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

Ανάλυση και εξαγωγή ρόλων σε μεγάλους γράφους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σεραφείμ Χ. Λάμπρου

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Μάρτιος 2018

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

Ανάλυση και εξαγωγή ρόλων σε μεγάλους γράφους

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σεραφείμ Χ. Λάμπρου

Επιβλέπουσα : Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28η Μαρτίου 2018.

.....	.....	.....
Θεοδώρα Βαρβαρίγου	Εμμανουήλ Βαρβαρίγος	Δημήτριος Ασκούνης
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2018

.....

Σεραφείμ Χ. Λάμπρου

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

Περίληψη

Σκοπός της εργασίας αυτής αποτελεί η ανάλυση της συμπεριφοράς μεγάλων γράφων και η προσπάθεια μοντελοποίησης της.

Ο γράφος αποτελεί υπολογιστικό τρόπο αναπαράστασης ενός δικτύου, όπου κάθε κόμβος αντιπροσωπεύει ένα στοιχείο ή μια οντότητα του δικτύου και οι ακμές αντιπροσωπεύουν συνδέσεις και σχέσεις μεταξύ των στοιχείων.

Με τον όρο συμπεριφορά, εννοούμε την προσπάθεια διαχωρισμού της λειτουργίας κάθε κόμβου που ανήκει στον γράφο. Παραδείγματα λειτουργίας ενός κόμβου μέσα στον γράφο μπορεί να είναι η κεντρικότητα, η ιδιότητα της γέφυρας, ή η πηγή μιας ροής μέσα στον γράφο. Περιλαμβάνεται επίσης η έννοια της εξέλιξης, δηλαδή πως ένας κόμβος θα μεταβεί από μια λειτουργία σε μία άλλη κατά τη διάρκεια του χρόνου

Μοντελοποίηση εννοούμε την προσπάθεια εύρεσης ενός τρόπου αναπαράστασης όλων των δυνατών συμπεριφορών των κόμβων σε έναν γράφο με τέτοιο τρόπο που να ευνοείται η ομαδοποίηση, η εξαγωγή συμπερασμάτων και η ερμηνεία της κάθε πιθανής συμπεριφοράς. Στη μελέτη μας, ο τρόπος μοντελοποίησης που εξετάζουμε είναι η ανάλυση και εξαγωγή ρόλων και επικεντρωνόμαστε κυρίως σε κοινωνικά δίκτυα.

Ρόλος είναι ένας τρόπος ομαδοποίησης συγκεκριμένων λειτουργιών και χαρακτηριστικών ενός κόμβου ή μιας ακμής. Με τον τρόπο αυτό, κόμβοι, ή ακμές με συγκεκριμένες λειτουργίες και χαρακτηριστικά τα οποία έχουν προκαθοριστεί, ομαδοποιούνται και θα ανήκουν στον ίδιο ρόλο.

Τελικά όλες οι παραπάνω διαδικασίες, μπορούν να αναχθούν σε πράξεις πάνω σε γράφους.

Αρχικά προσπαθήσαμε να ακολουθήσουμε ένα μέρος της ερευνητικής πορείας προκειμένου να ορίσουμε ορθά το πρόβλημα. Έγινε προσπάθεια σύνδεσης της εργασίας με την εξέλιξη ενός δικτύου. Μέσα από αυτή την ανάλυση προέκυψε, ως μοντέλο περιγραφής της συμπεριφοράς του δικτύου, η εξαγωγή και η ανάλυση ρόλων.

Αφού αναλύσαμε πλήρως την έννοια του ρόλου και της ισοδυναμίας, τόσο σε κοινωνικό, όσο και σε μαθηματικό και τεχνικό επίπεδο, παρουσιάσαμε μερικά μαθηματικά μοντέλα εφαρμογής τους. Δόθηκε μεγαλύτερη έμφαση σε ένα από τα μοντέλα – αλγόριθμους, προκειμένου να αναλυθεί η λειτουργία του.

Προέκυψαν έτσι δύο επιμέρους αλγόριθμοι, ένας για ανάλυση χαρακτηριστικών και ένας για την εξαγωγή ρόλων, ο Refex και ο Rolx.

Αναλύσαμε τι σημαίνει αναπαράσταση γράφου με χαρακτηριστικά, τι σημαίνουν τα χαρακτηριστικά και εξηγήσαμε τους επιμέρους υπολογισμούς των αλγορίθμων.

Στο τελικό στάδιο πραγματοποιήσαμε δύο πειράματα με εκτελέσεις των αλγορίθμων. Έγινε στατιστική ανάλυση των δεδομένων εισόδου, των παραμέτρων καθώς και τον αποτελεσμάτων και προσπαθήσαμε να εξαγάγουμε κάποια συμπεράσματα.

Τέλος αναφέραμε πρακτικές εφαρμογές και βελτιώσεις των παραπάνω εργαλείων.

Λέξεις Κλειδιά

Ανάλυση κοινωνικών δικτύων, εξαγωγή ρόλων, μοντελοποίηση συμπεριφοράς δικτύου, εξαγωγή χαρακτηριστικών, ανάλυση γράφων, πρόβλεψη δικτύου

Abstract

The purpose of this work is to analyze the behavior of large graphs and study ways of modeling that behavior.

A graph is a computational way to represent a network, where each node of the graph represents an element or an entity of the network and each edge of the graph represents a connection or a relationship between the elements.

Using the term behavior, we try to distinguish each node's unique functionality and its relative metrics to the other nodes and find a way to classify nodes which are equal in that aspect. Examples of functionalities and relative metrics could be the centrality of node, its position in the graph as a bridge, or being the source of a flow in the graph. Behavior also includes the concept of evolution that is how a node transforms from a functionality to another through time.

Modeling is a way to accumulate and represent every possible behavior of the nodes of a graph in a way to make classification, conclusion drawing and behavior interpretation easier. The model we choose to study in this work is role extraction and analysis, especially in social networks.

Role is a way of classifying certain functionalities and features of a node or an edge. Nodes with specific, predetermined functionalities or features will be grouped and belong to the same role in the graph.

Finally all the above procedures could be applied as mathematical operations on the graph.

We tried to track down part of the scientific research, concerning the problem, in order to define it properly. There was also a connection of this work with the link prediction problem in graph evolution. During the research, role extraction came up as a way to model and analyze graph behavior.

After thorough analysis of the meaning of role and equivalence, both from a social and a mathematical – technical aspect, we presented some mathematical models that apply roles. Emphasis was given to one of the models in order to analyze it.

The model was composed of two algorithms, the first one, Refex, for feature extraction and the second one, Rolx, for role extraction. We explained what features representation means on a graph and performed a step by step explanation of the algorithms.

Final step of our study was to see the algorithms in action. We performed two experiments on the algorithms with two different datasets. Statistical analysis was performed on the input dataset, algorithm parameters, as well as on outputs and we tried to draw some conclusion.

On the last part of this work, some every day applications of the algorithm, as well as some improvements were suggested

Keywords

Social network analysis, role extraction, network behavior modeling, feature extraction, graph analysis, network prediction, data mining

Πίνακας περιεχομένων

Κεφάλαιο 1 - Εισαγωγή στο πρόβλημα

1.1 Εισαγωγή

Σκοπός της παρούσας διπλωματικής εργασίας αποτέλεσε η μελέτη κοινωνικών και άλλων δικτύων με στόχο την έρευνα του προβλήματος της πρόβλεψης συνδέσμων (link prediction – data mining). Με την εξέλιξη των κοινωνικών δικτύων και την ικανότητα συλλογής όλο και περισσότερων δεδομένων, δεν αποτελεί έκπληξη ότι η μελέτη του προβλήματος αυτού, βρίσκεται στο επίκεντρο τα τελευταία χρόνια. Στη συγκεκριμένη εργασία, θα ακολουθήσουμε κάποιους σταθμούς στην πορεία της έρευνας για την ανάλυση κοινωνικών δικτύων με έμφαση στο κομμάτι της πρόβλεψης, και τελικά θα διερευνήσουμε σε βάθος την εξαγωγή ρόλων, που αποτελεί ένα μέρος της διαδικασίας πρόβλεψης συνδέσμων. Οι ρόλοι θα χρησιμοποιηθούν αργότερα ως βάση για προβλέψεις και συμπεράσματα. Θα αναφερθούν αρκετές από τις προτάσεις που υπήρξαν μέχρι σήμερα για εξαγωγή ρόλων, με γνώμονα και την ιστορική εξέλιξή τους και θα αναλυθεί με λεπτομέρεια, μία από τις πιο πρόσφατες.

Σε αυτό το σημείο, φαίνεται σωστό να δώσουμε έναν πρώτο, σύντομο και μη αυστηρό ορισμό για τον ρόλο.

> Ρόλος είναι ένας τρόπος ομαδοποίησης συγκεκριμένων λειτουργιών και χαρακτηριστικών ενός κόμβου ή μιας ακμής σε έναν γράφο. Με τον τρόπο αυτό κόμβοι, ή ακμές με συγκεκριμένες λειτουργίες και χαρακτηριστικά τα οποία έχουν προκαθοριστεί, θα ανήκουν στον ίδιο ρόλο.

Εξαγωγή ρόλων είναι η διαδικασία υπολογισμού και καθορισμού της βέλτιστης ομαδοποίησης λειτουργιών και χαρακτιριστικών.

Όταν αναφερόμαστε σε ανάλυση κοινωνικών δικτύων και εξαγωγή ρόλων, φαίνεται λογική η ενασχόληση με γράφους και έναν πολύ μεγάλο όγκο δεδομένων. Αναπόφευκτα λοιπόν η μελέτη μας, έρχεται σε επαφή με big data και την όσο το δυνατόν αποδοτικότερη διαχείρισή τους. Στη πορεία λοιπόν θα συναντήσουμε στοιχεία του αλγόριθμου map-reduce, καθώς και αρκετά μαθηματικά εργαλεία που βοηθούν στην κατεύθυνση αυτή.

Στο τέλος της εργασίας θα δοκιμάσουμε μία από τις μεθόδους μέχρι το τμήμα της εξαγωγής ρόλων

1.2 Μελέτη κοινωνικών δικτύων

Η μελέτη των κοινωνικών δικτύων προέκυψε ως τμήμα της εκτεταμένης ανάλυσης πολύπλοκων δικτύων και των ιδιοτήτων τους. Στον κλάδο της πληροφορικής, κοινωνικά δίκτυα ονομάζουμε δομές δεδομένων, όπου οι κόμβοι αντιπροσωπεύουν ανθρώπους ή οντότητες που σχετίζονται σε κάποιο κοινωνικό πλαίσιο, και οι ακμές εκφράζουν δραστηριότητα, συνεργασία ή επιρροή ανάμεσα στους κόμβους. Η ανάλυση των κοινωνικών δικτύων επομένως επωφελείται από την αναπαράστασή τους ως γράφους. Μπορούμε να σκεφτούμε πολλά τέτοια παραδείγματα. Μια τέτοια δομή θα ήταν ικανή να αναπαραστήσει τη δραστηριότητα του Facebook, όπου οι κόμβοι ισοδυναμούν με τα μέλη και οι ακμές με τις σχέσεις φιλίας, τα like ή τα mention, θα μπορούσε να αναπαραστήσει τα κανάλια και τους χρήστες του YouTube, ή τα retweets και τα favorites στο twitter.

Γίνεται εύκολα αντιληπτό, ότι τα κοινωνικά δίκτυα αποτελούν δυναμικές οντότητες, αλλάζουν γρήγορα συναρτήσει του χρόνου και η μελέτη τους παρουσιάζει αρκετές δυσκολίες τόσο στη διαχείριση όσο και στη χρήση των δεδομένων. Η κατανόηση του μηχανισμού και των αρχών, βάση των οποίων εξελίσσονται αποτελεί πρόκληση και θα προσπαθήσουμε να συνεισφέρουμε σε αυτήν την κατεύθυνση.

1.2.1 Ορισμός προβλήματος

Για αρχή ας ορίσουμε το πρόβλημα. Με δεδομένο ένα πλήθος στιγμιότυπων ενός κοινωνικού δικτύου σε ένα διάστημα Δt, θέλουμε να προβλέψουμε με ακρίβεια τις ακμές που θα παρουσιαστούν στο δίκτυο από μία χρονική τιμή μεταγενέστερη του Δt, έως την στιγμή t'. Προσπαθούμε δηλαδή να κατανοήσουμε σε ποιο βαθμό ένα κοινωνικό δίκτυο μπορεί να μοντελοποιηθεί με την χρήση εσωτερικών χαρακτηριστικών του δικτύου/γράφου. Ενστικτωδώς καταλαβαίνουμε ότι ένα τέτοιο μοντέλο είναι χρήσιμο, μόνο αν συμβάλει στην εξαγωγή χρήσιμων συμπερασμάτων.

1.2.2 Εφαρμογές του προβλήματος

Αποτελεσματικοί τρόποι πρόβλεψης συνδέσμων θα μπορούσαν να χρησιμεύσουν στην ανάλυση κοινωνικών δικτύων, ώστε να προτείνουν αξιόλογες συνεργασίες μεταξύ κόμβων, για θέματα ασφάλειας, όπως πρόβλεψη εγκληματικών ενεργειών, για τον εντοπισμό πιθανόν σχέσεων για τις οποίες δεν υπάρχουν ακόμα επαρκείς πληροφορίες.

1.2.3 Προσεγγίσεις του προβλήματος

Διάφορες προσεγγίσεις έχουν εμφανιστεί για την επίλυση του προβλήματος. Σε αυτήν την ενότητα θα κάνουμε μία επισκόπηση σε μερικές από αυτές και θα παρατηρήσουμε πως από τις πιο απλές προσεγγίσεις, φτάσαμε στις σύνθετες που χρησιμοποιούνται σήμερα. Σκοπός μας δεν είναι η σε βάθος μαθηματική ανάλυση τους, αλλά να κατανοήσει ο αναγνώστης τις ιδέες πίσω από κάθε προσέγγιση και πως μέσα από αυτές τις προσεγγίσεις φτάσαμε στο πιο ολοκληρωμένο μοντέλο που θα εξετάσουμε αναλυτικά στη συνέχεια.

The link prediction problem for social networks (Nowell – Kleinberg)

Μία από τις πρώτες προσεγγίσεις που προέκυψαν και στην ουσία έθεσε τις βάσεις για την περεταίρω μελέτη του προβλήματος ήταν αυτή των Nowell και Kleinberg το 2004. Με την χρήση των εσωτερικών μετρικών του δικτύου, εφαρμόζοντας θεωρία γράφων, αρχές της επιστήμης των υπολογιστών και κοινωνικές επιστήμες, προσπάθησαν να εντοπίσουν ποιες είναι οι πιο έγκυρες και χρήσιμες μετρικές στο κομμάτι της πρόβλεψης. Προσπάθησαν επομένως να αναγνωρίσουν τα μετρήσιμα χαρακτηριστικά του δικτύου τα οποία μπορούν να οδηγήσουν σε μία σχετικά ασφαλή εκτίμηση για τη μελλοντική κατάστασή του. Υπήρχε ειδικό ενδιαφέρον στο ποιες νέες ακμές θα δημιουργηθούν

Θεωρώντας δεδομένες τις ακμές ενός κοινωνικού δικτύου κατά τη διάρκεια ενός χρονικού διαστήματος, το οποίο είχε τον ρόλο του διαστήματος εκμάθησης, ο αλγόριθμος τους είχε ως έξοδο ακμές, που αναμένονταν να εμφανιστούν σε κάποιο επόμενο χρονικό διάστημα. Δεν μπορούσαν φυσικά να υπολογιστούν ακμές σε κόμβους που δεν υπήρχαν κατά το διάστημα εκμάθησης. Τελικά το αποτέλεσμα ήταν μία λίστα με όλα τα πιθανά ζευγάρια κόμβων που μπορούσαν να ενωθούν με ακμή, μαζί με μια βαθμολογία που υποδείκνυε τη πιθανότητα να αναπτύξουν τελικά την ακμή αυτή. Μερικές από τις μετρικές που δοκίμασαν με τον αλγόριθμό τους ήταν οι παρακάτω:

graph distance	(negated) length of shortest path between x and y
common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Adamic/Adar	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log\lvert\Gamma(z)\rvert}$
preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Katz_β	$\sum_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\mathsf{paths}_{x,y}^{\langle\ell\rangle}\rvert$ where $\mathsf{paths}_{x,y}^{\langle\ell\rangle}$:= {paths of length exactly ℓ from x to y} weighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle}$:= number of collaborations between x, y. unweighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle} := 1$ iff x and y collaborate.
hitting time	$-H_{x,y}$
stationary-normed	$-H_{x,y} \cdot \pi_y$
commute time	$-(H_{x,y} + H_{y,x})$
stationary-normed	$-(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$ where $H_{x,y}$:= expected time for random walk from x to reach y π_y := stationary distribution weight of y (proportion of time the random walk is at node y)
rooted PageRank_α	stationary distribution weight of y under the following random walk: with probability α, jump to x. with probability $1-\alpha$, go to random neighbor of current node.
SimRank_γ	$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} \mathsf{score}(a,b)}{\lvert\Gamma(x)\rvert\cdot\lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$

Οι μετρικές αυτές χρησιμοποιήθηκαν για την ανάδειξη της πιθανότητας εμφάνισης συνδέσμου μεταξύ ενός κόμβου x και ενός κόμβου y. Ο συμβολισμός Γ(x) αντιπροσωπεύει τους γείτονες του κόμβου x.

Μπορούμε να διακρίνουμε τριών ειδών μετρικές:

Μετρικές βασισμένες σε γειτονιές κόμβων. Τέτοιες είναι οι "common neighbors", "Jacard's coefficient", "Adamic/Adar" και "preferential attachment". Η ιδέα των μετρικών αυτών, βασίζεται στην άποψη, ότι δύο κόμβοι που έχουν πολλούς κοινούς γείτονες, είναι πιθανότερο να αναπτύξουν σχέση μεταξύ τους.

Μετρικές βασισμένες στο σύνολο όλων των δυνατών διαδρομών, όπως οι "Katz", "Hitting time" και "Page Rank". Οι μέθοδοι αυτοί βασίζονται στο πλήθος και το μήκος των διαδρομών μεταξύ δύο κόμβων ώστε να εξάγουν την πιθανότητα εμφάνισης δραστηριότητας μεταξύ τους. Η γενική ιδέα είναι ότι όσες περισσότερες διαδρομές υπάρχουν και όσο πιο κοντινές είναι μεταξύ τους, τόσο αυξάνεται η πιθανότητα μελλοντικής σχέσης.

Μετρικές που προκύπτουν ύστερα από τροποποίηση της αρχικής αναπαράστασης του δικτύου με αποτέλεσμα τη μείωση του θορύβου στα δεδομένα και στη συνέχεια εφαρμογή κάποιας από τις προηγούμενες μετρικές. Για παράδειγμα χρησιμοποιείται ένας low rank πίνακας γειτνίασης σε συνδυασμό με την "Katz" και την "common neighbors", είτε υπολογίζεται η βαθμολογία του κόμβου x, μέσω της ομοιότητας με κάποιον

κόμβο z, είτε πραγματοποιείται πρώτα clustering και pruning για αποφυγή του θορύβου σε ζεύγη με χαμηλή βαθμολογία.

Τα αποτελέσματα ήταν ενθαρρυντικά, χωρίς όμως την απαραίτητη βεβαιότητα της πρόβλεψης σε όλες τις περιπτώσεις. Όλες οι μετρικές κατέγραψαν καλύτερες επιδόσεις από την τυχαία επιλογή. Ξεχώρισαν οι "Adamic/Adar", η "Katz" και η low rank σε συνδυασμό με την "common neighbors".

Towards time-aware link prediction in evolving social networks (Angelova - Bedathur)

Ως λογική συνέχεια του παραπάνω μοντέλου, που βασιζόταν στο σύνολο του διαστήματος εκμάθησης, αποτελεί η έμφαση στη χρονική εξέλιξη ενός κοινωνικού δικτύου, ώστε σε συνδυασμό με μετρικές για πολλά στιγμιότυπα να προκύψουν μελλοντικά συμπεράσματα. Στο σημείο αυτό, θα εξετάσουμε την δουλειά των Angelova και Bedathur.

