσματα από την εκτέλεση πλήθους αλγοριθμων παραθέτονται παρακάτω (πίνακας 4.2).

Διάρκεια Ζωής	**bayes**					
	Bayesian Logistic Regression	**BayesNet**	**ComplementNaive Bayes**	**Naive Bayes**	**NaiveBayes Simple**	**NaiveBayes Updateable**
3''	96.66%	99.90%	95.08%	50.27%	50.25%	51.27%
5''	94.98%	99.87%	92.40%	50.81%	50.80%	50.81%
10''	63.97%	99.71%	90.01%	50%	50.02%	50%
Διάρκεια Ζωής	**functions**					
	Logistic	**RBF Network**	**Voted Perceptron**	**SMO**	**SPegasos**	**Simple Logistic**
3''	98.37%	57.09%	96.06%	60.15%	50.06%	98.52%
5''	98.37%	53.87%	94.35%	59.27%	50.05%	98.58%
10''	97.94%	51.94%	92%	58.50%	51.14%	98.62%
Διάρκεια Ζωής	**trees**					
	ADTree	**J48**		**Random Forest**		**Random Tree**
3''	99.91%	99.92%		99.91%		99.9%
5''	99.88%	99.92%		99.9%		99.9%
10''	99.78%	99.9%		99.8%		99.62%

Διάρκεια Ζωής	meta			
	AdaBoost M1	**AttributeSelected Classifier (J48)**	**Classification ViaRegression (MSP)**	**Dagging (SMO)**
3''	99.90%	99.90%	99.91%	50.46%
5''	99.87%	99.87%	99.88%	51.82%
10''	99.71%	99.71%	99.80%	50.55%
Διάρκεια Ζωής	**rules**			
	OneR		**DecisionTable**	
3''	97.55%		99.90%	
5''	96.69%		99.87%	
10''	95.35%		99.71%	

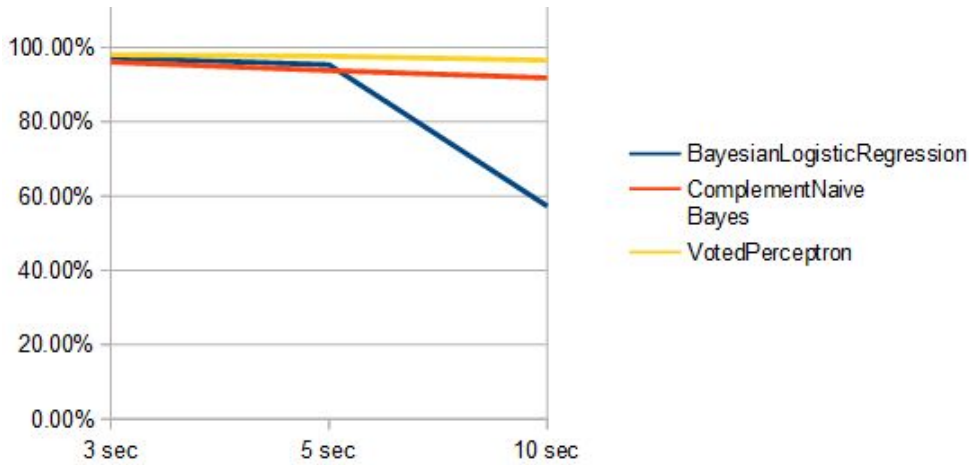
Πίνακας 4.2 Αποτελέσματα εκτέλεσης αλγορίθμων για το δεύτερο μέρος των πειραμάτων.