Κύριο χαρακτηριστικό της μελέτης αποτελεί η έμφαση στην χρονική εξέλιξη του δικτύου, βάση της λογικής, ότι γεγονότα που συνέβησαν παλαιότερα, θα έχουν λιγότερο αντίκτυπο σε μελλοντικές μεταβάσεις. Η προσέγγιση διαφέρει επίσης από την προηγούμενη και στο γεγονός ότι δίνεται έμφαση και στο είδος της σχέσης μεταξύ δύο κόμβων και στη διάκριση μεταξύ ύπαρξης μίας μόνο ακμής ή παράλληλων ακμών μεταξύ τους. Οι σχέσεις μεταξύ δύο κόμβων διαχωρίζονται σε σταθερές σχέσεις και διακριτά γεγονότα. Οι σταθερές σχέσεις αφορούν για παράδειγμα φιλία και εντοπίζονται σε απόσταση μήκους ενός ή δύο βημάτων από τον κόμβο προς εξέταση. Στα διακριτά γεγονότα, όπως ένα like, οι σχέσεις είναι πιθανότερο να εμφανιστούν σε κόμβους που είναι ήδη συνδεδεμένοι μεταξύ τους, με αποτέλεσμα παράλληλες ακμές. Γίνεται λοιπόν διαχωρισμός του προβλήματος στην πρόβλεψη νέων συνδέσμων και στην ενίσχυση παλαιών. Τέλος, δίνεται έμφαση κυρίως στην προσπάθεια πρόβλεψης των ακμών γύρω από ένα συγκεκριμένο κόμβο, παρά στο συνολικό εντοπισμό όλων των ακμών όπως στην προηγούμενη μελέτη.

Η μέθοδος έχει ως εξής. Οι γείτονες του προς εξέταση κόμβου ταξινομούνται, με βάση το πόσος χρόνος πέρασε από την προηγούμενη αλληλεπίδρασή τους, το πλήθος των αλληλεπιδράσεων μεταξύ τους και το αν υπήρχαν άλλοι συμμετέχοντες στην ίδια αλληλεπίδραση (γεγονός). Αυτά για την πρόβλεψη ενίσχυσης των σχέσεων. Χρησιμοποιήθηκε επίσης η ίδια λογική για την πρόβλεψη εντελώς νέων σχέσεων με τη δημιουργία βαρών και εφαρμογή τους στις μετρικές Adamic/Adar και Page rank. Για τη εύρεση του τελικού αποτελέσματος χρησιμοποιήθηκε το πιθανολογικό μοντέλο του Wang μαζί με πολλαπλασιαστές Lagrange.

Τα αποτελέσματα της έρευνας αυτή ήταν σημαντικά καλύτερα της προηγούμενης και άνοιξαν νέους ορίζοντες στην πρόβλεψη ρόλων.

A trust-aware system for personalized user recommendations in social networks (Eirinaki, Louta, Varlamis)

Η επόμενη μελέτη που θα εξετάσουμε δείχνει πως από τεχνικές όπως η προηγούμενες, μπορούμε να οδηγηθούμε σε ένα αφαιρετικό επίπεδο πιο πάνω και να εφαρμόσουμε τελικά το πρόβλημα της πρόβλεψης συνδέσμων. Όπως και στα προηγούμενα, έτσι και εδώ δεν θα αναλωθούμε σε μαθηματικές λεπτομέρειες και μας ενδιαφέρει το εννοιολογικό πλαίσιο της μελέτης.

Στην δουλεία των Eirinaki, Louta και Varlamis παρατηρούμε την ανάλυση του κοινωνικού φαινομένου της εμπιστοσύνης από την οπτική γωνία των social analytics, με στόχο την πρόταση δραστηριοτήτων και τελικά την αναγωγή του στο link prediction problem. Οι συγγραφείς θέλουν να φτιάξουν μία μέθοδο, που να υπολογίζει το trust μεταξύ χρηστών, με στόχο να τους προτείνει πιθανές αλληλεπιδράσεις.

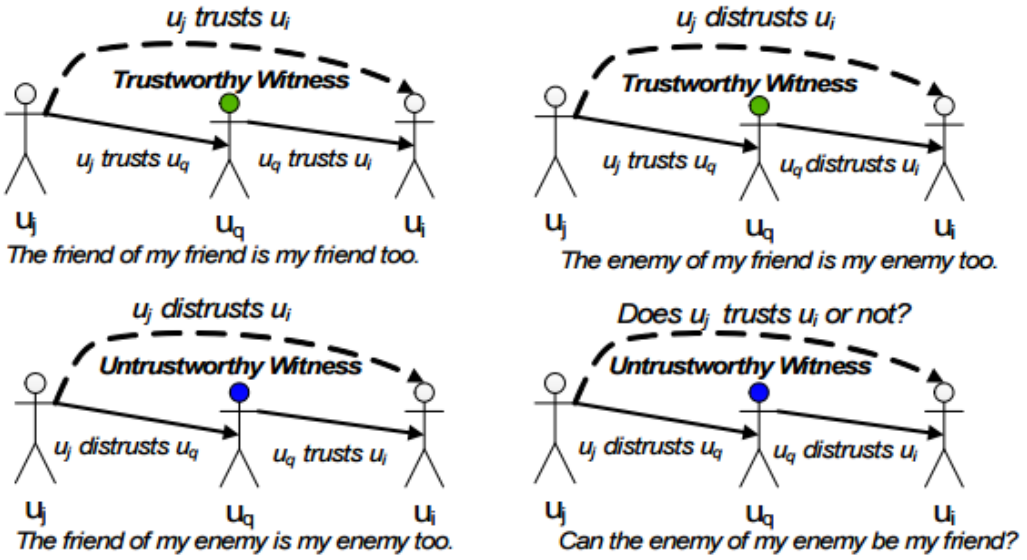
Το σύστημα αυτό βασίζεται σε μια τεχνική βαθμολόγησης των χρηστών, με βάση παρατηρήσεις, προηγούμενες εμπειρίες καθώς και γνώμες άλλων χρηστών. Προκειμένου να υπολογιστεί η φήμη κάθε χρήστη, υιοθετούνται ιδιότητες της εμπιστοσύνης όπως μεταβατικότητα, εξατομίκευση και τα συμφραζόμενα, ιδέες προερχόμενες από το πεδίο της κοινωνιολογίας. Η εμπιστοσύνη δεν θεωρείτε εντελώς μεταβατική, καθώς φθίνει με τον χρόνο. Εδώ υπάρχει σύνδεση με το στοιχείο του χρόνου και της εξέλιξης που εξετάσαμε στην προηγούμενη υποενότητα. Τέλος θεωρούν ότι η εμπιστοσύνη μπορεί να έχει θετικό πρόσημο, αρνητικό, ή να είναι ουδέτερη.

Η υλοποίηση χωρίζεται σε τρεις φάσεις. Στην πρώτη φάση αναλύεται το σχήμα σύνδεσης των χρηστών. Γίνεται διαφοροποίηση μεταξύ αλληλεπιδράσεων που εκφράζουν άμεση εμπιστοσύνη (καθαρή έκφραση γνώμης) ή έλλειψη αυτής και αυτών που την εκφράζουν με έμμεσο τρόπο (π.χ. ύπαρξη ενός like ή σχολίου). Έτσι σχηματίζονται τέσσερεις κατηγορίες σχέσεων εμπιστοσύνης.

- Άμεση από χρήστη σε χρήστη – εκφράζει πιο ισχυρούς δεσμούς, όπως φιλία.
- Άμεση από χρήση σε αντικείμενο – π.χ. το κουμπί +1
- Έμμεση από χρήστη σε αντικείμενο – π.χ. αναρτήσεις που περιέχουν συνδέσμους που οδηγούν σε κάποιο άλλο αντικείμενο.
- Έμμεση από χρήστη σε χρήστη – αλληλεπίδραση με αντικείμενα άλλου χρήστη, εκφράζουν έμμεση αλληλεπίδραση με τον χρήστη.

Στην δεύτερη φάση γίνεται βαθμολόγηση της φήμης κάθε χρήστη. Ποσοτικοποιούνται οι παραπάνω συνδέσεις και τελικά έχουμε ως αποτέλεσμα προσωπικές βαθμολογίες φήμης των υπόλοιπων μελών για τον κάθε χρήστη, οι οποίες εκφράζουν την τοπική άποψη του χρήστη για υπόλοιπα μέλη του δικτύου. Οι βαθμολογίες αυτές είναι διαφορετικές ανάλογα με το ποιος κρίνει, δηλαδή οι χρήστες Α και Β θα έχουν

διαφορετικές βαθμολογίες (εμπιστοσύνη) ως προς τον χρήστη Γ Στην παραπάνω βαθμολογία λαμβάνονται υπόψη η άποψη κάθε χρήστη για το προς εξέταση κάθε φορά κόμβο, καθώς και οι απόψεις των γειτόνων του κόμβου, ο οποίος κρίνει. Μέσω των άμεσων και έμμεσων (μέσω τρίτων) διασυνδέσεων των χρηστών, σχηματίζεται τελικά η σχετική βαθμολογία, δίνοντας μεγαλύτερη έμφαση στις πιο πρόσφατες αξιολογήσεις



.

Στην Τρίτη και τελευταία φάση, το σύστημα αυτό χρησιμοποίει τις παραπάνω αξιολογήσεις, ώστε να προτείνει νέες διασυνδέσεις στους χρήστες, θετικές ή αρνητικές (ένα καλό παράδειγμα, ώστε να κατανοήσουμε την λειτουργικότητα είναι τα date apps, που κάνουν προτάσεις, βάση στοιχείων της προσωπικότητας). Οι αρνητικές αξιολογήσεις μπορούν να χρησιμοποιηθούν, ως προειδοποιήσεις προς τους χρήστες.

1.3 Σύνοψη

Συνοψίζοντας, με τις παραπάνω μελέτες ως οδηγό, θέλαμε να υποδείξουμε τα εξής.

A. Τον ορισμό του προβλήματος της πρόβλεψης συνδέσμων και την αξιολόγηση τοπικών μετρικών του δικτύου για την επίλυση του.
B. Την εισαγωγή του χρόνου και της εξέλιξης ως δεδομένα για την ακριβέστερη επίλυση του προβλήματος.

C. Τέλος, θέλαμε να εξετάσουμε και την αντίστροφη πορεία, δηλαδή την προσπάθεια μεταφοράς και αναγωγής ενός κοινωνικού φαινομένου σε μεταβλητές του προβλήματος και ενός συνδυασμού μεταβλητών του δικτύου

Στη συνέχεια θα εξετάσουμε σε βάθος, ένα πιο ολοκληρωμένο πλαίσιο πρόβλεψης. Μέσα από τη χρήση μετρικών του δικτύου, μοντελοποιούνται οι στιγμιαίες συμπεριφορές κάθε κόμβου και τελικά ταξινομούνται σε ρόλους. Αφού το σύστημα παρακολουθήσει τη μετάβαση της συμπεριφοράς αυτής, δηλαδή πόσο συχνά ένας κόμβος μεταβαίνει από έναν ρόλο σε κάποιον άλλο, μπορεί να προβλέψει την μελλοντική εξέλιξη του δικτύου, αλλά και να εντοπίσει ασυνήθιστες (μη αναμενόμενες) συμπεριφορές. Το σύστημα αυτό είναι δυναμικό και προσαρμόζεται σε πολλών ειδών δίκτυα, είναι αυτόματο, καθώς δεν χρειάζεται απαραίτητα παραμέτρους από τον χρήστη, αλλά μπορεί να τις χρησιμοποιήσει αν το απαιτεί η εφαρμογή, μπορούμε να μεταφράσουμε τα αποτελέσματα του και μπορεί να λειτουργήσει και με εν εξέλιξη δίκτυα αντιμετωπίζοντάς τα ως streams.

Σε σχέση με τα προηγούμενα παρατηρούμε ότι και εδώ θα χρησιμοποιηθούν μετρικές του δικτύου, αλλά για τον εντοπισμό παγιωμένων συμπεριφορών. Θα χρησιμοποιηθεί το στοιχείο του χρόνου και της εξέλιξης. Τη μεγάλη διαφορά όμως αποτελεί η αντιμετώπιση των κοινωνικών φαινομένων ως “μαύρα κουτιά”, καθώς με το αυτόματο σύστημα εξαγωγής ρόλων, η υλοποίηση αυτή, εντοπίζει μόνη της μοτίβα (patterns), τα οποία αργότερα θα μεταφραστούν από τον χρήστη ως κοινωνικές συμπεριφορές. Επομένως δεν απαιτείται κωδικοποίηση των κοινωνικών φαινομένων σε μετρικές του δικτύου, αλλά τελικά η αποκωδικοποίηση των συμπερασμάτων σε κοινωνικά φαινόμενα.

Πριν προχωρήσομε σε αναλυτική περιγραφή του συστήματος θα δούμε κάποια πράγματα για την εξαγωγή ρόλων, που αποτελεί ίσως το σημαντικότερο χαρακτηριστικό των συστημάτων αυτών.

Έτσι λοιπόν στο κεφάλαιο 2 θα ορίσουμε πιο αναλυτικά τον ρόλο, θα δούμε τι σημαίνει ισοδυναμία κόμβων, θα διερευνήσουμε τρόπους αναπαράστασης ενός γράφου και τέλος πάρουμε μια γεύση της διαδικασίας εξαγωγής των ρόλων. Στο κεφάλαιο 3 θα δούμε σε βάθος τι σημαίνει εξαγωγή ρόλων από χαρακτηριστικά, θα παρουσιάσουμε μαθηματικά εργαλεία και θα σκιαγραφήσουμε ένα πλαίσιο γύρω από το οποίο λειτουργούν πολλές από τις μεθόδους εξαγωγής ρόλων. Στο Κεφάλαιο 4 θα εξετάσουμε βήμα βήμα δυο αλγορίθμους που πραγματοποιούν αυτή τη διαδικασία όταν δουλέψουν αθροιστικά και στο Κεφάλαιο 5 θα αναλύσουμε τα αποτελέσματα από δύο πρακτικές εφαρμογές των αλγορίθμων. Τέλος στο κεφάλαιο 6 θα προτείνουμε τρόπους βελτίωσης του πειράματος μας και θα αναφέρουμε κάποιες πρακτικές εφαρμογές των αλγορίθμων.

Κεφάλαιο 2 - Εξαγωγή ρόλων

2.1 Ορισμός έννοιας του ρόλου

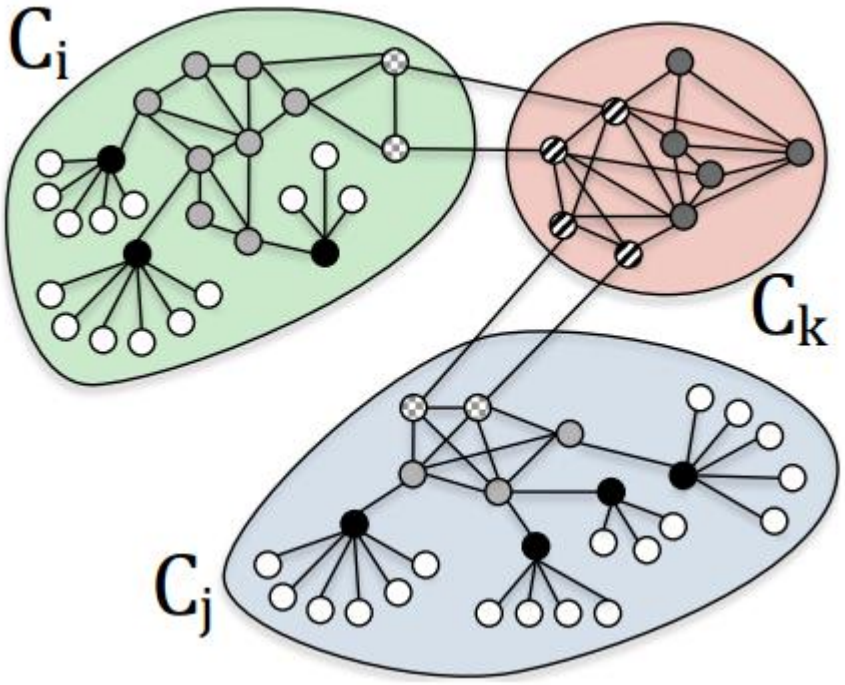
Η ιδέα του προσδιορισμού ρόλων προέρχεται από το πεδίο της κοινωνιολογίας, όπου οι ρόλοι χρησιμοποιούνται για να εκφράσουν μια συγκεκριμένη λειτουργία ενός μέλους σε μια κοινωνία (πχ δάσκαλος, παππούς, μέντορας). Καθώς οι ρόλοι αυτοί, εκφράζουν κοινωνικές λειτουργίες, εύλογα αποτελούν σημαντικό παράγοντα στην ανάλυση κοινωνικών δικτύων.

Ρόλους μπορούμε να ανακαλύψουμε όχι μόνο σε κοινωνικά δίκτυα, αλλά σε τεχνολογικά, βιολογικά, στο web και σε πολλά άλλα. Ενώ οι ρόλοι αποτελούν ακρογωνιαίο λίθο για graph mining αναλύσεις, μπορεί να έχουν και άλλες πρακτικές εφαρμογές, όπως ο εντοπισμός ανωμαλιών, η στόχευση του περιεχομένου για συγκεκριμένους ρόλους χρηστών, στην κατανομή καθηκόντων και τις εσωτερικές λειτουργίες μιας εταιρείας. Φαίνεται λογικό ότι σε εφαρμογές κατηγοριοποίησης, αυτόματης μάθησης και δειγματοληψίας, μπορούν να αποτελέσουν πολύ χρήσιμα εργαλεία στην κατεύθυνση αυτή.

Προκειμένου να συνεχίσουμε θα πρέπει να ορίσουμε την εξαγωγή ρόλων. Εξαγωγή ρόλων αποτελεί η διαδικασία διαχωρισμού των κόμβων ενός δικτύου σε ομάδες με ίδια κατασκευαστικά χαρακτηριστικά. Δύο κόμβοι θεωρούνται κατασκευαστικά όμοιοι, αν ενώνονται με το υπόλοιπο δίκτυο με τους ίδιους τρόπους. Στις διάφορες μελέτες που έχουν γίνει, προκειμένου να αποκτήσει πρακτική εφαρμογή ο παραπάνω ορισμός, η έννοια της ομοιότητας συνήθως χαλαρώνει. Χρησιμοποιώντας αυτή τη διαπίστωση, εξαγωγή ρόλων είναι ο διαχωρισμός των κόμβων σε ομάδες με παρόμοια κατασκευαστικά χαρακτηριστικά. Έτσι οι ρόλοι ενός δικτύου εκφράζουν πρότυπα σύνδεσης των κόμβων με το υπόλοιπο δίκτυο, όπως για παράδειγμα, κόμβοι/ακμές κέντρα ενός αστέρα, περιφερειακοί κόμβοι, σχεδόν κλίκες, γέφυρες που ενώνουν ξεχωριστά κομμάτια του γράφου. Ενώ με την πρώτη ματιά, οι ρόλοι φαίνονται να έχουν τοπική σημασία, στην

πραγματικότητα αποτελούν πολύπλοκα μοτίβα που εξαρτώνται από το πεδίο και τις διαδικασίες εφαρμογής τους.

Σε αυτό το σημείο, καλό είναι να διαχωρίσουμε την έννοια των ρόλων από αυτό των κοινοτήτων. Οι κοινότητες αποτελούν σύνολα κόμβων με περισσότερες ακμές μέσα στο σύνολο, παρά έξω από αυτό, ενώ οι ρόλοι είναι σύνολα κόμβων με περισσότερα κοινά κατασκευαστικά χαρακτηριστικά μέσα στο σύνολο παρά έξω από αυτό. Οι ρόλοι εκφράζουν κατασκευαστικές δομές, όπως γέφυρες, και κλίκες, ενώ οι κοινότητες εκφράζουν μεγέθη όπως η πυκνότητα και η εγγύτητα.



Τα σύνολα Ci, Cj, Ck αποτελούν κοινότητες, ενώ τα διαφορετικά χρώματα και μοτίβα που είναι ζωγραφισμένα στο εσωτερικό των κόμβων αναδεικνύουν ομάδες ρόλων

Στη συνέχεια θα αναφέρουμε διάφορες τεχνικές για τον υπολογισμό ρόλων, βασιζόμενοι σε χαρακτηριστικά του γράφου. Σε επόμενο κεφάλαιο θα καταλήξουμε στη μέθοδο που δοκιμάστηκε από εμάς, την οποία θα δούμε πιο αναλυτικά. Έχει αποδειχτεί από έρευνες, πως είναι αδύνατον μια μέθοδος εξαγωγής ρόλων, να λειτουργεί καλύτερα από όλες τις υπόλοιπες σε όλες τις περιπτώσεις, όπως και για οποιοδήποτε πρόβλημα βελτιστοποίησης. Παρακάτω θα κάνουμε μια αναδρομή και σύγκριση σε μεθόδους εξαγωγής ρόλων, με περισσότερη έμφαση στους ρόλους, που προέρχονται από αναπαράσταση του δικτύου μέσω χαρακτηριστικών και όχι απευθείας από τον αρχικό γράφο.

Μπορούμε να διακρίνουμε τρεις κατηγορίες εξαγωγής ρόλων, οι οποίες ακολούθησαν ιστορικά η μία την άλλη.

A. Συστήματα που παράγουν ρόλους απευθείας από την αναπαράσταση του γραφήματος.
B. Συστήματα που αναπαριστούν τον γράφο μέσω χαρακτηριστικών και στη συνέχεια από αυτά εξάγονται οι ρόλοι.
C. Ένας συνδυασμός αυτών των δύο. (υβριδική προσέγγιση)

Η μέθοδος αναπαράστασης με χαρακτηριστικά, περιλαμβάνει όλες τις μεθόδους που μπορούν να χρησιμοποιηθούν για την αναγωγή ενός γράφου σε ένα σύνολο χαρακτηριστικών. Τα χαρακτηριστικά αυτά υπολογίζονται από μετρικές των κόμβων ή των ακμών, καθώς και με την χρήση μη σχετικών με τον γράφο χαρακτηριστικών, τα οποία μπορεί να είναι δεδομένα από τη γνώση του προς μελέτη πεδίου του προβλήματος.

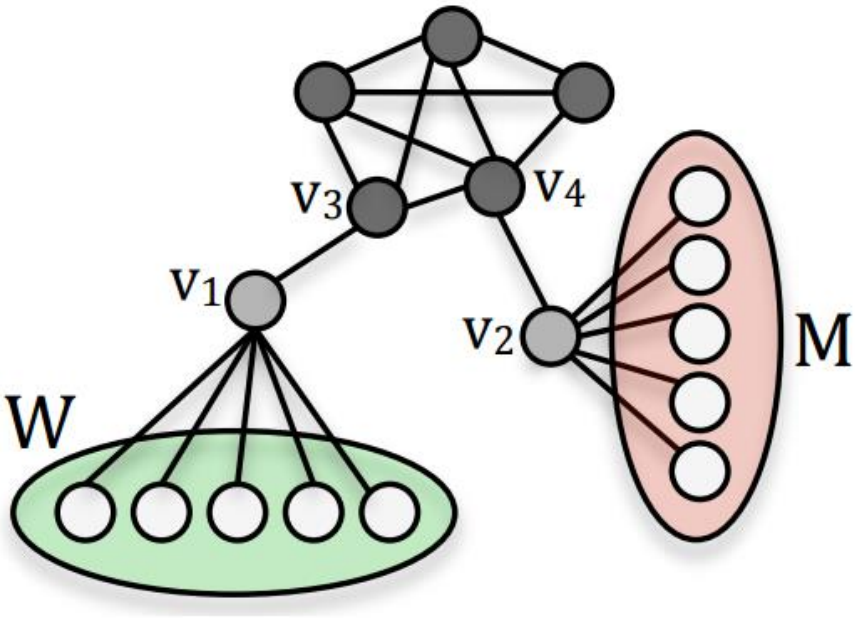
Προκειμένου ο αναγνώστης να κατανοήσει καλύτερα τον διαχωρισμό των μεθόδων, θα αναλύσουμε περεταίρω την ισοδυναμία των ρόλων. Όλοι οι κόμβοι που χαρακτηρίζονται από συγκεκριμένο ρόλο, θα πρέπει να είναι ισοδύναμοι, βάση μιας προκαθορισμένης σχέσης ισοδυναμίας. Το ερώτημα λοιπόν που προκύπτει απαιτεί τον αυστηρό καθορισμό μιας τέτοιας σχέσης. Ουσιαστικά οι τρεις μέθοδοι που αναφέρθηκαν παραπάνω αντιστοιχούν σε διαφορετικούς ορισμούς της σχέσης αυτής.

2.2 Ισοδυναμία κόμβων

2.2.1 Κατασκευαστική ισοδυναμία (structural equivalence)

Η κατασκευαστική ισοδυναμία ρόλων, η οποία αναφέρθηκε έμμεσα και πιο πάνω, δηλώνει ως ισοδύναμους κόμβους αυτούς που έχουν το ίδιο μοτίβο σύνδεσης με ακριβώς τους ίδιους γείτονες. Επομένως δύο ισοδύναμοι κόμβοι θα έχουν ακριβώς τους ίδιους γείτονες. Με τη λογική αυτή, τα μέλη του συνόλου W του παρακάτω σχήματος είναι κατασκευαστικά ισοδύναμα. Είναι προφανές ότι κατασκευαστικά ισοδύναμοι κόμβοι είναι δυσδιάκριτοι μεταξύ τους, καθώς έχουν τον ίδιο βαθμό (degree), τον ίδιο συντελεστή ομαδοποίησης (clustering coefficient), κεντρικότητα (centrality), ανήκουν στις ίδιες κλίκες και ούτω καθεξής. Επομένως αυτή η μορφή ισοδυναμίας είναι πολύ αυστηρή και δεν μπορεί να εφαρμοστεί σε μεγάλους, πραγματικούς γράφους. Μεγαλύτερο μειονέκτημα αποτελεί η σύγχυση της ισοδυναμίας και της ομοιότητας, με την εγγύτητα(closeness), που προκύπτει από την ιδιότητα των ακριβώς ίδιων γειτόνων, για κόμβους του ίδιου ρόλου. Στην πράξη, κόμβοι με την παραπάνω ιδιότητα δεν μπορούν ποτέ να απέχουν περισσότερα από δύο βήματα μεταξύ τους. Ως αποτέλεσμα

έχουν γίνει πολλές προσπάθειες χαλάρωσης της κατασκευαστικής ισοδυναμίας.



2.2.2 Αυτομορφική ισοδυναμία(automorphic equivalence)

Ο ισομορφισμός αποτελεί, μια αντιστοίχηση από έναν γράφο σε έναν άλλο, κατά την οποία διατηρείται η δομή του γράφου. Δηλαδή αν έχουμε έναν γράφο , όπου ισχύει u → v και εφαρμόσουμε τον ισομορφισμό p, τότε θα ισχύει ότι p(u) → p(v). Ο αυτομορφισμός είναι ένας ισομορφισμός από έναν γράφο στον ίδιο τον γράφο, οπότε διατηρούνται οι συμμετρίες. Ένας κόμβος u είναι αυτομορφικά ισοδύναμος με έναν κόμβο v, αν υπάρχει τέτοιος αυτομορφισμός ώστε u = p(v). Έτσι η αυτομορφική ισοδυναμία μπορεί να θεωρηθεί ως μια χαλάρωση της κατασκευαστικής ισοδυναμίας. Στην ουσία η κατασκευαστική ισοδυναμία εκφράζει αν δυο κόμβοι μπορούν να αλλάξουν θέση μεταξύ τους, ενώ διατηρούνται οι σύνδεσμοι τους, ενώ η αυτομορφική ισοδυναμία υποδεικνύει σύνολα κόμβων τα οποία μπορούν να αλλάξουν θέση μεταξύ τους ως υπο-γράφοι. Στο παραπάνω σχήμα τα σύνολα W και M μπορούν να αλλάξουν αμοιβαία θέση μεταξύ τους, οπότε τα μέλη τους εκφράζουν έναν ρόλο. Παρομοίως οι κόμβοι V_1 και V_2 αποτελούν έναν ακόμα ρόλο, αφού και αυτοί μπορούν να αλλάξουν θέση μεταξύ, από τη στιγμή, που το ίδιο συμβαίνει με τα W και M.

2.2.3 Τακτική ισοδυναμία(regular equivalence)

Με την τακτική ισοδυναμία, η έννοια του ρόλου χαλαρώνει ακόμα περισσότερο, ώστε να καλυφθεί ακριβέστερα η λειτουργικότητα του κοινωνικού ρόλου. Η ιδέα από πίσω, είναι ότι κόμβοι που συνδέονται με ισοδύναμους μεταξύ τους κόμβους, θα έχουν τον ίδιο ρόλο στην συμπεριφορά του δικτύου. Αποτελεί αρκετά διαφορετική προσέγγιση από την κατασκευαστική ισοδυναμία και οι κόμβοι θεωρούνται ότι έχουν παρόμοιες συμπεριφορές αν συνδέονται με παρόμοιους τρόπους σε κόμβους, με μεταξύ τους ίδιο ρόλο. Έτσι δεν υπάρχει δέσμευση ούτε για ίδιους γείτονες, ούτε για ίδιο αριθμό γειτόνων, αρκεί να συνδέονται με κόμβους που έχουν ισοδύναμους ρόλους μεταξύ τους. Υπό το πρίσμα αυτό, παρατηρούμε στο σχήμα, ότι τα σύνολα W και M είναι τακτικά ισοδύναμα και οι κόμβοι V_1 και V_2 σχηματίζουν έναν ακόμη ρόλο. Οι κόμβοι V_3 και V_4 απαρτίζουν μια κλάση ρόλων ακόμα και οι εναπομείναντες κόμβοι αποτελούν μέρος ενός τέταρτου ρόλου. Είναι για παράδειγμα εμφανές ότι στον τρίτο ρόλο, οι κόμβοι V_3 και V_4 έχουν τουλάχιστον μία ακμή σε κόμβους του δεύτερου ρόλου και είναι επίσης ενωμένοι με κόμβους του τέταρτου ρόλου.

Η τακτική ισοδυναμία μπορεί να είναι είτε ακριβής, είτε προσεγγιστική, οπότε ενδέχεται να υπάρξουν πολλοί σωστοί τρόποι ομαδοποίησης των κόμβων σε ρόλους για έναν δεδομένο γράφο. Ακόμα, συνεχίζει να υπάρχει αυστηρότητα στο γεγονός ότι ο ρόλος ενός κόμβου είναι συνυφασμένος με όλους τους κόμβους του ρόλου αυτού, παρά με ένα μέρος αυτών. Έτσι ακλούθησε ο ορισμός της στοχαστικής ισοδυναμία.

2.2.4 Στοχαστική ισοδυναμία(stochastic equivalence)

Στη στοχαστική ισοδυναμία, κόμβοι διαμορφώνουν ομάδες ρόλων, αν έχουν την ίδια πιθανότητα κατανομής, εμφάνισης ακμών με άλλους κόμβους, για μια δεδομένη κατανομή πιθανότητας των ακμών του γράφου. Πιο απλά, οι κόμβοι διαχωρίζονται σε ρόλους, ώστε η πιθανότητα ενός κόμβου να συνδέεται με όλους τους άλλους κόμβους του γράφου, να είναι η ίδια για τα μέλη του κάθε ρόλου. Μια ακόμα έκφανση αυτού, αποτελεί ότι η πιθανότητα κατανομής του γράφου πρέπει να μένει ίδια, αν αλλάξουν αμοιβαία οι θέσεις δυο κόμβων του γράφου, που ανήκουν στον ίδιο ρόλο. Λόγω της απλότητας του προηγούμενου σχήματος, οι ρόλοι που θα διακρίναμε, είναι ίδιοι με αυτούς που προέκυψαν από την τακτική ισοδυναμία. Με αφορμή τον παραπάνω ορισμό, προέκυψαν τα stochastic blockmodels, με τα οποία χαλαρώνει και επεκτείνεται περισσότερο η κατασκευαστική ισοδυναμία των κόμβων.

Στην υποενότητα που ακολουθεί αναλύουμε τα blockmodels και κάνουμε αναφορά στις μεθόδους εξαγωγής ρόλων απευθείας από την αναπαράσταση του γράφου του δικτύου.

2.3 Εξαγωγή ρόλων απευθείας από την αναπαράσταση του γράφου.

Όπως αναφέρθηκε και προηγουμένως, οι μέθοδοι αυτής της κατηγορίας, έχουν ως είσοδο απευθείας την αναπαράσταση του γράφου, χωρίς κάποια επεξεργασία στο ενδιάμεσο.

2.3.1 Blockmodels

Τα blockmodels αναπαριστούν το δίκτυο μέσω ενός γραφήματος διάδρασης ρόλων (role – interaction graph ή διαφορετικά image matrix), όπου οι κόμβοι του νέου αυτού γραφήματος αντιπροσωπεύουν ρόλους (ή blocks ή θέσεις) και οι ακμές δραστηριότητες μεταξύ αυτών. Αυτή η αρκετά μικρότερη οντότητα μπορεί εκφράσει καλύτερα τα φαινόμενα που διέπουν το αρχικό δίκτυο. Οι μέθοδοι γύρω από τα blockmodels, κατηγοριοποιούν τους κόμβους, με βάση, είτε την κατασκευαστική, είτε την στοχαστική ισοδυναμία.

Μία από τις μεθόδους που χρησιμοποιούν blockmodels είναι η CONCOR (convergence of iterated correlations). Υπολογίζει την αυτοσυσχέτιση του πίνακα γειτνίασης ενός γράφου Corr(A, A) και ονομάζει τον νέο πίνακα C_0. Στη συνέχεια ο πίνακα συσχέτισης C_1 υπολογίζεται από την αυτοσυσχέτιση του C_0. Επαναλαμβάνεται αυτή η διαδικασία, μέχρις ότου όλες οι τιμές του τελευταίου πίνακα να είναι είτε 1 είτε -1. Έτσι έχουμε χωρίσει τους κόμβους σε δύο μπλοκ. Επαναλαμβάνουμε αυτή τη διαδικασία για τα δύο μπλοκ, ανάλογα με το πόσους ρόλους θέλουμε να εξάγουμε. Παρατηρούμε ότι το κριτήριο αυτό βασίζεται στην κατασκευαστική ισοδυναμία και αποτελεί μια αρκετά απλοϊκή μέθοδο.

Τα stochastic blockmodels από την άλλη υλοποιούν την ιδέα της στοχαστικής ισοδυναμίας. Το πλεονέκτημα του μοντέλου, είναι ότι επιτρέπει την παρέκκλιση από τις αρχικές παρατηρήσεις και χαλαρώνει το κριτήριο της ιδανικής ισοδυναμίας. Αρχικά καθορίζεται πόσοι ρόλοι θα υπάρξουν, Υπάρχει μια πιθανότητα, που αντιπροσωπεύει την εμφάνιση ακμής μεταξύ δυο κόμβων που ανήκουν σε διαφορετικούς ρόλους ή blocks, ή καλύτερα ένα εύρος πιθανοτήτων. Αν k το σύνολο των κόμβων, έχουμε έναν πίνακα $k_x k$, που εκφράζει αυτές τις πιθανότητες, και είτε παρέχεται από τον χρήστη, είτε υπολογίζεται από τα δεδομένα. Τα μοντέλα αυτά μπορούν να χρησιμοποιηθούν για την παραγωγή τυχαίων γράφων με βάση αυτές τις πιθανότητες. Δίνεται επίσης η δυνατότητα συμμετοχής των κόμβων σε πολλαπλούς ρόλους με περεταίρω επεξεργασία των αποτελεσμάτων, χρησιμοποιώντας δηλαδή μεταδεδομένα.

2.3.2 Ομοιότητα γραμμής/στήλης του πίνακα γειτνίασης.

Αν και τα πιο διάσημα μοντέλα της κατηγορίας που εξετάζουμε είναι τα blockmodels, υπάρχουν και άλλες τεχνικές που υπολογίζουν ομοιότητα μεταξύ των γραμμών του πίνακα γειτνίασης. Διακρίνουμε δύο γενικά βήματα σε όλες αυτές τις μεθόδους.

A. Η ομοιότητα, ή η διαφορά, υπολογίζεται μεταξύ όλων των δυνατών ζευγαριών από γραμμές του πίνακα. Μετρικές που αναδεικνύουν αυτή την ομοιότητα αποτελούν η ευκλείδεια απόσταση, ή και η συσχέτιση.
B. Γίνεται διαχωρισμός των κόμβων με βάση τις τιμές του πίνακα που προέκυψαν από το πρώτο βήμα.

Συνηθισμένες μέθοδοι για το δεύτερο βήμα αποτελούν οι hierarchical clustering και multi-dimensional scaling.

Υπάρχουν επίσης και τεχνικές, οι οποίες υπολογίζουν τους ιδιοπίνακες του πίνακα γειτνίασης και χρησιμοποιούν ένα υποσύνολο από αυτούς, ώστε να εξάγουν ρόλους. Οι ιδιοπίνακες που έχουν ενδιαφέρον είναι όχι μόνο αυτοί με τις μεγαλύτερες ιδιοτιμές, αλλά και εκείνοι που αντιπροσωπεύουν κατασκευαστικά μοτίβα, όπως κέντρα αστέρων, ακμές αστέρων, κλίκες και τα λοιπά. Τις περισσότερες φορές, οι ιδιοπίνακες, που προέρχονται από τις μεγαλύτερες ιδιοτιμές αντιπροσωπεύουν κάποιον συγκεκριμένο ρόλο (συνήθως σχεδόν κλίκες).

Το μεγαλύτερο μειονέκτημα των μεθόδων, που βασίζονται σε blockmodels, είναι η δυσκολία της εφαρμογής τους σε μεγάλους και πολύπλοκους γράφους, όπως είναι το γράφημα φίλων του Facebook. Για παράδειγμα η εφαρμογή της μεθόδου MMSB μπορεί να διαρκέσει και μία ολόκληρη μέρα για τον υπολογισμό χιλίων κόμβων, καθώς έχει πολυπλοκότητα O(n^4), όπου n ο αριθμός των κόμβων. Οι μέθοδοι ομοιότητας γραμμών, αν και η εκτέλεσή τους είναι πολύ λιγότερο απαιτητική σε χρόνο, έχουν το μειονέκτημα, πως οι ρόλοι που θα προκύψουν θα έχουν λιγότερο νόημα και θα είναι δύσκολο να μεταφραστούν σε κοινωνικά φαινόμενα. Επίσης η ακρίβεια του αποτελέσματος εξαρτάται σε μεγάλο βαθμό από την εφαρμογή και το πεδίο στο οποίο ανήκει το δίκτυο.

2.4 Εξαγωγή ρόλων μέσω εξαγωγής χαρακτηριστικών

2.4.1 Εισαγωγή

Από την στιγμή που η μέθοδος που εφαρμόσαμε ανήκει στην κατηγορία αυτή, θα είμαστε περισσότερο αναλυτικοί. Θα ορίσουμε τις μεθόδους αυτές με λεπτομέρεια και θα τις κατηγοριοποιήσουμε. Θα γίνει επίσης σύνδεση και επέκταση στους ορισμούς της ισοδυναμίας ρόλων που αναφέραμε παραπάνω.

Οι ρόλοι που προήλθαν από χαρακτηριστικά, προκύπτουν ύστερα από μετατροπή της αναπαράστασης του γράφου σε μια αναπαράσταση από χαρακτηριστικά. Έτσι η ισοδυναμία κόμβων, μετατρέπεται σε ισοδυναμία χαρακτηριστικών. Βλέπουμε, ότι η προσέγγιση αυτή, είναι πολύ διαφορετική από την προηγούμενη, η οποία εξήγαγε ρόλους, απευθείας από την αναπαράσταση του γράφου. Το σύνολο των χαρακτηριστικών συνήθως προκύπτει από συναρτήσεις – αντιστοιχίσεις, που εφαρμόζονται στον γράφο. Έχουμε δηλαδή ότι $f(G) = X$, όπου G είναι ο προς μελέτη γράφος, f ένα σύνολο μετατροπών που θα εφαρμόσουμε σε αυτόν και X τα χαρακτηριστικά των κόμβων ,που θα προκύψουν. Σε περίπτωση που γνωρίζουμε ήδη κάποια χαρακτηριστικά του γράφου, μπορούμε να χρησιμοποιήσουμε και αυτά ως είσοδο και η παραπάνω σχέση γίνεται $f(G, X_{in}) = X_{out}$.

Στη συνέχεια, θα πρέπει να μετακινηθούμε από την ισοδυναμία κόμβων, της αναπαράσταση με γράφο, προς την ισοδυναμία κόμβων της αναπαράστασης με χαρακτηριστικά.

Οι μέθοδοι εξαγωγής ρόλων βασισμένες σε χαρακτηριστικά, θα πρέπει να αντιστοιχούν σε κάποιους από τους ορισμούς ισοδυναμίας κόμβων, που ορίστηκαν παραπάνω. Θα μπορούσε επίσης να υπάρξει επέκταση των ορισμών αυτών ώστε να ταιριάξουν στο πεδίο της νέας αναπαράστασης με χαρακτηριστικά. Επιπρόσθετα δύναται να χαλαρώσουμε περισσότερο τους ορισμούς αυτούς, ώστε να επιτρέψουμε μεγαλύτερη ευελιξία και αποτελεσματικότητα. Γενικά παρακάτω ακολουθείται η δεύτερη προσέγγιση.