Όπως ήδη αναφέραμε και στο προηγούμενο πείραμα, το αρχείο καταγραφής αλληλεπιδράσεων μεταξύ των χρηστών στο δίκτυο του Twitter περιλαμβάνει αραιά δεδομένα, δηλαδή ακμές που εμφανίζονται τη χρονική στιγμή t μπορεί να εμφανισθούν ακόμη και μετά από κάποια λεπτά ή και ποτέ ξανά στο μέλλον. Παρατηρώντας τα δεδομένα με βάση τις προηγούμενες χρονικές στιγμές, δηλαδή 3, 5 και 10 δευτερόλεπτα πριν εμφανισθεί μια ακμή, βλέπουμε ότι παρουσιάζουν παρόμοια αποτελέσματα με το πρώτο πείραμα.

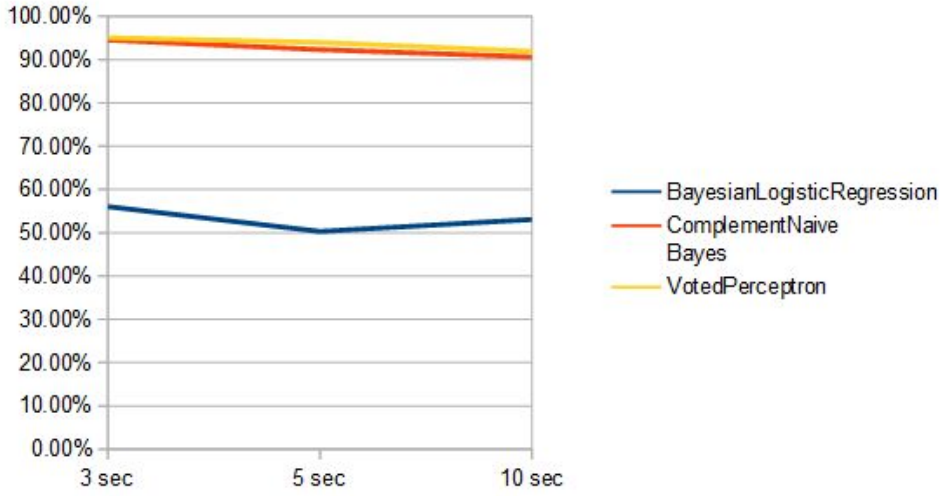
Πιο συγκεκριμένα, στους αλγορίθμους bayes άλλοι εμφανίζουν πολύ χαμηλά ποσοστά όπως οι NaiveBayes, NaiveBayesSimple και NaiveBayesUpdateable και άλλοι υψηλότερα ποσοστά, όπως οι BayessianLogisticRegression, BayesNet και ComplementNaiveBayes γεγονός που οφείλεται στον τρόπο υπολογισμού των αλγορίθμων. Παράλληλα, οι αλγόριθμοι functions αποδεικνύουν ότι οι προηγούμενες χρονικές στιγμές δεν προσδίδουν κάποιο ασφαλές συμπέρασμα σχετικά με την πρόβλεψη ακμών, αφού τα ποσοστά αξιοπιστίας τους δεν ξεπερνούν το 60.15%.

Τέλος, όπως και στο πρώτο πείραμα, οι αλγόριθμοι trees και κάποιοι meta που χρησιμοποιούν δέντρα απόφασης για την κατασκευή του μοντέλου, εμφανίζουν πολύ υψηλά ποσοστά, που ίσως και εδώ να δικαιολογεί την υπερεκπαίδευση του κάθε μοντέλου που κατασκευάζουν. Για παράδειγμα, για την παρατήρηση των 10 χρονικών στιγμών πριν εμφανισθεί μια ακμή κατασκευάζονται δέντρα απόφασης με αριθμό χαρακτηριστικών ίσο με 120 και για κάθε χαρακτηριστικό δημιουργείται και ένα μονοπάτι. Με τον τρόπο αυτό, το κάθε δέντρο απόφασης που δημιουργείται βασίζεται στις διάφορες εγγραφές που χρησιμοποιούνται για την κατασκευή του μοντέλου.

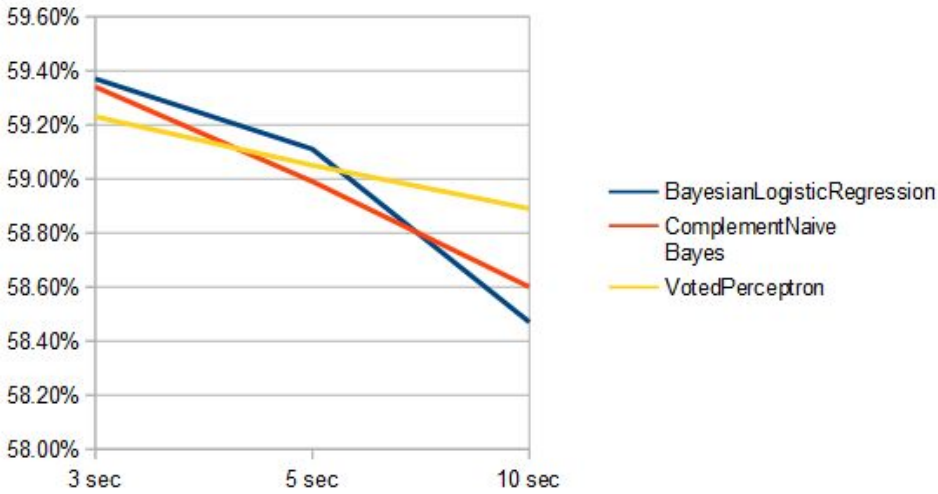
Αποφασίσαμε, λοιπόν, να παρατηρήσουμε την απόδοση των διαφορετικών βαθμών εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών ξεχωριστά για τους αλγόριθμους εκείνους που δεν εμφανίζουν σταθερά ή τυχαία αποτελέσματα για τις 3,5 και 10 χρονικές στιγμές πριν εμφανισθεί μια ακμή. Καταλήξαμε στα διαγράμματα που παρουσιάζονται στα σχήματα 4.4, 4.5 και 4.6 που παραθέτονται παρακάτω και περιγράφουν την απόδοση των εισερχόμενων, των εξερχόμενων και των αμφίδρομων ακμών κατά την εκτέλεση τριών διαφορετικών αλγορίθμων (BayesianLogisticRegression, ComplementNaiveBayes και VotedPerceptron).



Σχήμα 4.4 Αξιοπιστία βαθμών εισερχόμενων κόμβων κατά την εκτέλεση αλγορίθμων για τα προηγούμενα 3, 5 και 10 δευτερόλεπτα.



Σχήμα 4.5 Αξιοπιστία βαθμών εξερχόμενων κόμβων κατά την εκτέλεση αλγορίθμων για τα προηγούμενα 3, 5 και 10 δευτερόλεπτα.



Σχήμα 4.6 Αξιοπιστία βαθμών αμφίδρομων κόμβωνκατά την εκτέλεση αλγορίθμων για τα προηγούμενα 3, 5 και 10 δευτερόλεπτα.

Όπως και στο πρώτο πείραμα, έτσι και εδώ παρατηρούμε ότι καλύτερη απόδοση παρουσιάζει ο βαθμός εισερχόμενων και εξερχόμενων κόμβων με πιο σταθερή διαφοροποίηση μεταξύ των 3 και των 10 δευτερολέπτων. Αντίθετα, ο βαθμός αμφίδρομων κόμβων και εδώ εμφανίζει μια τυχαία πρόβλεψη ακμών, αφού το ποσοστό του ακόμη από τα 3 προηγούμενα δευτερόλεπτα παρατήρησης δεν ξεπερνά το 59%.