2.4.2 Ορισμός ισοδυναμίας κόμβων σε αναπαράσταση με χαρακτηριστικά.

Αν f_1, f_2,...,f_m, ένα σύνολο κατασκευαστικών χαρακτηριστικών (βαθμός, απόσταση, κτλ) και u και v δυο τυχαίοι κόμβοι, τότε ένας αυστηρός ορισμός της ισοδυναμίας των κόμβων θα ήταν:

$$Αν\ \forall i, όπου\ 1 \leq i\ \leq m, ισχυεί\ fi(u) = fi(v), τότε\ u \equiv v$$

Ο ορισμός αυτός είναι αυστηρός, υπό το πρίσμα, ότι δυο κόμβοι είναι ισοδύναμοι, αν και μόνο αν παρουσιάζουν ταυτόσημους πίνακες χαρακτηριστικών. Όπως και στους προηγούμενους ορισμούς ισοδυναμίας,

έτσι και εδώ μπορούμε να χαλαρώσουμε το κριτήριο ισοδυναμίας, κρατώντας τις ιδιότητες που μας ενδιαφέρουν. Θα μεταβούμε λοιπόν από τον αυστηρό ορισμό της ισοδυναμίας κόμβων, στον πιο χαλαρό της ομοιότητας κόμβων σε μια αναπαράσταση χαρακτηριστικών. Σύμφωνα με αυτό, δύο κόμβοι θα έχουν τον ίδιο ρόλο, αν έχουν παρόμοιες τιμές χαρακτηριστικών.

Επομένως: Δύο κόμβοι u και v θα είναι όμοιοι αν $S(x_u, x_v) \cong 1$, όπου S συνάρτηση μέτρησης της ομοιότητας και τα x_u και x_v οι πίνακες των τιμών των χαρακτηριστικών των δύο κόμβων. Η συνάρτηση ομοιότητας θα πάρει την τιμή 1 (μέγιστη) σε περίπτωση που τα x_u και x_v είναι ταυτόσημα.

Έτσι αν τα χαρακτηριστικά είναι αυστηρώς χαρακτηριστικά του γράφου και αντιπροσωπευτικά των κατασκευαστικών του ιδιοτήτων, τότε οι κόμβοι u και v θα είναι κατασκευαστικά όμοιοι. Εδώ να αναφέρουμε, ότι μαθηματικά εργαλεία που θα αναφέρουμε παρακάτω όπως, low-rank approximation και hierarchical clustering, χρησιμοποιούν μια συνάρτηση ομοιότητας, όπως η ευκλείδεια απόσταση, ή η Frobernius Norm. Να επισημάνουμε επίσης, ότι ο παραπάνω ορισμός είναι ανεξάρτητος από τους γείτονες των κόμβων και βασίζεται αποκλειστικά στα χαρακτηριστικά. Η ανεξαρτησία αυτή, επιτρέπει και στους ρόλους να μην εξαρτώνται ο ένας από τον άλλον, όπως συνέβαινε στους ορισμούς της ισοδυναμίας της προηγούμενης υποενότητας. Επιπρόσθετα, ο ορισμός βασίζεται σε μη αυστηρές συγκρίσεις με αποτέλεσμα τη χαλάρωση του προηγούμενου αυστηρού ορισμού, καθιστώντας το χρήσιμο για πρακτική εφαρμογή και την έκφραση μιας μεγαλύτερης ποικιλίας ρόλων, οι οποίοι μπορούν να εξηγηθούν ευκολότερα. Αυτό οφείλεται και στην δυνατότητα κατασκευής των ρόλων από ένα πιθανώς άπειρο σύνολο χαρακτηριστικών.

2.4.3 Παραδείγματα χρήσης

Ακολουθούν μερικά παραδείγματα εφαρμογής της μεθόδου των χαρακτηριστικών.

Η πρώτη εφαρμογή που προτάθηκε και βασιζόταν στις παραπάνω ιδέες ήταν από τους Batagelj, Ferilgoj και Doreian (Direct and indirect methods for structural equivalence). Στη μελέτη ορίζεται η σχέση μεταξύ ενός συνόλου κατασκευαστικών χαρακτηριστικών και κατασκευαστικής ισοδυναμίας και δίνεται έμφαση στην ανάγκη σταθερότητας της σχέσης αυτής. Η έμμεση αυτή μέθοδος ορισμού της ισοδυναμίας εφαρμόστηκε στα blockmodels, που εξετάσαμε παραπάνω, τα οποία σχεδιάζονταν με τέτοιο τρόπο ώστε να αντιστοιχούν ακριβώς σε κάποιον από τους αρχικούς ορισμούς της ισοδυναμίας. Άλλες έρευνες χρησιμοποίησαν έναν πιο χαλαρό ορισμού του ρόλου, βασισμένο σε χαρακτηριστικά, ο οποίος δεν ήταν συνυφασμένος με την έννοια της ισοδυναμίας της προηγούμενης ενότητας. Οι προσεγγίσεις αυτές δημιουργούν έναν μεγάλο πίνακα με τοπικά χαρακτηριστικά, ειδικά επιλεγμένα για κοινωνικά δίκτυα και στη συνέχεια χρησιμοποιούν NMF (non-negative matrix factorization), ώστε να χωρίσουν τους κόμβους σε

ρόλους. Οι ίδια τεχνική έχει χρησιμοποιηθεί κα για την διερεύνηση δυναμικών – εν εξελίξει δικτύων, αλλά και για βελτίωση του προβλήματος της ταξινόμησης των κόμβων.

Μέχρι τώρα παρατηρήσαμε ότι οι παραδοσιακές προσεγγίσεις με την αναπαράσταση του γραφήματος δεν μπορούν να εντοπίσουν πολύπλοκους ρόλους σε μεγάλα δίκτυα λόγω των περιορισμών τους. Αναφέραμε την ικανότητα των μεθόδων βασισμένων σε αναπαράσταση με χαρακτηριστικά, να προσαρμοστούν σε τέτοιους ρόλους, οι οποίοι αντιστοιχούν περισσότερο στα πραγματικά δίκτυα και παρέχουν την ευελιξία εφαρμογής τους στην πράξη. Παρόλα αυτά, γίνεται αρκετά δυσκολότερος ο προσδιορισμός των χαρακτηριστικών που θα ληφθούν υπόψη, για τον εντοπισμό των ρόλων, που απαιτεί η κάθε εφαρμογή, ή έρευνα. Στο πρόβλημα αυτό, απαιτείται καθοδήγηση από κάποιον ειδικό στο πεδίο προς μελέτη.

2.5 Υβριδικές προσεγγίσεις

Όπως λογικά καταλαβαίνει κάποιος, οι υβριδικές προσεγγίσεις χρησιμοποιούν στοιχεία και των δύο προηγούμενων ομάδων από μεθόδους. Αυτό που διαχωρίζει τις υβριδικές μεθόδους μεταξύ τους είναι αν η αναπαράσταση του γραφήματος χρησιμοποιήθηκε πριν ή μετά την κατασκευή ρόλων από χαρακτηριστικά.

Η πρώτη κατηγορία χρησιμοποιεί μεθόδους εύρεσης ρόλων πρώτα στην αναπαράσταση του γράφου και ύστερα πραγματοποιείται εξαγωγή τους από χαρακτηριστικά. Για παράδειγμα μπορούν να χρησιμοποιηθούν blockmodels για την εξαγωγή ρόλων απευθείας από τον γράφο και στη συνέχεια να τους χρησιμοποιήσουμε ως αρχικά χαρακτηριστικά (είσοδο) στην μέθοδο των χαρακτηριστικών, έτσι ώστε να παραχθούν πιο χρήσιμα αποτελέσματα. Το πλεονέκτημα της μεθόδου είναι, ότι μπορεί να υπάρξει ένας παραμετρικός υπολογισμός χαρακτηριστικών, ενώ το μειονέκτημα, ότι λόγω των περιορισμών των blockmodels δεν μπορεί να επεκταθεί σε πολύ μεγάλους γράφους

Η δεύτερη κατηγορία χρησιμοποιεί ουσιαστικά πολλές ροές δεδομένων ώστε να επηρεάσει τη διαδικασία διαχωρισμού των ρόλων. Υπήρξε δηλαδή μία αναπαράσταση με χαρακτηριστικά, αλλά, καθώς εμφανίστηκαν άλλες πηγές δεδομένων, όπως γράφοι και νέα σύνολα χαρακτηριστικών, θέλουμε να χρησιμοποιήσουμε και αυτές για τον καθορισμό των ρόλων. Με γνώμονα τεχνικές σαν και αυτή, μια πιθανή εξέλιξη του ερευνητικού πεδίου θα μπορούσε να είναι ο καθορισμός δυναμικών ρόλων.

Κεφάλαιο 3 - Πλαίσιο εξαγωγής ρόλων από χαρακτηριστικά

3.1 Ορισμός πλαισίου

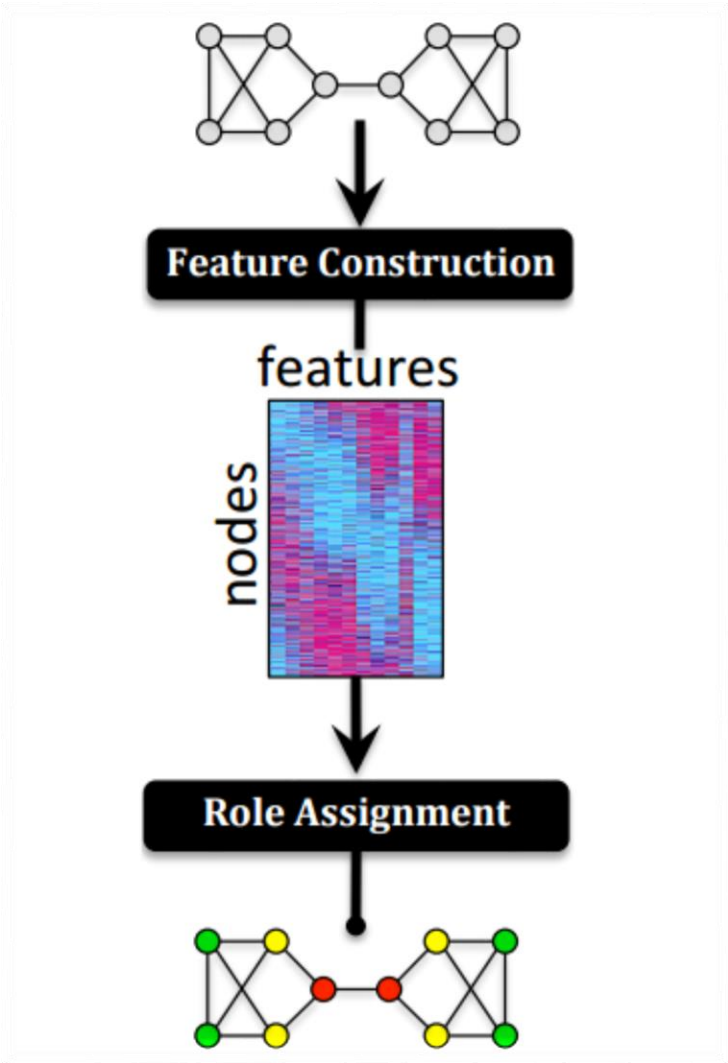
Στη συνέχεια θα παρουσιάσουμε ένα πλαίσιο πάνω στο οποίο μπορεί να γίνει εξαγωγή ρόλων βασισμένη σε χαρακτηριστικά. Η υλοποίηση που θα δοκιμάσουμε στο τελευταίο μέρος της εργασίας ακολουθεί το πλαίσιο αυτό και φαίνεται λογικό να παρουσιαστεί και αυτό ως μέρος της διπλωματικής.

Το πλαίσιο λοιπόν που θα παρουσιάσουμε αποτελείται από δύο διακριτά τμήματα:

A. Μετατροπή του γράφου σε ένα σύνολο χαρακτηριστικών των κόμβων.
B. Διαχωρισμός των κόμβων σε ρόλους με βάση τον πίνακα χαρακτηριστικών του κάθε κόμβου.

Ο γράφος λοιπόν αρχικά θα μετατραπεί σε μια αναπαράσταση με βάση τα χαρακτηριστικά του κάθε κόμβου και στη συνέχεια θα ανατεθούν ρόλοι στους κόμβους, μέσω κάποιας μεθόδου παραγοντοποίησης πινάκων ή αλγορίθμου clustering.

Στην παρακάτω εικόνα γίνονται πιο εμφανής οι ιδέες του πλαισίου που θα περιγράψουμε:



Παραπάνω φαίνεται διαδικασία που ακολουθεί το πλαίσιο. Το αρχικό γράφημα αναπαρίσταται από έναν πίνακα χαρακτηριστικών των κόμβων και στη συνέχεια με τη χρήση του πίνακα αυτού, πραγματοποιείται ανάθεση ρόλων

Αξίζει να αναφερθεί ότι μπορούν να χρησιμοποιηθούν διάφορα συστήματα εκμάθησης χαρακτηριστικών στο πρώτο βήμα. Τα συστήματα μπορεί να ψάχνουν για αυστηρώς τοπικά χαρακτηριστικά (πχ degree) , ή πιο γενικά (πχ centralities). Μπορούν να προέρχονται από τους άμεσους γείτονες (egonet) ή από τους γείτονες σε περισσότερα από ένα βήματα απόσταση. Η επιλογή εξαρτάται από το πεδίο εφαρμογής, τους περιορισμούς σε υπολογιστική ισχύ και τις επιμέρους ιδιαιτερότητες της κάθε εφαρμογής. Το ίδιο μπορεί να συμβεί και για την ανάθεση των ρόλων, καθώς κάποιος μπορεί να επιλέξει NMF, είτε SVD για την παραγοντοποίηση.

Το κύριο χαρακτηριστικό που προσπαθεί να εκμεταλλευτεί η μέθοδος αυτή είναι η ευελιξία. Για τον λόγο αυτό η μέθοδος μπορεί να εφαρμοστεί και στον εντοπισμό ρόλων για τις ακμές του γράφου. Επιπλέον οι διάφοροι περιορισμοί του προβλήματος, ανάλογα την εφαρμογή, όπως sparseness, diversity και locality μπορούν να ενσωματωθούν στη διαδικασία, είτε με την

επιλογή κατάλληλων χαρακτηριστικών, είτε με την επιλογή καταλληλότερων ρόλων.

Παρουσιάζουμε συνοπτικά τα πλεονεκτήματα του πλαισίου αυτού με βάση τη μέχρι τώρα ανάλυση:

- Ευελιξία των ρόλων
- Οι ρόλοι μπορούν να καθοριστούν ξεχωριστά για κάθε εφαρμογ.ή
- Η πολυπλοκότητα και η αποτελεσματικότητα μπορούν να ισορροπήσουν ανάλογα τους περιορισμούς.
- Μπορούν να εντοπιστούν τυχαία μοτίβα βασιζόμενα σε μεταδεδομένα.
- Αρχικά χαρακτηριστικά του δικτύου μπορούν εύκολα να εισαχθούν στον υπολογισμό, ως επιπλέον χαρακτηριστικά, είτε πριν, είτε μετά την εκμάθηση των υπόλοιπων χαρακτηριστικών.

3.2 Κατασκευή χαρακτηριστικών για απόδοση ρόλων

Συνεχίζουμε, αναλύοντας περισσότερο το πρώτο μέρος του πλαισίου. Σκοπός του βήματος αποτελεί η συστηματική κατασκευή χαρακτηριστικών βασιζόμενοι, είτε στην κατασκευή του γράφου, είτε σε άλλα γνωστά χαρακτηριστικά. Υποθέτουμε την ύπαρξη ενός γράφου G, οι κόμβοι του οποίου χαρακτηρίζονται από κάποιες ιδιότητες X_n και οι ακμές του από κάποιες άλλες ιδιότητες X_e. Ένα παράδειγμα μιας ιδιότητας είναι η ηλικία του κόμβου – ατόμου, ή το φύλο του (αρσενικό η θηλυκό). Τελικά, μέσω της διαδικασίας θα προσθέσουμε επιπλέον χαρακτηριστικά.

Από το πεδίο του machine learning, μαθαίνουμε ότι η κατασκευή χαρακτηριστικών αποσκοπεί στον εντοπισμό ιδιοτήτων μεγάλης συσχέτισης με τη μεταβλητή, που προσπαθούμε να προβλέψουμε και είναι ιδανικά, ασυσχέτιστες μεταξύ τους. Μεταφέροντας τον ορισμό αυτό στα δίκτυα, ορίζουμε ως στόχο της κατασκευής χαρακτηριστικών, την παραγωγή ενός συνόλου χαρακτηριστικών που αντικατοπτρίζουν τις βασικές κατασκευαστικές δομές του δικτύου και περιγράφουν επαρκώς τα μοτίβα που παρατηρούνται σε αυτό. Μπορούμε να επεκτείνουμε τον ορισμό, ώστε να καλύπτει και φαινόμενα, που αφορούν μια συγκεκριμένη εφαρμογή ή πεδίο. Για παράδειγμα, διαφορετικές ιδιότητες μας ενδιαφέρουν στα κοινωνικά δίκτυα και διαφορετικές στα τεχνολογικά, οπότε τελικώς η επιλογή τους, εξαρτάται σε μεγάλο βαθμό από τους στόχους της έρευνάς μας. Στη συνέχεια λοιπόν θα αναφέρουμε τεχνικές κατασκευής χαρακτηριστικών που μπορούν να εφαρμοστούν σε μεγάλη γκάμα εφαρμογών.

3.2.1 Βήματα εκμάθησης χαρακτηριστικών

Διακρίνουμε πέντε βήματα – επιλογές που πρέπει να ακολουθηθούν για την εκμάθηση μιας αναπαράστασης χαρακτηριστικών.

A. Τύποι χαρακτηριστικών. Επιλογή του είδους των χαρακτηριστικών που θα χρειαστούμε.
B. Τελεστές χαρακτηριστικών. Επιλογή των τελεστών, που θα εφαρμόσουμε σε αυτούς τους τύπους χαρακτηριστικών.
C. Στρατηγική διερεύνησης των χαρακτηριστικών. Δηλαδή στρατηγική έρευνας και εντοπισμού στο πεδίο των πιθανών χαρακτηριστικών, η οποία μπορεί να είναι εξαντλητική, τυχαία, ή καθοδηγούμενη
D. Επιλογή – αξιολόγηση των τελικών χαρακτηριστικών. Αποφασίζουμε, πως θα αξιολογηθούν τα διάφορα χαρακτηριστικά και την διαδικασία, με την οποία θα παραλείψουμε ορισμένα από αυτά

Επιλογή τύπων χαρακτηριστικών.

Διακρίνουμε τέσσερεις κύριους τύπους χαρακτηριστικών. Κατασκευαστικά χαρακτηριστικά, χαρακτηριστικά βάση των τιμών των ακμών, χαρακτηριστικά βάση των τιμών των κόμβων και μη σχεσιακά χαρακτηριστικά, τα οποία δεν βασίζονται σε πληροφορίες συνδέσμων, αλλά μόνο αυτούσιες τιμές χαρακτηριστικών των κόμβων.

Τα κατασκευαστικά χαρακτηριστικά μπορούν να υπολογιστούν χρησιμοποιώντας μόνο τον αρχικό γράφο και όχι άλλα χαρακτηριστικά. Τέτοια παραδείγματα είναι: degree, clustering coefficient και betweeness. Θα μπορούσε να χρησιμοποιηθεί επίσης οποιοδήποτε μοτίβο του γράφου, για παράδειγμα τρίγωνα, αστέρες, ή χαρακτηριστικά βασιζόμενα σε μονοπάτια και αποστάσεις. Είναι προφανές ότι μπορεί να παραχθεί ένας τεράστιος αριθμός τέτοιων χαρακτηριστικών.

Τα χαρακτηριστικά βάση τιμών των ακμών υπολογίζονται, χρησιμοποιώντας μόνο τις τιμές των χαρακτηριστικών των ακμών, που έχει ένας κόμβος. Η διαδικασία αυτή μπορεί να γενικευτεί και για ακμές σε βήματα μήκος δύο, τρία κτλ. Για παράδειγμα, αν έχουμε τις τιμές των χαρακτηριστικών των ακμών, είτε από πριν, είτε τα υπολογίσαμε κατασκευαστικά, μπορούμε να εφαρμόσουμε σε αυτά κάποιον μέσο όρο ή κάποιο άθροισμα, ώστε να εκμαιεύσουμε ένα νέο χαρακτηριστικό. Όπως και πριν, με τη μέθοδο αυτή μπορούμε να υπολογίσουμε χαρακτηριστικά, είτε κόμβων, είτε ακμών.

Αντίστοιχα μπορούν να υπολογιστούν χαρακτηριστικά βάση των τιμών των χαρακτηριστικών σε γειτονικούς κόμβους, ή σε κόμβους που απέχουν ένα, δύο, ν βήματα από τον κόμβο που εξετάζουμε. Αν θέλουμε, ας πούμε, να υπολογίσουμε την πολιτική επιρροή ενός προσώπου στο Facebook, μπορούμε να το κάνουμε υπολογίζοντας την πολιτική επιρροή των γειτονικών του κόμβων. Λαμβάνοντας υπόψη κόμβους σε μεγαλύτερη απόσταση μπορούμε να δώσουμε σε αυτούς μικρότερο βάρος. Η τεχνική αυτή

θυμίζει αρκετά την προσέγγιση της εμπιστοσύνης που συναντήσαμε στο εισαγωγικό κεφάλαιο.

Τα μη σχεσιακά χαρακτηριστικά υπολογίζονται από τις τιμές των ήδη υπαρχόντων χαρακτηριστικών, που έχει ο κόμβος, και συνήθως είναι τα τελευταία που υπολογίζονται. Τέτοια χαρακτηριστικά, αποτελούν οι μέσοι όροι ή τα αθροίσματα μερικών χαρακτηριστικών του κόμβου. Πιο πρακτικά, αν έχουμε για ένα χρήση του Facebook τον μέσο όρο που βρίσκεται online και τον μέσο αριθμό posts την ημέρα, βγάζει νόημα να προσθέσουμε αυτά τα δύο, ώστε να φτιάξουμε ένα νέο χαρακτηριστικό. Τα χαρακτηριστικά αυτής της κατηγορίας συνήθως χρησιμοποιούνται για την ρύθμιση του συστήματος εξαγωγής ρόλων για μια συγκεκριμένη εφαρμογή.

Επιλογή τελεστών

Οι τελεστές καθορίζουν τον χώρο, στον οποίο θα ψάξουμε για χαρακτηριστικά, για την εξαγωγή ρόλων. Παίρνουν δηλαδή ως είσοδο τους τύπους χαρακτηριστικών που επιλέξαμε πριν και παράγουν νέα. Οι κύριοι τύποι φαίνονται στον παρακάτω πίνακα, μαζί με κάποια παραδείγματα. Είναι φανερό, πως μπορούν να υπολογιστούν άπειρα νέα χαρακτηριστικά με βάση τους τελεστές.

Operators	Examples
Rel. aggr.	MODE, MEAN, COUNT, ...
Set ops	Union, multiset, inters., ...
Subgraph pat.	k-star, k-clique, k-motif, ...
Dim. redu.	SVD, PMF, NMF, ICA, PCA, ...
Similarity	Cosine sim, mutual info, ...
Paths/walks	random-walks, k-walks, ...
Text analy.	LDA, Link-LDA/PLSA, ...

Αρκετοί από τους τελεστές μπορούν να χρησιμοποιηθούν για την εξαγωγή κατασκευαστικών χαρακτηριστικών πάνω σε κόμβους και πάνω σε ακμές. Κάποιοι άλλοι βασίζονται σε μη σχεσιακή πληροφορία όπως για παράδειγμα η ανάλυση κειμένου και δεν μπορούν να χρησιμοποιηθούν για κατασκευαστικά χαρακτηριστικά. Επιπρόσθετα πολλοί από τους τελεστές μπορούν να εφαρμοστούν αναδρομικά, όπως τα αθροίσματα και οι μέσοι όροι, το clustering και η μείωση των διαστάσεων.

Προκειμένου να παραχθεί ένα σύνολο χαρακτηριστικών, που να έχει νόημα, καλό θα ήταν, να επιλεγεί ένας μικρός αριθμός τελεστών, βασισμένος σε γνώση του πεδίου, ή σε υποθέσεις που θέλουμε να εξετάσουμε.

Επιλογή στρατηγικής

Ύστερα από την επιλογή τύπου χαρακτηριστικών και τους τελεστές που θα εφαρμόσουμε, πρέπει να διαλέξουμε και μια στρατηγική διερεύνησης τους, που θα υποδείξει τελικά, ποια από αυτά τα χαρακτηριστικά θα διαλέξουμε. Μια εξαντλητική στρατηγική θα επιλέξει όλα τα δυνατά χαρακτηριστικά, βάσει τύπων και τελεστών, ενώ μια τυχαία στρατηγική θα κρατήσει ένα μέρος τους, στηριζόμενη σε κάποια μέθοδο δειγματοληψίας. Η καθοδηγούμενη στρατηγική, αποφασίζει με βάση κάποια ευρεστική τεχνική, που υποδεικνύει τα πιο χρήσιμα χαρακτηριστικά. Και στις τρεις περιπτώσεις, κάθε χαρακτηριστικό υπόκειται σε μια αξιολόγηση, η οποία καθορίζει τη χρησιμότητα του στον καθορισμό ρόλων. Καθώς οι ρόλοι εξαρτώνται σε μεγάλο βαθμό από το πεδίο και την εκάστοτε εφαρμογή, θα πρέπει και η στρατηγική αξιολόγησης να ρυθμιστεί κατάλληλα με κριτήριο την ιδιότητα αυτή.

Παρόλα αυτά, μπορούμε να αναφέρουμε μερικές γενικές ιδιότητες για την κατασκευή γενικευμένων συνόλων χαρακτηριστικών. Κάθε τέτοιο σύνολο πρέπει να είναι αντιπροσωπευτικό του δικτύου προς εξέταση και απαιτούμε να είναι το μικρότερο δυνατό. Όπως θα δούμε όμως και αργότερα, οι ιδιότητες αυτές είναι σημαντικές και στη διαδικασία ανάθεσης των ρόλων. Στις τεχνικές low-rank approximation η διαχείριση όμοιων χαρακτηριστικών θα συμβεί αυτόματα με την σύμπτυξη τους σε χαμηλότερη διάσταση, ενώ τα πλεονάζοντα χαρακτηριστικά θα θεωρηθούν θόρυβος και θα απορριφτούν.

Η κατασκευή των χαρακτηριστικών μπορεί να συμβεί είτε αυτόματα είτε χειροκίνητα. Ο χειροκίνητος τρόπος αποτελεί εξαντλητική μέθοδο και δεν γίνεται αξιολόγηση, ή αναδρομικός υπολογισμός άλλων χαρακτηριστικών. Αρκεί επομένως ο καθορισμός τύπων και τελεστών. Έτσι το σύνολο των χαρακτηριστικών θα είναι προκαθορισμένο και βασισμένο στη γνώση των ειδικών του προς μελέτη πεδίου. Πλεονέκτημα της χειροκίνητης μεθόδου αποτελεί η ευκολία μετάφρασης των χαρακτηριστικών και των ρόλων, που θα προκύψουν. Επιπλέον, στην περίπτωση, που υπάρχει ισχυρή γνώση του αντικειμένου, θα οδηγηθούμε σε μια πιο σταθερή και αποτελεσματική εφαρμογή. Μειονεκτήματα αποτελούν τα λάθη, στα οποία μπορεί να υποπέσει ο ειδικός και την απώλεια κάποιου σημαντικού τελικά χαρακτηριστικού, καθώς και, ότι οι προϋποθέσεις του προβλήματος υπόκεινται σε αλλαγές με την πάροδο του χρόνου. Το κόστος από τις αναγκαίες αλλαγές και τη συνεχή ανάγκη επαναρρύθμισης της λειτουργίας του συστήματος, θα ήταν πολύ μεγάλο.

Από την άλλη μεριά στον αυτόματα τρόπο υπολογισμού, τα χαρακτηριστικά κατασκευάζονται αυτόματα, χωρίς την χρήση παραμέτρων καθορισμένων από τον χρήστη. Με την χρήση κυρίως αναδρομικών χαρακτηριστικών, δημιουργούνται συνεχώς νέες ιδιότητες, οι οποίες ελέγχονται αυτόματα για ομοιότητα μεταξύ τους και απορρίπτονται οι πλεονάζουσες. Η διαδικασία

αυτή επαναλαμβάνεται, μέχρι να μην υπάρχουν άλλα χρήσιμα χαρακτηριστικά προς κατασκευή. Η προσέγγιση αυτή είναι καταλληλότερη για μεγάλης κλίμακας αναλύσεις, όπου οι ρόλοι θα πρέπει να γενικευτούν μεταξύ διάφορων δικτύων, ή όπου δεν υπάρχουν αρκετές πληροφορίες για τα προς ανάλυση δίκτυα. Οι ρόλοι σαφώς θα είναι δυσκολότερο να εξηγηθούν, αλλά θα εντοπιστούν περισσότερα τυχαία κατασκευαστικά μοτίβα, γεγονός σημαντικό για τον εντοπισμό ανωμαλιών.

Τελικά, η πιο μεστή προσέγγιση θα ήταν η ύπαρξη ενός ειδικού για την επιλογή μερικών σημαντικών αρχικών χαρακτηριστικών και στη συνέχεια η χρήση κάποιου αυτοματοποιημένου συστήματος για την κατασκευή νέων τυχαίων χαρακτηριστικών. Έτσι θα προσεγγιστούν σωστά και οι σημαντικοί ρόλοι από τους ειδικούς, αλλά θα εντοπιστούν και τα τυχαία κατασκευαστικά μοτίβα του γράφου.

Επιλογή – αξιολόγηση τελικών χαρακτηριστικών

Μετά την έρευνα στον χώρο τον πιθανών χαρακτηριστικών, έχουμε καταλήξει σε ένα σύνολο από υποψήφια προς χρήση χαρακτηριστικά. Επόμενο βήμα στη διαδικασία αποτελεί η εύρεση ενός τρόπου αξιολόγησης των χαρακτηριστικών αυτών και τελικά η επιλογή ή απόρριψή τους. Κύριοι στόχοι της διαδικασίας αυτής πρέπει να είναι η μείωση των χαρακτηριστικών στα απολύτως απαραίτητα και η αποκοπή χαρακτηριστικών που προσθέτουν θόρυβο στο σύστημα.

Στο σημείο αυτό χρειάζεται να εξηγήσουμε και να ορίσουμε τον πίνακα ομοιότητας. Ο πίνακας ομοιότητας μπορεί να θεωρηθεί σαν ένα γράφος ομοιότητας μεταξύ των χαρακτηριστικών, τα βάρη του οποίου εκφράζουν την ομοιότητα μεταξύ δύο χαρακτηριστικών. Χρησιμοποιώντας μαθηματικούς ορισμούς, με δεδομένη μια συνάρτηση ομοιότητας S (έχουμε εξηγήσει αναλυτικά σε προηγούμενη ενότητα την έννοια της συνάρτησης ομοιότητας) και ένα σύνολο, ή πίνακα χαρακτηριστικών X, ορίζουμε τον πίνακα ομοιότητας ως $s = \forall(i,j) \in F, S(x_i, x_j)$. Υπάρχουν διάφορες συναρτήσεις ομοιότητας, που μπορούν να χρησιμοποιηθούν, όπως οι Pearson correlation και Spearman rank correlation. Ιδιαίτερα χρήσιμη είναι επίσης η logarithmic binning, για πολλά πραγματικά sparse δίκτυα, όπου παρατηρούνται ιδιαιτερότητες στις τοπικές μετρικές των κόμβων, όπως κοινωνικά και πληροφοριακά δίκτυα. Άλλες τεχνικές θα μπορούσαν να είναι η MIC και η BIC.

Μέχρι τώρα έχουμε εξετάσει κυρίως αυτόματες προσεγγίσεις, που βασίζονται στην αντιπροσωπευτικότητα των χαρακτηριστικών και την ελαχιστοποίηση τους. Πάραυτα, η διαδικασία εκμάθησης χαρακτηριστικών μπορεί να καθοδηγηθεί από την πρότερη γνώση της εφαρμογής και του πεδίου των ρόλων.

Με δεδομένο πια τον πίνακα ομοιότητας s, ο οποίος εκφράζει την ομοιότητα μεταξύ όλων των ζευγών χαρακτηριστικών, έχει έρθει η στιγμή να αποφασίσουμε ποια από αυτά τελικά θα κρατήσουμε. Το πρόβλημα αυτό μπορεί να θεωρηθεί pruning problem. Οδηγούμενοι από την ανάγκη επιλογής του μικρότερου δυνατού συνόλου χαρακτηριστικών, το pruning συνήθως γίνεται αυτόματα χρησιμοποιώντας ένα όριο της τάξης του 0.5 στη συνάρτηση ομοιότητας. Το όριο αυτό καθορίζει το επίπεδο στο οποίο δύο χαρακτηριστικά θα θεωρηθούν όμοια, ότι δηλαδή, συνεισφέρουν το ίδιο στην επιλογή ρόλων. Το όριο αυτό μπορεί φυσικά να ρυθμιστεί κατάλληλα, ανάλογα την εφαρμογή.

3.2.2 Σύνοψη και συνολική παρουσίαση

Παρακάτω παρουσιάζουμε έναν γενικευμένο επαναληπτικό αλγόριθμο που εφαρμόζει τα προηγούμενα. Ο αλγόριθμος έχει ως είσοδο έναν γράφο και κάποιες αρχικές ιδιότητές και στην έξοδό βγάζει ένα σύνολο χαρακτηριστικών, για την χρήση τους σε καθορισμό ρόλων. Τα χαρακτηριστικά αυτά περιγράφουν με επάρκεια τα μοτίβα και τις βασικές ιδιότητες του κάθε γράφου.

Role Feature Learning Template

Input: $G = (V, E, \mathbf{X}^{\text{attr}})$ – Initial graph and attributes, P – Set of primitive operators, Φ – Set of relational iterative operators, $\mathsf{S}(\cdot)$ – Score function, maxiter – Maximum number of iterations allowed, λ – Threshold for searching
$F_0 \leftarrow$ PRIMITIVES(G,P)
Let $\mathbf{X}$ be the feature data computed from the primitives
for $t \leftarrow 1$ **to** maxiter **do**
  $F_t, \mathbf{X} \leftarrow$FEATURESEARCH($F_{t-1}, \mathbf{X}, \Phi$)
  $F_t \leftarrow F_t \cup F_{t-1}$
  $\mathcal{G}_F \leftarrow$CREATEFEATUREGRAPH($F_t, \mathbf{X}, \mathsf{S}, \lambda$)
  $\mathcal{C} \leftarrow$ Partition the feature graph $\mathcal{G}_F$ (e.g., conn. components)
  for each $\mathcal{C}_k \in |\mathcal{C}|$ **do** ▷ Prune features
    Find the earliest (or min corr.) feature f_i s.t. $\forall f_j \in \mathcal{C}_k : i < j$.
    $F_t \leftarrow (F_t \setminus \mathcal{C}_k) \cup \{f_i\}$
  Remove features from $\mathbf{X}$ that were pruned (not in F_t)
  if $F_t = F_{t-1}$ **then terminate search** ▷ no new features
return $\mathbf{X}$ and F_t ▷ feature matrix $\mathbf{X} \in \mathbb{R}^{n \times f}$ and list F_t

procedure CREATEFEATUREGRAPH($F_{t-1}, \mathbf{X}, \mathsf{S}$), λ)
  Set $\mathcal{G}_F = (V_F, E_F)$ – the initial feature-graph
  Set V_F to be the set of features from F and $E_F = \emptyset$
  for each pair of features $(f_i, f_j) \in F_{t-1}$ **do**
    if $\mathsf{S}(f_i, f_j) \geq \lambda$ **then** Add edge (i, j) to E_F

Χρησιμοποιείται ένα σύνολο πολύ βασικών τελεστών (Ρ) και παράγονται χαρακτηριστικά που προστίθενται στα ήδη γνωστά. Ύστερα από pruning των χαρακτηριστικών, ο αλγόριθμος συνεχίζει επαναληπτικά. Σε κάθε επανάληψη υπολογίζονται νέα χαρακτηριστικά με τη χρήση σχεσιακών τελεστών (Φ) και προστίθενται στα προηγούμενα. Υπολογίζεται η ομοιότητα μεταξύ των χαρακτηριστικών, η οποία παριστάνεται σε γράφο και αποκόβονται οι ακμές μεταξύ των χαρακτηριστικών που δεν σχετίζονται μεταξύ τους. Έτσι έχουμε ως αποτέλεσμα έναν γράφο χαρακτηριστικών, όπου μεγάλα βάρη στις ακμές δηλώνουν συσχέτιση. Χρησιμοποιούμε τον γράφο αυτό για να αποκόψουμε μη απαραίτητα χαρακτηριστικά, τα οποία προσθέτουν θόρυβο. Για το λόγο αυτό κάνουμε διαμέριση (partitioning) στον γράφο. Πλέον, κάθε ανεξάρτητο κομμάτι του γράφου αποτελεί ένα σύνολο συσχετισμένων χαρακτηριστικών. Έτσι από κάθε τέτοιο κομμάτι μπορούμε να κρατήσουμε ένα μόνο χαρακτηριστικό, είτε αυτό που εμφανίστηκε νωρίτερα στη διαδικασία, είτε αυτό που σχετίζεται λιγότερο με τα υπόλοιπα.

Παρατηρούμε ότι σε κάθε επανάληψη, το σύνολο των χαρακτηριστικών που απομένει, είναι αντιπροσωπευτικό όλων των χαρακτηριστικών που έχουν προκύψει κατά τη διαδικασία. Ο αλγόριθμος σταματάει όταν πλέον δεν προκύπτουν νέα χαρακτηριστικά.

Ο αλγόριθμος λειτουργεί αποτελεσματικά σε κατευθυνόμενους και μη γράφους, οι οποίοι μπορεί να περιλαμβάνουν βάρη, χρονικές τιμές και να έχουν έναν τυχαίο αριθμό κόμβων και ακμών. Πολλά από τα μέρη του αλγόριθμου μπορούν εύκολα να αντικατασταθούν με άλλες τεχνικές, που επιτυγχάνουν το ίδιο αποτέλεσμα. Τέλος ο αλγόριθμος μπορεί να χρησιμοποιηθεί και για μεταφερόμενη μάθηση με τη χρήση ενός συνόλου χαρακτηριστικών, που προέκυψε από την εκτέλεση του σε έναν άλλο γράφο.

3.3 Ανάθεση ρόλων

Χρησιμοποιώντας τη γνώση που αποκτήσαμε στα προηγούμενα, στην ενότητα αυτή θα καθορίσουμε τον τρόπο ανάθεσης των ρόλων, Δηλαδή με ποιο τρόπο, κόμβοι με παρόμοιους πίνακες χαρακτηριστικών θα ανήκουν στον ίδιο ρόλο. Θα επικεντρωθούμε σε δύο μεγάλες ομάδες μεθόδων, οι οποίες χρησιμοποιούν την αναπαράσταση των χαρακτηριστικών του αρχικού γράφου. Αυτές είναι οι τεχνικές clustering και οι τεχνικές low rank approximation. Στη συνέχεια θα αναφέρουμε και τρόπους επιλογής του κατάλληλου αριθμού ρόλων. Θα παρατηρήσουμε επίσης, ότι οι περισσότερες τεχνικές ανάθεσης ρόλων μπορούν να μάθουν τους ορισμούς του κάθε ρόλου, ώστε να χρησιμοποιηθούν και σε άλλο γράφο.

3.3.1 Μέθοδοι clustering

Υπάρχουν αρκετοί αλγόριθμοι clustering, που μπορούν να χρησιμοποιηθούν για ανάθεση ρόλων, με χρήση της αναπαράσταση του γράφου με χαρακτηριστικά. Οι δύο κύριοι τύποι είναι οι hierarchical clustering και οι αλγόριθμοι διαμέρισης όπως ο k-means. Μερικοί από αυτούς, όπως ο k-means πραγματοποιούν αυστηρή ανάθεση, με την έννοια, ότι ένας κόμβος θα ανήκει αυστηρά σε έναν ρόλο, ενώ άλλοι όμως ο C-means επιτρέπουν στον κόμβο να καταλαμβάνει πολλούς ρόλους.