5 Συζήτηση

5.1 Συμπεράσματα

Είναι τελικά τα χαρακτηριστικά που επιλέξαμε, δηλαδή οι βαθμοί των εισερχόμενων, εξερχόμενων και αμφίδρομων ακμών, σημαντικοί παράγοντες που καθορίζουν την πρόβλεψη ακμών σε γράφο που εξελίσσεται στο χρόνο; Τα αποτελέσματα έδειξαν ότι οι οι βαθμοί αυτοί δε μπορούν με ασφάλεια να προβλέψουν την ύπαρξη ή μη ακμής στο μέλλον, λόγω ασυμφωνίας των παραγόμενων μοντέλων. Αλγόριθμοι functions και bayes είχαν σχετικά καλύτερη απόδοση σε σχέση με τα δέντρα απόφασης και τους meta αλγορίθμους, που παρήγαγαν υπερεκπαιδευμένα μοντέλα.

Επίσης, η διάρκεια ζωής των ακμών, δηλαδή τα 3, 5, 10 και 272 δευτερόλεπτα που παραμένουν ενεργές οι ακμές αμέσως μετά την εμφάνισή τους, δεν αποτελεί έναν αποδοτικό τρόπο παρακολούθησης της εξέλιξης του γράφου. Πιο συγκεκριμένα, τα 3, 5 και 10 δευτερόλεπτα εμφανίζουν μικρές διαφοροποιήσεις ειδικά σε σχέση με τα 272 δευτερόλεπτα ζωής, που καλύπτει ένα μικρό ποσοστό από ακμές που κατά μέσο όρο επανεμφανίζονται κατά τις 12 ώρες καταγραφής. Παρόμοια, οι 3, 5 και 10 χρονικές στιγμές πριν εμφανισθεί μια ακμή δεν προσφέρουν έναν αξιόλογο τρόπο παρατήρησης των ακμών στι χρόνο. Ακμές που εμφανίζονται τη χρονική στιγμή t, ίσως να μην έχουν εμφανισθεί ξανά στο παρελθόν και σε αντίθετη περίπτωση η εμφάνισή τους έχει γίνει πολύ πιο πριν στο παρελθόν.

Τέλος, παρατηρώντας τα χαρακτηριστικά που επιλέξαμε ξεχωριστά, συμπεραίνουμε ότι οι βαθμοί των εισερχόμενων και εξερχόμενων κόμβων μπορούν σε πολύ λίγες περιπτώσεις και με μικρή ακρίβεια να προβλέψουν την ύπαρξη ή μη ακμής στο μέλλον. Αντίθετα, ο βαθμός των αμφίδρομων ακμών ενός κόμβου με σιγουριά δεν καταφέρνει να προβλέψει την ύπαρξη ακμής με αξιοπιστία αλλά τυχαία.

5.2 Μελλοντικές Βελτιώσεις

Λαμβάνοντας υπόψην ότι τα δεδομένα που χρησιμοποιήσαμε για την πρόβλεψη ακμών σε εξελισσόμενο γράφο είναι αραιά, δηλαδή η επανεμφάνιση ακμών κατά τη διάρκεια των 12 ωρών καταγραφής γίνεται μετά από μεγάλα χρονικά διαστήματα ή και ποτέ, μια βελτίωση που θα μπορούσε να αλλάξει τα αποτελέσματα ίσως προέρχεται από τον καθαρισμό των δεδομένων. Πιο συγκεκριμένα, ακμές που επανεμφανίζονται μετά από το πέρας της μιας ώρας και πιο μετά θα μπορούσαν να μη ληφθούν υπόψη. Με τον τρόπο αυτό, τα χαρακτηριστικά που επιλέξαμε, δηλαδή οι βαθμοί των εισερχόμενων, εξερχόμενων και αμφίδρομων κόμβων θα εμφάνιζαν διαφορετικά αποτελέσματα, και θα αποφεύγαμε τις περιπτώσεις όπου υπάρχει μηδενισμός των χαρακτηριστικών αυτών μετά από 3, 5, 10 ή και 272 δευτερόλεπτα ζωής. Οι ακμές θα είναι χρονικά πιο κοντά μεταξύ τους και οι λιγότερο ενεργοί κόμβοι δε θα επηρέαζαν το αποτέλεσμα της πρόβλεψης.