3.3.2 Προσεγγίσεις low-rank

Αποτελούν τη δεύτερη επιλογή για την ανάθεση ρόλων. Με δεδομένο έναν πίνακα χαρακτηριστικών $X \in R^{nxf}$, όπου n το πλήθος των κόμβων και f τα χαρακτηριστικά, υπολογίζουμε έναν πίνακα rank – r, όπου r < f, το πλήθος των ρόλων, ο οποίος προσεγγίζει όσο το δυνατόν καλύτερα τον αρχικό πίνακα. Μπορούν να χρησιμοποιηθούν διάφορα εργαλεία για τη μείωση της διάστασης, όπως SVD, Principal Component Analysis, Spectral Decomposition, Probabilistic Matrix Factorization, Non-negative Matrix Factorization.

Στις περισσότερες από τις τεχνικές υπάρχει ένας πίνακας U, όπου οι γραμμές αντιπροσωπεύουν την συμμετοχή κάθε κόμβου στους υπάρχοντες ρόλους. Στις περισσότερες low rank προσεγγίσεις, ένας κόμβος μπορεί να συμμετέχει σε πολλαπλούς ρόλους. Σε επόμενο κεφάλαιο θα εξετάσουμε αναλυτικότερα και θα εφαρμόσουμε τον αλγόριθμο NMF. Το αποτέλεσμα τον μεθόδων αυτών συνήθως είναι ένας πίνακας $U \in R^{nxr}$ που αντιπροσωπεύει τη συμμετοχή των κόμβων σε κάθε ρόλο και ένας πίνακας $S \in R^{rxf}$, που καθορίζει τις τιμές χαρακτηριστικών, από τις οποίες αποτελείται ο κάθε ρόλος.

Ανάλογα την εφαρμογή, τα προσδοκώμενα αποτελέσματα, τις υπολογιστικές απαιτήσεις σε μνήμη και επεξεργαστική ισχύ, χρειάζεται να χρησιμοποιηθεί διαφορετικός αλγόριθμος.

3.3.3 Επιλογή αριθμού ρόλων

Διαφορετικές τεχνικές έχουν προταθεί για την επιλογή κατάλληλου αριθμού ρόλων. Μερικές είναι ευρεστικές, ενώ άλλες βασίζονται στην στατιστική, όπως οι AIC και η MDL. Γενικά έχει παρατηρηθεί, ότι ένας μικρός αριθμός ρόλων, από 2 έως 15 είναι αρκετός για τους περισσότερους τύπους δικτύων. Σε επόμενο κεφάλαιο θα αναφέρουμε περισσότερα πράγματα για την τεχνική MDL.

Κεφάλαιο 4 - Refex και Rolx - Εφαρμογή της θεωρίας

4.1 Εισαγωγή

Στο παρόν κεφάλαιο θα παρουσιάσουμε έναν πραγματικό αλγόριθμο που εφαρμόζει τις αρχές και το πλαίσιο που αναλύθηκε στο προηγούμενο κεφάλαιο και στο τέλος θα αναλύσουμε τον τρόπο που τον εφαρμόσαμε εμείς στην πράξη σε πραγματικά δεδομένα κοινωνικών δικτύων. Ο αλγόριθμος αυτός αποτελείται από δύο ξεχωριστούς αλγόριθμους.

- Τον Refex, με τον οποίο εξάγουμε τον πίνακα χαρακτηριστικών των κόμβων ενός γράφου.
- Τον Rolx, ο οποίος βοηθάει στην εξαγωγή ρόλων με βάση τον πίνακα χαρακτηριστικών του προηγούμενου βήματος.

4.2 Αλγόριθμος Refex (Recursive feature extraction)

4.2.1 Εισαγωγή

Ο αλγόριθμος Refex αποτελεί κλασικό παράδειγμα εφαρμογής των βημάτων εκμάθησης χαρακτηριστικών του πλαισίου που ορίσαμε προηγουμένως. Περιλαμβάνει την εύρεση ενός συνόλου τυπικών χαρακτηριστικών μέσα στον γράφο και την κατασκευή αναδρομικών χαρακτηριστικών μέσω της εφαρμογής κάποιων τελεστών. Υλοποιεί μια στρατηγική διερεύνησης των χαρακτηριστικών, ώστε κάποια στιγμή η διαδικασία να τελειώσει και στο τέλος επιλέγονται τα χαρακτηριστικά που θα χρησιμεύσουν στο επόμενο βήμα.

4.2.2 Χαρακτηριστικά γειτνίασης

Χρησιμεύουν ως "σπόροι" για τον υπολογισμό των αναδρομικών χαρακτηριστικών του αλγορίθμου. Αποτελούνται από τοπικά χαρακτηριστικά και χαρακτηριστικά του egonet. Τα τοπικά χαρακτηριστικά συνήθως αποτελούν μετρικές σχετικές με το βαθμό κάθε κόμβου, ενώ τα χαρακτηριστικά του egonet αφορούν τον υπογράφο του κόμβου μαζί με τους γείτονες του, όπως τις ακμές που εισέρχονται στο egonet, τις ακμές που εξέρχονται και τα λοιπά.

Το egonet αποτελείται από τους γείτονες κάθε κόμβου. Έτσι στο egonet επιπέδου μηδέν, ανήκουν όλοι οι κόμβοι που απέχουν απόσταση ενός βήματος από τον προς εξέταση κόμβο, στο egonet επιπέδου ένα αυτοί που απέχουν το πολύ δύο βήματα και ούτω καθεξής.

Σε περίπτωση κατευθυνόμενων γράφων, ή για γράφους με βάρη προσαρμόζουμε τις μετρικές κατάλληλα ώστε να ανταποκρίνονται σωστά στην κάθε κατάσταση.

4.2.3 Αναδρομικά χαρακτηριστικά

Εξαγωγή αναδρομικών χαρακτηριστικών

Τα αναδρομικά χαρακτηριστικά προκύπτουν εφαρμόζοντας κάποιον τελεστή στα ήδη υπάρχοντα χαρακτηριστικά των γειτόνων του κόμβου. Παραδείγματα τελεστών είναι το άθροισμα και ο μέσος όρος. Για μια πιο εκτεταμένη λίστα τελεστών, ανατρέξτε στο προηγούμενο κεφάλαιο.

Στην περίπτωση του Refex χρησιμοποιούνται δύο τύποι αναδρομικών χαρακτηριστικών. Αθροίσματα και μέσοι όροι. Τα χαρακτηριστικά στα οποία μπορούν να εφαρμοστούν τελεστές δεν περιορίζονται στα χαρακτηριστικά γειτνίασης ή σε κατασκευαστικά χαρακτηριστικά, αλλά μπορούν να συμμετέχουν και γνωρίσματα που προϋπήρχαν ως είσοδος στο σύστημα μας, καθώς και διαφορετικά αναδρομικά χαρακτηριστικά.

Αποκοπή άχρηστων αναδρομικών χαρακτηριστικών (Vertical logarithmic binning)

Καθώς ο αριθμός των αναδρομικών χαρακτηριστικών μπορεί δυνητικά να είναι άπειρος και μεγαλώνει εκθετικά σε κάθε επανάληψη εφαρμογής των επιλεγμένων τελεστών, θα πρέπει να υπάρξει ένας τρόπος ώστε να αποκόβονται τα χαρακτηριστικά που δεν μας προσφέρουν παραπάνω γνώση από τα ήδη υπάρχοντα. Η ιδέα πίσω από την αποκοπή αφορά την αναζήτηση ζευγαριών από χαρακτηριστικά, τα οποία παρουσιάζουν υψηλή συσχέτιση για όλους τους κόμβους.

Κατά την υλοποίηση του Refex και κυρίως για λόγους υπολογιστικής πολυπλοκότητας, οι τιμές των χαρακτηριστικών αντιστοιχίζονται σε μικρούς ακέραιους με τη χρήση της μεθόδου vertical logarithmic binning. Η μέθοδος εφαρμόζεται ως εξής:

Για κάθε χαρακτηριστικό ορίζεται μια παράμετρος p, όπου $0 < p < 1$. Για το χαρακτηριστικό f, οι p * |V| κόμβοι με την χαμηλότερη τιμή του f, όπου |V| το σύνολο των κόμβων του αρχικού γράφου, αντιστοιχίζονται στην τιμή 0. Οι p * Ve, κόμβοι με την αμέσως μεγαλύτερη τιμή, όπου Ve οι εναπομένοντες κόμβοι, αντιστοιχίζονται στην τιμή 1, το αμέσως επόμενο τμήμα p των εναπομενόντων κόμβων στην τιμή 2 και ούτω καθεξής. Η διαδικασία συνεχίζεται έως ότου οι τιμές του χαρακτηριστικού f για όλους τους κόμβους αντιστοιχιστούν σε ακέραιες τιμές από το 0 έως το $\log_{p-1}(|V|)$.

Μόλις η παραπάνω διαδικασία ολοκληρωθεί για κάποιο χαρακτηριστικό, ο αλγόριθμος ψάχνει για ζεύγη χαρακτηριστικών, οπού για κάθε κόμβο, η τιμή τους δεν διαφέρει παραπάνω από ένα όριο s. Τα χαρακτηριστικά αυτά θεωρούνται όμοια. Στη συνέχεια κατασκευάζεται ένας νέος γράφος , όπου ο κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό του προηγούμενου γράφου και κάθε ακμή συνδέει όμοια μεταξύ τους χαρακτηριστικά, όπως ορίστηκαν παραπάνω. Κάθε συνδεδεμένο τμήμα του γράφου αντικαθίσταται με ένα από τα χαρακτηριστικά από τα οποία αποτελείται, δίνοντας προτεραιότητα στο χαρακτηριστικό που προήλθε χρονικά πρώτο στις επαναλήψεις.

Στην περίπτωση που δεν προκύψει κανένα χρήσιμο χαρακτηριστικό ο αλγόριθμος σταματάει και επιστρέφει τα υπάρχοντα χαρακτηριστικά μέχρι εκείνο το σημείο.

4.2.4 Παράμετροι

Ο Refex απαιτεί δύο παραμέτρους για να λειτουργήσει, το p για το logarithmic binning, όπως ορίστηκε παραπάνω και το s ως όριο ομοιότητας.

Το p αντιστοιχεί σε τιμές από 0 έως 1. Θέτοντας το p πολύ κοντά στο 1 οδηγεί σε πιο επιθετική αποκοπή χαρακτηριστικών, το οποίο δια μπορούσε να οδηγήσει σε εξάλειψη της διακριτότητας μεταξύ των χαρακτηριστικών. Φτάνοντας το p πολύ κοντά στο 0, μπορεί να οδηγήσει στην διατήρηση πάρα πολλών χαρακτηριστικών με αποτέλεσμα την αύξηση του χρόνου εκτέλεσης και της πολυπλοκότητας των αποτελεσμάτων. Από τους δημιουργούς του αλγορίθμου, ως μια λογική τιμή του p προτείνεται το 0,5.

Σχετικά με το s, ο Refex εφαρμόζει χαλάρωση του ορίου σε κάθε επανάληψη. Για μικρούς γράφους μικρότερους από 100000 κόμβους, χρησιμοποιείται s = 0 για την πρώτη επανάληψη. Σε κάθε επόμενη επανάληψη, το s αυξάνεται κατά 1. Αυτό εγγυάται ότι ο αλγόριθμος θα σταματήσει ύστερα από το πολύ $\log_{p-1}(|V|)$ επαναλήψεις, καθώς σε εκείνο το σημείο η μέγιστη τιμή κάθε χαρακτηριστικού θα είναι ίση με s.

4.2.5 Ανάλυση πολυπλοκότητας Refex

Έστω n ο αριθμός των κόμβων, m ο αριθμός των ακμών, M ο μέγιστος βαθμός μεταξύ όλων των κόμβων, f ο αριθμός των χαρακτηριστικών και d_i ο βαθμός του κόμβου i.

Η υπολογιστική πολυπλοκότητα του αλγορίθμου μπορεί να χωριστεί σε δύο βήματα.

- Υπολογισμός των χαρακτηριστικών γειτνίασης
- Υπολογισμός των αναδρομικών χαρακτηριστικών σε κάθε επανάληψη

Ο υπολογισμός των χαρακτηριστικών γειτνίασης αναμένεται να είναι O(n) για πραγματικούς γράφους, σύμφωνα με απόδειξη που δίδεται από τους δημιουργούς του αλγορίθμου.

Σε κάθε αναδρομική επανάληψη ο αλγόριθμος έχει πολυπλοκότητα

O(f * (m + n * f)

4.3 Αλγόριθμος Rolx (Role extraction)

4.3.1 Εισαγωγή

Ο αλγόριθμος Rolx έναν γραμμικό ως προς τον αριθμό των ακμών αλγόριθμο εκμάθησης για αυτοματοποιημένη εξαγωγή ρόλων ενός δικτύου, με είσοδο δεδομένα του δικτύου. Ο Rolx έχει δύο βασικά χαρακτηριστικά. Αποφασίζει αυτόματα τους ρόλους του δικτύου, χωρίς προϋπάρχουσα γνώση των ρόλων που ίσως υπάρχουν και στη συνέχεια επιτρέπει σε έναν κόμβο να συμμετέχει σε πολλαπλούς ρόλους.

Χρησιμοποιείται σε μια πληθώρα εφαρμογών που αφορούν την ανάλυση δικτύων, όπως σε network transfer learning, μέτρηση της κατασκευαστικής ομοιότητας δύο δικτύων, κατανόηση της πραγματικής συμπεριφοράς πίσω από ένα δίκτυο κτλ. Για παράδειγμα εξάγοντας ρόλους από ένα γράφο, μπορούμε να υλοποιήσουμε ένα αλγόριθμο διαφοροποίησης (classification) για κάποιον άλλον γράφο, οδηγώντας στην μηχανική εκμάθηση μιας ολόκληρης κατηγορίας δικτύων.

Ο Rolx εντοπίζει ρόλους μέσα από έναν πίνακα χαρακτηριστικών των κόμβων και συνοψίζεται σε δύο βήματα. Εφόσον υπάρχει ήδη ο πίνακας χαρακτηριστικών, αρχικά υπολογίζεται το διάνυσμα κάθε ρόλου στον χώρο των χαρακτηριστικών και στη συνέχεια αποφασίζεται το ποσοστό συμμετοχής κάθε κόμβου στον κάθε ρόλο. Τα χαρακτηριστικά που χρησιμοποιούνται ως είσοδος εξαρτώνται κάθε φορά από την φύση του εκάστοτε προβλήματος και δικτύου.

Παρακάτω αναλύουμε τον τρόπο με τον οποίο ο αλγόριθμος υλοποιεί τις παραπάνω λειτουργίες. Με το κομμάτι της εξαγωγής χαρακτηριστικών δεν θα ασχοληθούμε, καθώς καλύφθηκε επαρκώς στην προηγούμενη ενότητα.

4.3.2 Ομαδοποίηση χαρακτηριστικών σε ρόλους

Ύστερα από την εξαγωγή χαρακτηριστικών έχουμε ως αποτέλεσμα n διανύσματα, δηλαδή έναν για κάθε κόμβο, οπού το καθένα έχει f αριθμητικές συνιστώσες, όσες και τα χαρακτηριστικά του γράφου. Χρειαζόμαστε έναν τρόπο να ομαδοποιήσουμε τους κόμβους που έχουν παρόμοια διανύσματα χαρακτηριστικών.

Ο τρόπος που χρησιμοποιεί ο αλγόριθμος είναι η παραγοντοποίηση πινάκων. Έστω ένας πίνακας $V_{n \times f}$, όπου αντιπροσωπεύει τους κόμβους με τα χαρακτηριστικά τους, θα υπολογιστεί μια προσέγγιση GF ~= V, όπου $G_{n \times r}$ αντιπροσωπεύει την κατανομή των κόμβων σε ρόλους και $F_{r \times f,}$ συμβολίζει τη συμμετοχή των χαρακτηριστικών σε κάθε ρόλο. Το μαθηματικό εργαλείο που χρησιμοποιείται για την παραγοντοποίηση του πίνακα V είναι το NMF.

NMF(Non – negative matrix factorization)

Ουσιαστικά αποτελεί ένα σύνολο αλγορίθμων και μεθόδων για την παραγοντοποίηση πινάκων. Παίρνει δηλαδή ως είσοδο τον πίνακα V και δίνει ως έξοδο δύο πίνακες G και F, ώστε G x F ~= V, όπου να τονίσουμε ότι οι V, F και G περιέχουν στοιχεία με τιμές μεγαλύτερες ή ίσες του 0 (non negative). Ο περιορισμός των μη αρνητικών τιμών έχει ως συνέπεια να οδηγούμαστε σε sparse πίνακες G και F, που περιγράφουν τον αρχικό πίνακα, το οποίο έχει μεγαλύτερο νόημα για την απόδοση ρόλων και βοηθάει στη συνέχεια στην καλύτερη αποκωδικοποίηση τους ώστε να εντοπιστεί το πραγματικό νόημα της λειτουργίας του ρόλου αυτού.

Μαθηματικά, ο αλγόριθμος αναζητά δυο μη αρνητικούς πίνακες G και F, χαμηλής τάξης ώστε να λύσει το πρόβλημα βελτιστοποίησης¨

$\text{Argmin}_{G,F} ||V - GF||_{fro}$, όπου $||.||$, η νόρμα Frobenius.

Αντί για τη νόρμα Frobenius, προκειμένου να μετρήσουμε ομοιότητα, μπορεί να χρησιμοποιηθεί και η μέθοδος Kullback Leibler divergence, η οποία είναι καταλληλότερη για sparse πίνακα V. Στην εκδοχή του NMF που χρησιμοποιήθηκε από εμάς, προτιμήθηκε η Kullback Leibler divergence.

Το πρόβλημα που προκύπτει με την επιλογή της παραγοντοποίσης πινάκων για την ομαδοποίηση χαρακτηριστικών, είναι ότι το μέγεθος του μοντέλου (ο αριθμός των ρόλων δηλαδή), θα πρέπει να προκαθοριστεί. Καθώς είναι μη πρακτικό να γνωρίζουμε πριν πριν τον βέλτιστο αριθμό ρόλων στην επόμενη

υποενότητα αναλύουμε τον τρόπο επιλογής του μεγέθους του μοντέλου, δηλαδή τον αριθμό των ρόλων

Να υπενθυμίσουμε σε αυτό το σημείο ότι σκοπός της εργασίας αυτής δεν είναι η επεξήγηση των μαθηματικών εργαλείων, αλλά οι μέθοδοι ανάλυσης των δικτύων για εξαγωγή ρόλων. Στην υλοποίηση του αλγορίθμου που πραγματοποιήσαμε, τα μαθηματικά εργαλεία αντιμετωπίστηκαν ως «μαύρα κουτιά» και από πλευράς κώδικα, ως έτοιμα πακέτα της python. Αν ο αναγνώστης επιθυμεί να εντρυφήσει περισσότερο στο μαθηματικό κομμάτι, μπορεί να ανατρέξει στη βιβλιογραφία στο τέλος της εργασίας.

MDL(Minimum description length)
Αν σκεφτούμε ότι η εξαγωγή ρόλων, πρακτικά μοντελοποιεί και συμπιέζει τον πίνακα V, θα πρέπει να επιλέξουμε ένα μέγεθος μοντέλου που να εξασφαλίζει την μέγιστη και βέλτιστη συμπίεση. Για την εύρεση λοιπόν του βέλτιστου μοντέλου ο Rolx χρησιμοποιεί το κριτήριο MDL (Minimum description length), το οποίο ουσιαστικά αποφασίζει τον αριθμό των ρόλων.

Θα αναφέρουμε λίγες λεπτομέρειες σχετικά με το πώς λειτουργεί το κριτήριο. Για να υπολογιστεί το μέγεθος της περιγραφής του μοντέλου (description length) απαιτούνται δύο μέρη. Πρώτα ο αριθμός των bits που απαιτούνται για την περιγραφή του ίδιου του μοντέλου (ονομάζουμε το κόστος αυτό M) και στη συνέχεια το κόστος της περιγραφής των λαθών ανακατασκευής του V – GF, έτσι ώστε να έχουμε συμπίεση χωρίς απώλειες (ονομάζουμε το κόστος αυτό E. Το συνολικό κόστος λοιπόν είναι L = M + E.

Με βάση τα παραπάνω, χρησιμοποιώντας b bits ανά τιμή του πίνακα, το κόστος περιγραφής του μοντέλου είναι n * r * b για τον πίνακα G και αντίστοιχα r * f * b για τον πίνακα F. Συνολικά λοιπόν και για τους δύο πίνακες θα έχουμε M = b * r * (n + f).