Ένα ακόμη βήμα για τη βελτίωση των πειραμάτων θα ήταν η εύρεση ενός άλλου, παρόμοιου συνόλου δεδομένων με τα οποίο θα πραγματοποιούσαμε τη δοκιμή των μοντέλων που δημιουργήσαμε κατά την εκτέλεση των διαφορετικών τεχνικών και αλγορίθμων κατηγοριοποίησης. Τα δεδομένα αυτά θα περιλαμβάνουν έναν εξελισσόμενο γράφο και τις αλληλεπιδράσεις μεταξύ των χρηστών, όπως και στην περίπτωσή μας με το δίκτυο του Twitter και τις κοινοποιήσεις περιεχομένου των χρηστών. Με τα δεδομένα αυτά, θα ελένξουμε την αξιοπιστία κάθε αλγορίθμου ξεχωριστά και κατά πόσο είναι ακριβής η κατηγοριοποίηση που πραγματοποιεί. Έτσι, θα είμαστε σε θέση να συμπεράνουμε με μεγαλύτερη σιγουριά την περίπτωση που τα χαρακτηριστικά που επιλέξαμε αποτελούν έναν τρόπο να προβλέψουμε ακμές σε ένα γράφο που εξελίσσεται στο χρόνο.

Βιβλιογραφία

Adamic L., Adar E., (2001), 'Friends and neighbors on the web. Social Networks', 25:211-230, 2001.
Retrieved from *http://www.hpl.hp.com/research/idl/papers/web10/fnn2.pdf*

Boyd D.M., N.B. Ellison, (2007) 'Social network sites: Definition, history, and scholarship', Journal of Computer-Mediated Communication Vol. 13, Issue 1, p210-230, Blackwell.2007
Retrieved from http://www.danah.org/papers/JCMCIntro.pdf

Chen Wei, Wang Yajun, Yang Siyu, (2009), 'Efficient Influence Maximization in Social Networks', Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. p 199-20
Retrieved from http://research.microsoft.com/en-us/people/weic/kdd09_influence.pdf

Fowler, James H. and Nicholas A. Christakis. (2008), 'Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study' British Medical Journal 337, no. a2338: 1-9
Retrieved from
https://dash.harvard.edu/bitstream/handle/1/3685822/Christakis_DynamicSpreadHappiness.pdf?sequence=2

Fu W., Song L., Xing E.. (2009), 'Dynamic mixed membership blockmodel for evolving networks.' In *ICML*, pages 329–336. ACM, 2009.
Retrieved from *http://www.cs.cmu.edu/~epxing/papers/2009/fu_song_xing_icml09.pdf*

Hasan M. A., V. Chaoji, S. Salem, and M. Zaki, (2006), 'Link prediction using supervised learning'. In *Proc. of SDM '06 workshop on Link Analysis, Counterterrorism and Security*, 2006.
Retrieved from
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.118.2474&rep=rep1&type=pdf

Leroy V., B. B. Cambazoglu, and F. Bonchi., (2010), 'Cold start link prediction.', In *KDD '10*

Leskovec J., L. Backstrom, R. Kumar, and A. Tomkins, (2008), 'Microscopic evolution of social networks.', Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 462–470, New York, NY, USA, 2008. ACM.

Retrieved from *https://cs.stanford.edu/people/jure/pubs/microEvol-kdd08.pdf*

Leskovec J., Horvitz E., (2008), 'Planetary-scale views on a large instant-messaging network.', *Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.
Retrieved from *http://cs.stanford.edu/people/jure/pubs/msn-www08.pdf*

Liben-Nowell D., J. Kleinberg., (2003), 'The link prediction problem for social networks.', In *CIKM '03*, pages 556–559, 2003.
Retrieved from
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.7831&rep=rep1&type=pdf

Lichtenwalter R. N., J. T. Lussier, and N. V. Chawla. (2010), 'New perspectives and methods in link prediction.', In *KDD '10*, 2010.