Ο υπολογισμός του κόστους ανακατασκευής των λαθών στην προσέγγιση απαιτεί μαθηματικούς υπολογισμούς σε αρκετά υψηλότερο επίπεδο από τον σκοπό της εργασίας. Θα αναφέρουμε επιγραμματικά ότι το τελικό αποτέλεσμα είναι $E = \sum_{i,j}(V_{i,j} log \frac{V_{ij}}{(GF)_{i,j}} - V_{i,j} + (GF)_{i,j})$. Για τον υπολογισμό ομοιότητας χρησιμοποιείται KL divergence, καθώς τα λάθη στη σύγκριση ομοιότητας V – GF δεν ακολουθούν κανονική κατανομή. Επειδή οι τιμές των πινάκων του μοντέλου είναι δεκαδικοί αριθμοί υψηλής ακρίβειας, χρησιμοποιείται κβαντοποίηση Lloyd-Max με $log_2(n)$ κβάντα, σε συνδυασμό με κωδικούς Huffman, προκειμένου να ενισχυθεί η συμπίεση

4.3.3 Υπολογιστική πολυπλοκότητα Rolx
Ορίζουμε n τον αριθμό των κόμβων, m τον αριθμό των ακμών, f το πλήθος των χαρακτηριστικών και r το πλήθος των ρόλων που προέκυψαν. Η

χρονική πολυπλοκότητα του Rolx είναι γραμμική ως προς τις ακμές και πιο συγκεκριμένα είναι O(mf + nfr).

Η εξαγωγή χαρακτηριστικών όπως αναφέρθηκε και προτίτερα έχει χρονικό κόστος O(f(m + nf)).

Για τον υπολογισμό του σφάλματος απαιτείται O(nrf) για τον πολλαπλασιασμό δύο πινάκων, (n x r) ο ένας και (r x f) ο άλλος.

Η κβαντοποίηση έχει κόστος O(nfilog(K)), όπου K ο αριθμός των κβάντων και i ο αριθμός των επαναλήψεων που τρέχουμε τον αλγόριθμο κβαντοποίησης. Το K που χρησιμοποιήσαμε είναι ίσο με log(n), οπότε συνολικά θα έχουμε O(nfilog(log(n))). Σε αυτό το σημείο παρατηρούμε ότι ο όρος O(log(log(n)) είναι ένας πολύ μικρός αριθμός. Οπότε προσεγγιστικά η πολυπλοκότητα της κβαντοποίησης γίνεται O(nfi)

Σχετικά με την κωδικοποίηση Huffman η πολυπλοκότητα είναι O(nf + Klog(K)). Επειδή όμως και πάλι ο όρος O(Klog(K)) είναι πολύ μικρός, μπορούμε να τον αγνοήσουμε.

Τέλος, το κομμάτι του NMF έχει πολυπλοκότητα στην χειρότερη περίπτωση $O(nfr + nr^2 + fr^2) = O(nfr)$

Αν τα αθροίσουμε όλα αυτά μεταξύ τους έχουμε O(mf + nfr).

Κεφάλαιο 5 - Εκτέλεση πειράματος

5.1 Εισαγωγή

Προκειμένου να παρατηρήσουμε στην πράξη τους αλγορίθμους της εξαγωγής ρόλων, πραγματοποιήθηκε εκτέλεση του αλγορίθμου πάνω σε δεδομένα κοινωνικών δικτύων που βρέθηκαν στο διαδίκτυο. Στο κεφάλαιο αυτό θα κάνουμε μια παρουσίαση των δεδομένων και των αποτελεσμάτων του αλγόριθμου και θα προσπαθήσουμε να καταλήξουμε σε κάποια συμπεράσματα.

Ο αλγόριθμος εκτελέστηκε σε δύο σετ δεδομένων όπου παρουσιάζονται στην επόμενη υποενότητα. Και τα δύο σετ αφορούν δεδομένα του facebook για τα οποία δεν υπάρχει πρότερη γνώση πέρα από το ότι αφορούν ένα δίκτυο φίλων στο facebook.

5.2 Πρώτο πείραμα

5.2.1 Πληροφορίες για τα δεδομένα

Το δίκτυο που αντιπροσωπεύουν τα δεδομένα αφορά ένα δίκτυο φιλίας στο facebook. Πρόκειται για ένα σχετικά αραιό και μικρό δίκτυο, ώστε να διευκολύνει στην εξαγωγή συμπερασμάτων σχετικά με τη λειτουργία του αλγόριθμου εξαγωγής ρόλων. Πηγή δικτύου: http://networkrepository.com/socfb-nips-ego.php

Στον παρακάτω τμήματα δίδονται μερικά στατιστικά στοιχεία του γράφου, ώστε να έχει μια πιο ολοκληρωμένη άποψη ο αναγνώστης.

Πλήθος κόμβων	2888
Πλήθος ακμών	2982
Κατευθυνόμενος	Όχι

Μέσος βαθμός	2,064
Διάμετρος δικτύου	9
Modularity	0,809
Μέσο μήκος μονοπατιού	3,867

Στατιστικά στοιχεί δεδομένων πρώτου πειράματος. Η εξαγωγή τους έγινε με το πρόγραμμα Gephi.

Όπως παρατηρούμε στον πίνακα, έχουμε έναν μικρού μεγέθους, αραιό γράφο, ώστε να μας βοηθήσει στην εξαγωγή συμπερασμάτων σχετικά με την ορθή λειτουργία του αλγόριθμου. Η αναλογία κόμβων είναι περίπου 1:1, το οποίο δημιουργεί ένα ευανάγνωστο πλαίσιο για τη συνέχεια. Ο γράφος είναι μη κατευθυνόμενος, καθώς η φιλία στο facebook αντιμετωπίζεται πάντα ως διμερής σχέση.

Τα προβλήματα που αναμένουμε από αυτή την επιλογή είναι αστοχίες του αλγόριθμου εξαιτίας του μικρού βάθους του γράφου και αδυναμία αναγωγής των αποτελεσμάτων σε συμπεράσματα σχετικά με προβλήματα πραγματικών μεγάλων δικτύων.

5.2.2 Παράμετροι εκτέλεσης

Ο αλγόριθμος έτρεξε με σταθερά p = 0,5, αρχική τιμή s ίση με 0 και αύξηση της κατά ένα ανά επανάληψη, μέγιστο όριο επαναλήψεων τις 100, το οποίο δεν εξαντλήθηκε, καθώς η MDL σταμάτησε νωρίτερα τον αλγόριθμο.

5.2.3 Αρχικά χαρακτηριστικά

Παρακάτω δίδεται λίστα με τα αρχικά χαρακτηριστικά που υπολογίστηκαν στο πρώτο τμήμα του refex. Πρόκειται για κατασκευαστικά χαρακτηριστικά κάθε κόμβου και μετρικές του egonet κάθε κόμβου. Χαρακτηριστικά του egonet υπολογίστηκαν σε βάθος (επίπεδο) μηδέν και βάθος ένα.

Συμβολισμός	Εξήγηση
wn	Πλήθος εσωτερικών κόμβων egonet
weu	Πλήθος μοναδικών ακμών μέσα στο egonet
wet	Πλήθος συνολικών ακμών μέσα στο egonet
xesu	Πλήθος μοναδικών ακμών που εξέρχονται από το egonet
xest	Πλήθος συνολικών ακμών που εξέρχονται από το egonet

xedu	Πλήθος μοναδικών ακμών που εισέρχονται στο egonet
xedt	Πλήθος συνολικών ακμών που εισέρχονται στο egonet

Αρχικά κατασκευαστικά χαρακτηριστικά

Το πλήθος των κόμβων του egonet είναι απαραίτητο ώστε να φανεί πόσο «διάσημος» είναι ένας κόμβος. Άλλωστε μεγάλο πλήθος δηλώνει, πολλούς γνωστούς, άρα μεγαλύτερη αναγνωρισιμότητα.

Οι μοναδικές ακμές του egonet δείχνουν κατά πόσο οι κόμβοι του egonet επικοινωνούν άμεσα μεταξύ τους και αν η επικοινωνία τους οφείλεται στην λειτουργία ενός κεντρικού κόμβου. Μικρή τιμή σε αυτήν την μετρική σε συνδυασμό με μεγάλο πλήθος κόμβων στο egonet δείχνει κεντρικότητα και επιρροή για τον προς εξέταση κόμβο.

Το πλήθος των συνολικών ακμών για το δικό μας δίκτυο δεν έχει τόσο μεγάλη σημασία, αλλά περιλήφθηκε για λόγους πληρότητας του αλγορίθμου

Σχετικά με το πλήθος μοναδικών ακμών που εξέρχονται από το egonet, δείχνει αν ένα ολόκληρο υποδίκτυο μπορεί να έχει τον ρόλο της πηγής σε ένα μεγαλύτερο δίκτυο. Αντίστοιχα το πλήθος μοναδικών ακμών που εισέρχονται στο egonet, δείχνει ένα είδος προορισμού. Αν και οι δύο αυτές τιμές είναι αρκετά μεγαλύτερες από τα υπόλοιπα egonets του δικτύου, τότε το συγκεκριμένο egonet έχει κεντρικό ρόλο στη διάδοση της πληροφορίας και στη σύνδεση κομματιών του δικτύου.

5.2.4 Χαρακτηριστικά από τελεστές και αναδρομικά χαρακτηριστικά

Στη δημιουργία νέων χαρακτηριστικών με βάση τελεστές, χρησιμοποιήθηκε μόνο το άθροισμα με τον τρόπο που φαίνεται στον παρακάτω πίνακα:

Συμβολισμός	Εξήγηση
wea	Άθροισμα του wet για κάθε κόμβο του egonet
xesa	Άθροισμα του xest για κάθε κόμβο του egonet
xeda	Άθροισμα του xedt για κάθε κόμβο του egonet
xea	Άθροισμα του xeda και xesa
xeu	Άθροισμα του xesu και xedu
xet	Άθροισμα του xest και xedt

Χαρακτηριστικά με τελεστές

Στην γέννηση των αναδρομικών χαρακτηριστικών, χρησιμοποιήθηκαν δύο τελεστές, αθροίσματα και μέσοι όροι. Ένα παράδειγμα τέτοιου χαρακτηριστικού είναι το εξής:

xedu1-0-s0-1-m0-2-m0 : Βρίσκουμε το xedu επιπέδου 1 για όλους τους κόμβους. Στη συνέχεια κατά την επανάληψη 0, πήραμε το άθροισμα του xedu για όλους τους κόμβους του egonet. Κατά την επανάληψη 1, πήραμε τον μέσο όρο των αθροισμάτων του xedu. Κατά την επανάληψη δύο πήραμε τον μέσο όρο, των μέσων όρων, των αθροισμάτων του xedu. Τα χαρακτηριστικά αξιολογούνται ως χρήσιμα για την ανάλυση ή όχι με βάση μαθηματικά εργαλεία που επεξηγήθηκαν σε προηγούμενο κεφάλαιο

Φυσικά όλη αυτή η διαδικασία γίνεται αυτόματα και στην πραγματικότητα χαρακτηριστικά σαν και αυτά δεν έχουν αναγνωρίσιμη πραγματική σημασία. Χρησιμεύουν για την καλύτερη αναπαράσταση και συμπίεση του γράφου, καθώς και στη δημιουργία ρόλων

5.2.5 Αποτελέσματα

Μετά την εκτέλεση του αλγορίθμου έχουμε τα εξής ευρήματα:

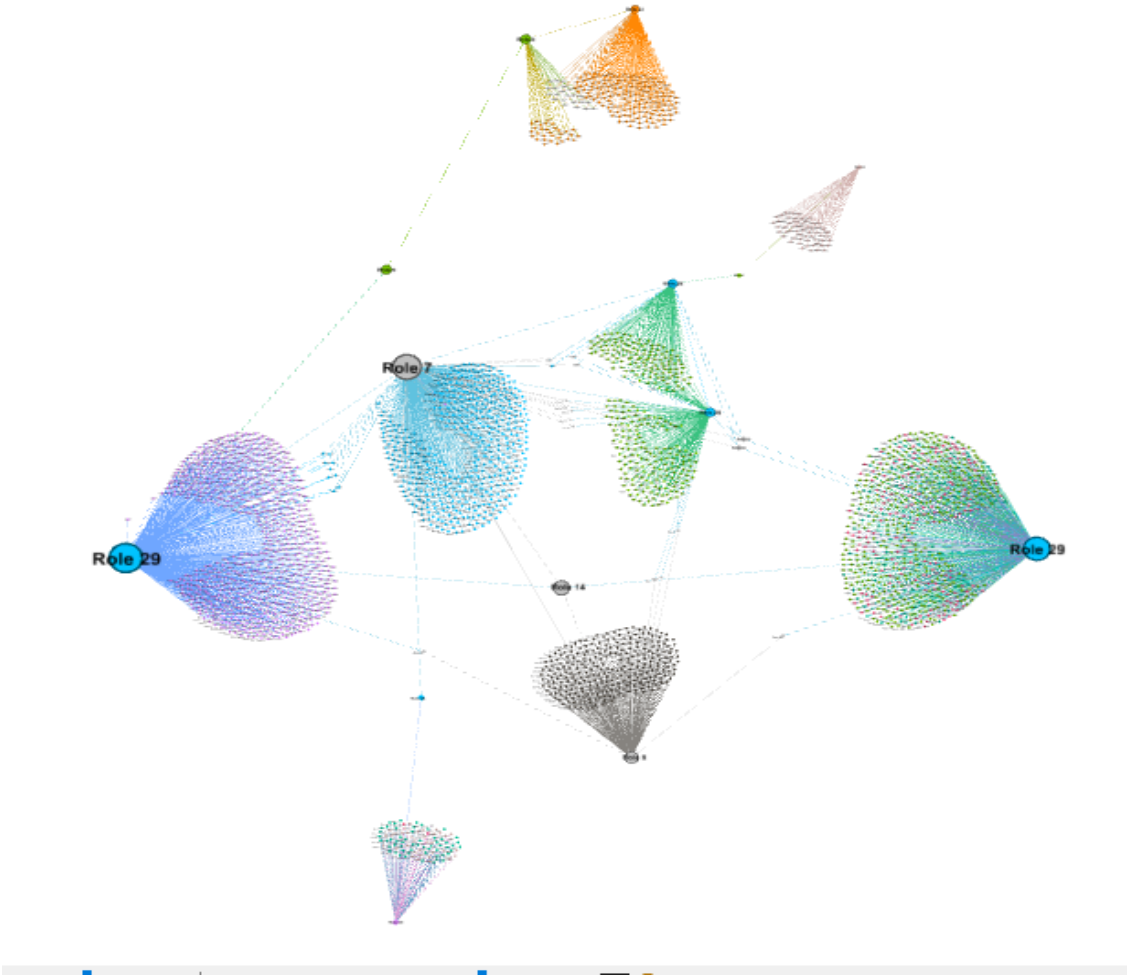
Εντοπίστηκαν 44 διαφορετικοί ρόλοι, μέσα από 143 χαρακτηριστικά. Όπως προαναφέρθηκε, κάθε ρόλος μπορεί να περιλαμβάνει συμμετοχή σε διάφορα χαρακτηριστικά και κάθε κόμβος να συμμετέχει σε διαφορετικούς ρόλους. Για χάρη απλότητας και για διευκόλυνση της ανάλυσης κρατήσαμε τον κυρίαρχο ρόλο για κάθε κόμβο, καθώς και τα πέντε κυρίαρχα χαρακτηριστικά για κάθε ρόλο. Σε αυτό το σημείο φαίνεται ήδη πως η MDL ίσως να μην είναι κατάλληλο εργαλείο, καθώς το πλήθος των ρόλων που προέκυψαν είναι αρκετά μεγάλο.

Στην επόμενη σελίδα θα δείτε τη συνολική εικόνα του γράφου, με τα χρώματα να αντιπροσωπεύουν τον κυρίαρχο ρόλο. Μέσα από στιγμιότυπα του γράφου θα προσπαθήσουμε να ερμηνεύσουμε τη σημασία μερικών ρόλων, αντικείμενο αρκετά δύσκολο επί της αρχής.

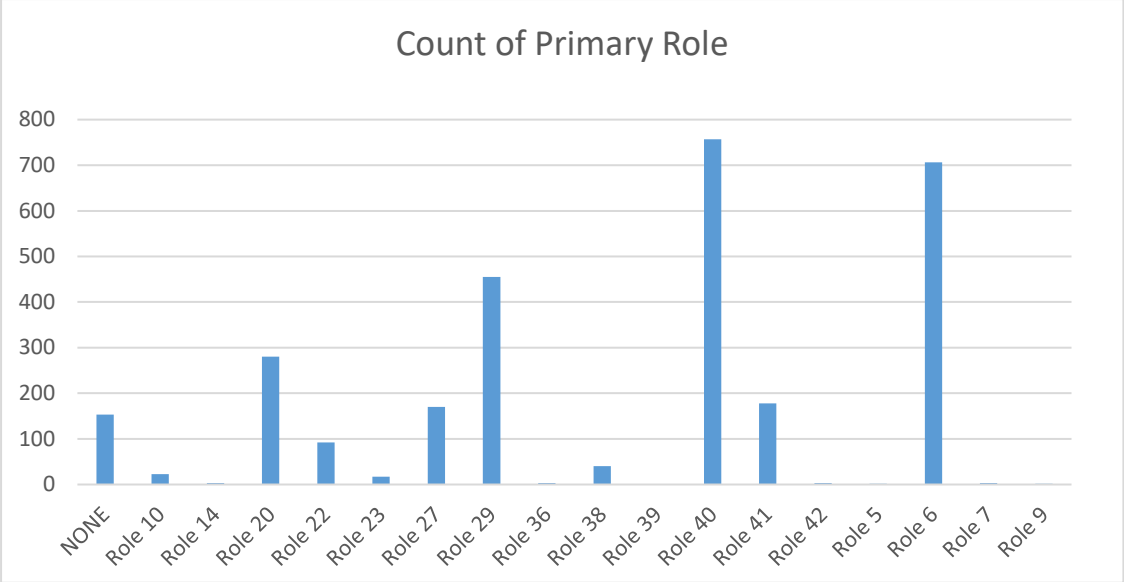
Στην πράξη και στις πραγματικές εφαρμογές του αλγορίθμου, τα αποτελέσματα αυτά δεν υπάγονται απαραίτητα σε ανθρώπινη ερμηνεία. Μπορούν να χρησιμοποιηθούν ως ενδιάμεσα στάδια για ενδιάμεσες εφαρμογές, ή για τον μηχανικό και αυτόματα εντοπισμό διαφορών και ανωμαλιών. Στο τέλος αυτού του κεφαλαίου θα αναφερθούν παραδείγματα τέτοιων εφαρμογών.

Με βάση τα παραπάνω τίθεται το ερώτημα της ορθότητας ενός αλγορίθμου. Αν δεν μπορώ να ελέγξω ότι τα αποτελέσματα έχουν κάποιο νόημα, πως θα είμαι σε θέση να γνωρίζω αν ο αλγόριθμος μου δουλεύει; Σύμφωνα λοιπόν με τις μετρήσεις και αναλύσεις των δημιουργών του, καθώς και τον ειδικών του κάθε κλάδου που αντιπροσωπεύει το εκάστοτε δίκτυο, υπάρχουν πολύ σημαντικές ενδείξεις ότι τα αποτελέσματα έχουν νόημα και είναι

αντιπροσωπευτικά του κάθε δικτύου. Το πλήθος των συστημάτων που χρησιμοποιούν Refex και Rolx για την αναπαράσταση και ερμηνεία των συμπεριφορών ενός δικτύου, ενισχύουν την άποψη αυτή.