Liu Lu, Tang Jie, Han Jiawei , (2012), 'Learning Influence from Heterogeneous Social Networks', Data Mining and Knowledge Discovery Journal, Volume 25, Issue 3, pp 511-544
Retrieved from
http://keg.cs.tsinghua.edu.cn/jietang/publications/DMKD12-Liu-Tang-et-al-Learning-Influence-from-Heterogeneous-Social-Networks.pdf

Mitzenmacher M., (2001), 'A brief history of lognormal and power law distributions', In Proceedings of the Allerton Conference on Communication, Control, and Computing, 2001.

Newman M., (2003), 'The structure and function of complex networks', *SIAM Review*, 45:167, 2003.
Retrieved from *http://arxiv.org/pdf/cond-mat/0303516.pdf*

Romero Daniel M., Galuba Wojciech, Asur Sitaram, Huberman Bernardo A., (2011), 'Influence and Passivity in Social Media', Volume 6913 of the series Lecture Notes in Computer Science pp 18-33 Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Proceedings, Part III
Retrieved from *http://www.hpl.hp.com/research/scl/papers/influence/influence.pdf*

Rossi Ryan A., Gallagher Brian, (2013), 'Modeling Dynamic Behavior in Large Evolving Graphs', Proceeding WSDM '13, Proceedings of the sixth ACM international conference on Web search and data mining, p 667-676

Retrieved from http://ryanrossi.com/papers/wsdm13-dbmm.pdf

Tang Jie, Sun Jimeng, Wang Chi, Yang Zi, (2009), 'Social Influence Analysis in Large-scale Networks', p 807-816, Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
Retrieved from
https://www.cs.cmu.edu/~ziy/pubs/KDD09-Tang-et-al-Social-Influence-Analysis.pdf

Tylenda Tomasz, Angelova Ralitsa, Bedathur Srikanta, (2009), 'Towards Time-aware Link Prediction in Evolving Social Networks', p.9, Proceedings of the 3rd Workshop on Social Network Mining and Analysis
Retrieved from
http://halma.mpi-inf.mpg.de/intranet/ag5/ag5publ.nsf/e127ff338913b2a3c12565f4005ef860/818fecf002c66936c12575fd00685881/$FILE/main.pdf

Sarkar Purnamrita, Chakrabarti Deepayan, Jordan Michael., (2014), 'Nonparametric link prediction in large scale dynamic networks.', Electron. J. Statist. 8 , no. 2, 2022--2065, Proceedings of the 29 th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012
Retrieved from http://arxiv.org/pdf/1109.1077v3.pdf

Sun Yizhou, Han Jiawei, Aggarwal Charu C., Chawla Nitesh V., (2012), "When Will It Happen? — Relationship Prediction in Heterogeneous Information Networks", Proc. 2012 ACM Int. Conf. on Web Search and Data Mining (WSDM'12), Seattle, WA
Retrieved from http://charuaggarwal.net/wsdm333-sun.pdf

Wang C., V. Satuluri, and S. Parthasarathy. (2007), 'Local probabilistic models for link prediction.', In *ICDM '07*, pages 322–331, 2007.
Retrieved from
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.572.1954&rep=rep1&type=pdf

Wickramaarachchi Charith , Kumbhare Alok, Frincu Marc, Chelmis Charalampos, Prasanna, Viktor K, (2015), "Real-Time Analytics for Fast Evolving Social Graphs", Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on, p.829-834
Retrieved from https://ganges.usc.edu/svn/pg/pubs/preprint/Charith-Scale-2015.pdf

Xing E., Fu W., Song L.. (2010), 'A state-space mixed membership blockmodel for dynamic network tomography'. *Ann. Appl. Stat.*, 4(2):535–566, 2010.

Retrieved from *http://www.cs.cmu.edu/~epxing/papers/2010/xing_fu_song_aoas09.pdf*