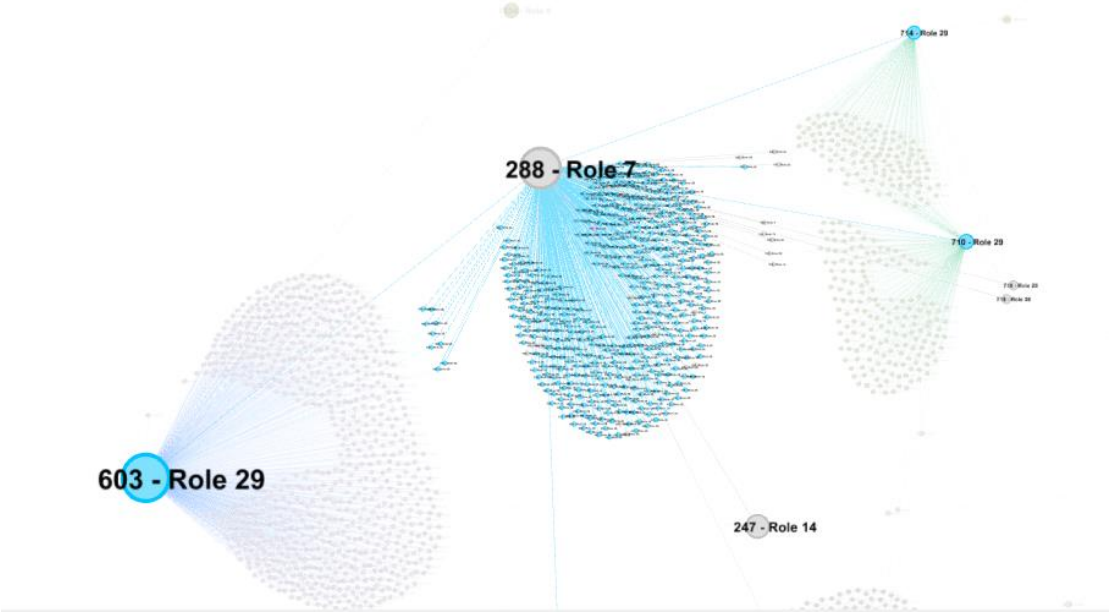


Το δίκτυο με τους εντοπισμένους ρόλους. Τα χρώματα αντιπροσωπεύουν ρόλους και το μέγεθος των κόμβων βαθμό. Η απεικόνιση έγινε σε κοινότητες για λόγους παρουσίασης



Στη συνέχεια θα επικεντρωθούμε σε περιοχές του γράφου που παρουσιάζουν ενδιαφέρον, προκειμένου να εξηγήσουμε σημασιολογικά μερικούς από τους ρόλους.

Σταγμιότυπο Α



Εδώ δίνουμε έμφαση στους γαλάζιους κόμβους, που αντιπροσωπεύουν τον ρόλο 29. Δεν είναι δύσκολο να καταλάβουμε ότι αποτελούν τις ακτίνες αστέρα γύρω από τον κεντρικό κόμβο 288 που έχει τον ρόλο 7, ο οποίος αντιπροσωπεύει το κέντρο του αστέρα

Κύρια χαρακτηριστικά ρόλου 29:

xedt0	xedu0-0-m0-1-s0-2-s0	xedu0-0-m0-1-s0-2-m0	wn1-0-m0-1-m0-2-s0	wn1-0-m0

Κύρια χαρακτηριστικά ρόλου 7:

weu0-0-s0-1-m0-2-s0	weu0-0-s0-1-s0	xedu1-0-s0-1-s0-2-s0	wet0-0-s0-1-m0-2-m0-3-s0	wet0-0-s0-1-m0-2-m0-3-s0

Μπορούμε όμως να εντοπίσουμε και κάποιες φαινομενικές ανωμαλίες:

Τα καφέ πλαίσια περιέχουν κόμβους που είναι και οι ίδιοι κέντρα αστέρων. Αυτό συμβαίνει διότι έχουμε κρατήσει μόνο τους κυρίαρχους ρόλους κάθε κόμβου. Έτσι παρόλο που αλγόριθμος έχει αναθέσει παραπάνω από έναν ρόλο στους κόμβους αυτούς, ως πρωτεύοντας ρόλος φαίνεται ο ρόλος 29

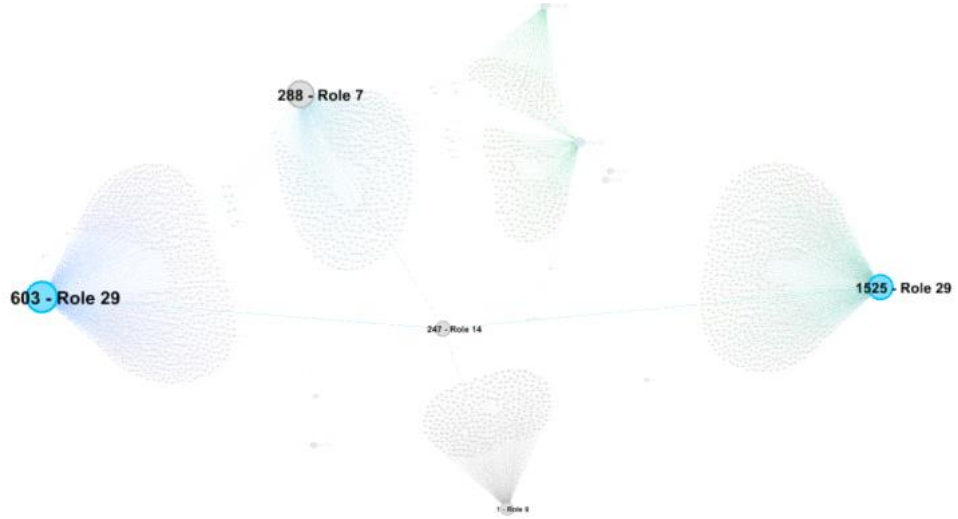
Τα πράσινα πλαίσια περιλαμβάνουν κόμβους γέφυρες, πάραυτα κάποιοι φαίνονται να ανήκουν στον ρόλο 29, καθώς δεν εμφανίζονται στην εικόνα δευτερεύοντες ρόλοι..

Ο κόμβος 247, παρόλο που είναι και αυτός μια από τις ακτίνες του αστέρα, ανήκει στον ρόλο 14, καθώς αντιπροσωπεύει το κέντρο ενός άλλου μικρότερου αστέρα, όπως θα δούμε στο στιγμιότυπο B

Κύρια χαρακτηριστικά ρόλου 14:

xedu0	wet1-0-m0-1-s0-2-m0	xedt0-0-m0	xedu0-0-s0-1-m0	xedt0-0-m0-1-s0

Στιγμιότυπο Β



Παρατηρούμε τον κόμβο 247 από πριν ότι όχι μόνο είναι κέντρο μικρότερου αστέρα, αλλά αποτελεί και κεντρικό κόμβο που συνδέει μεταξύ τους τέσσερεις πολύ μεγάλες κοινότητες. Θα μπορούσαμε να πούμε ότι αποτελεί κέντρο του γράφου κατά κάποιο τρόπο.

Κύρια χαρακτηριστικά ρόλου 14:

xedu0	wet1-0-m0-1-s0-2-m0	xedt0-0-m0	xedu0-0-s0-1-m0	xedt0-0-m0-1-s0

Στιγμιότυπο Γ

Εδώ βλέπουμε μία κοινότητα η οποία περιλαμβάνει δύο κέντρα. Έχουμε ένα κέντρο τον ρόλο 6, δεύτερο κέντρο με ρόλο 41, ίδιο με τις ακτίνες του κέντρου. Από τα μέλη της κοινότητας έχουμε αυτά που ανήκουν στο ένα κέντρο με χρώμα πορτοκαλί (ρόλος 41), αυτά που ανήκουν στο άλλο κέντρο και έχουν πάλι ρόλο 41 όπως είναι λογικό και τέλος αυτά που ανήκουν και στα δύο κέντρα και αποτελούν γέφυρες, τα οποία έχουν ρόλο 10. Παρατηρώντας καλύτερα το δίκτυο (δεν φαίνεται στην εικόνα, βλέπουμε ότι το κέντρο με ρόλο 6 συνδέεται με το υπόλοιπο δίκτυο μέσω ενός άλλου κόμβου γέφυρας κάτι που δεν συμβαίνει στο κέντρο με ρόλο 41

Κύρια χαρακτηριστικά ρόλου 6:

xedu1-0-m0	xedu1-0-m0-1-s0-2-m0	xedu1-0-m0-1-m0-2-m0	xedu1-0-m0-1-s0-2-m0-3-m0	xedu1-0-m0-1-s0-2-s0

Κύρια χαρακτηριστικά ρόλου 36:

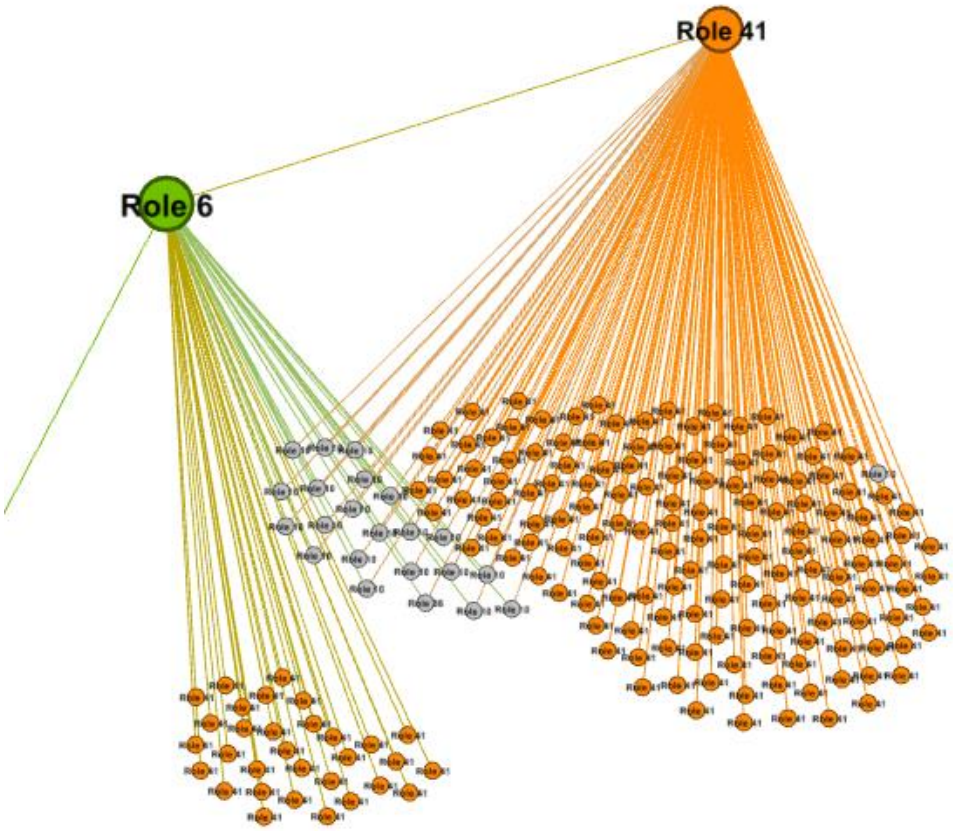
wet0-0-m0-1-s0	wet0	wet0-0-s0-1-s0	wet0-0-m0-1-m0-2-m0-3-s0	weu0-0-m0-1-s0

Κύρια χαρακτηριστικά ρόλου 41:

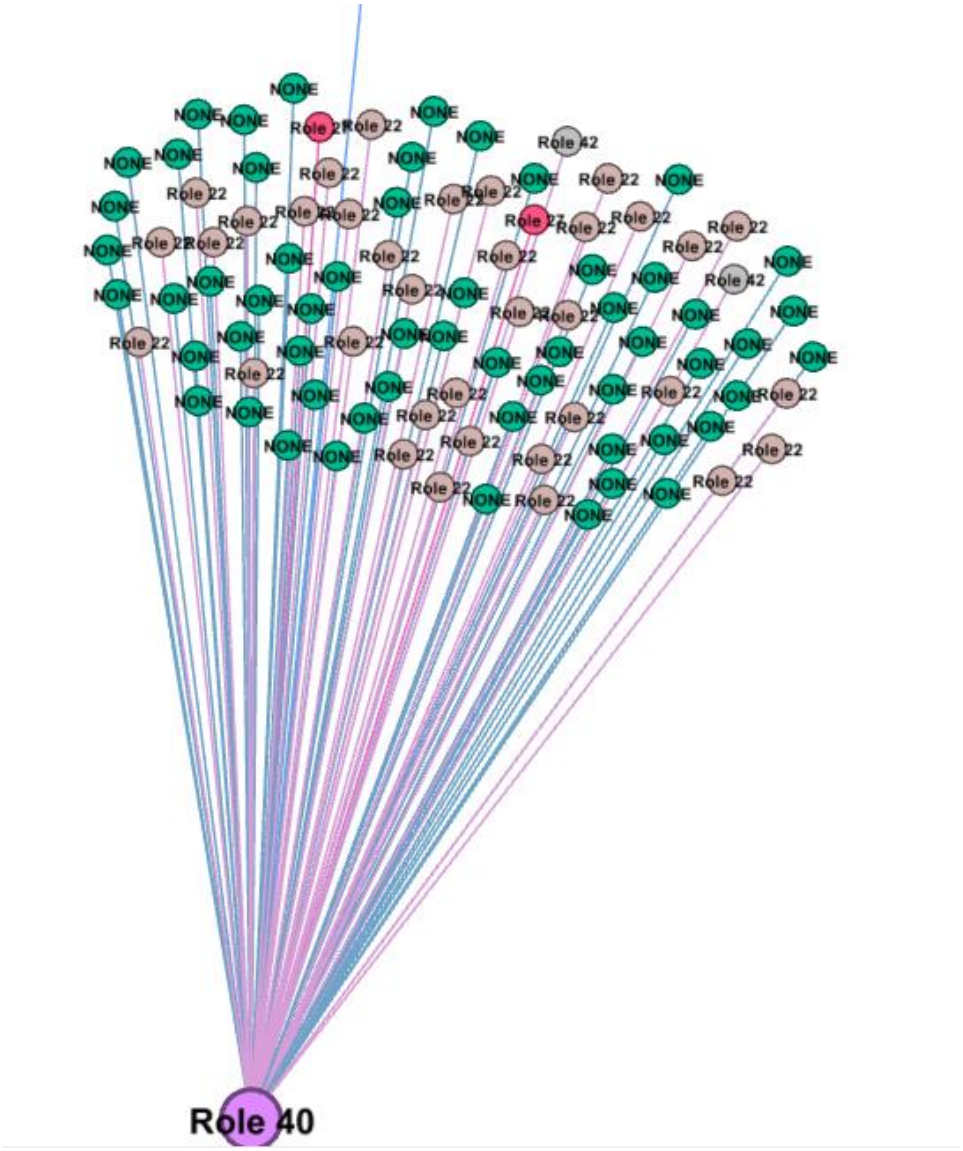
wn0-0-m0-1-m0-2-s0-3-m0	wn0-0-m0-1-m0-2-m0-3-m0	wn0-0-m0-1-m0-2-s0	wn0-0-m0-1-m0-2-m0-3-s0	wn0-0-m0-1-m0-2-m0

Κύρια χαρακτηριστικά ρόλου 10:

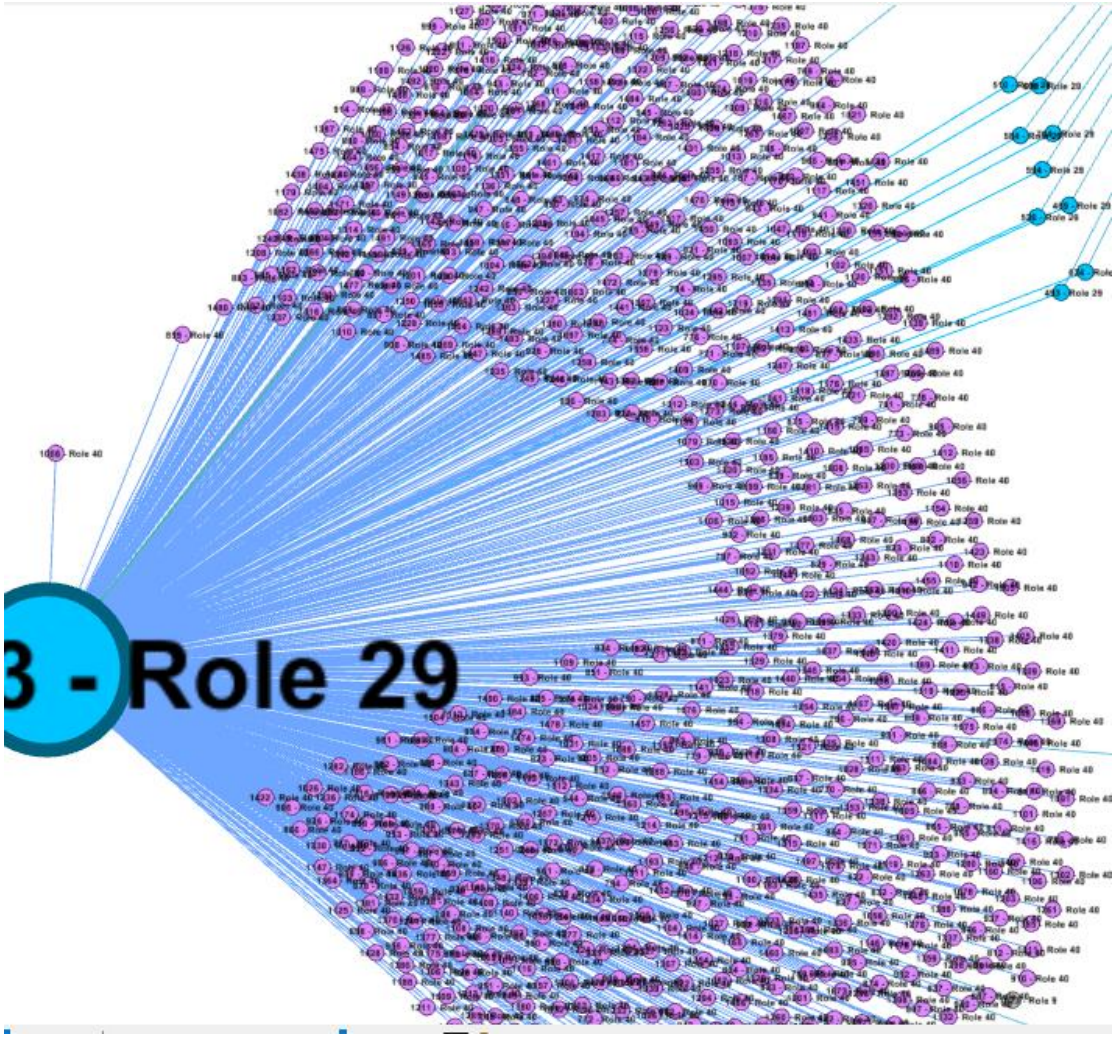
weu0-0-s0-1-m0	weu0-0-s0-1-m0-2-m0	weu0-0-m0	wn0-0-m0-1-s0-2-m0-3-m0	weu0-0-s0-1-s0



Στιγμιότυπο Δ



Οι εικόνες στο στιγμιότυπο αυτό επιλέχθηκαν για να δείξουν είτε σφάλματα του αλγόριθμου, είτε αδυναμία εξήγησης από ανθρώπινο μάτι. Περιπτώσεις σαν και αυτές υποδεικνύουν την ανάγκη ύπαρξης ενός ειδικού πάνω στο πεδίο που ανήκει το δίκτυο, για την πιο έγκυρη σημασιολογία των εντοπισμένων ρόλων.



Και στις δύο εικόνες, οι μωβ κόμβοι αντιπροσωπεύουν τον ρόλο 40. Μόνο που η λειτουργικότητα των κόμβων μέσα στο δίκτυο δείχνει πολύ διαφορετική στις δύο εικόνες. Στην πρώτη ο κόμβος φαίνεται να αντιπροσωπεύει κόμβο κάποιου αστέρα, ενώ στη δεύτερη ακτίνες.

Κύρια χαρακτηριστικά ρόλου 40:

wet1-0-m0	xedu0-0-m0-1-m0	wet1-0-m0-1-m0	xedu0-0-m0	wet1-0-m0-1-s0-2-m0

Στην πρώτη εικόνα όλοι οι κόμβοι φαίνονται απλές ακτίνες ενός αστέρα. Δεν φαίνεται να εξηγείται κάπως ο διαχωρισμός τους σε πολλαπλούς ρόλους. Ούτε το γεγονός ότι κάποιοι κόμβοι δεν έχουν ρόλο.

Κύρια χαρακτηριστικά ρόλου 22:

xedu1-0-s0-1-m0-2-m0-3-m0	xedu1-0-m0-1-m0-2-m0-3-m0	xedu1	xedu1-0-s0-1-m0-2-m0	xedu1-0-s0-1-m0-2-m0-3-s0

Κύρια χαρακτηριστικά ρόλου 27:

wet0-0-s0-1-m0-2-m0-3-s0	weu0-0-s0-1-s0	weu0-0-s0-1-m0-2-m0	weu0-0-s0-1-s0-2-m0-3-m0

Στην προσπάθειά μας να μελετήσουμε καλύτερα την σημασιολογία του ρόλου, πραγματοποιήσαμε ανάλυση των κόμβων μέσα από το Gephi.

Με αφορμή το προηγούμενο στιγμιότυπο, απομονώσαμε τους κόμβους με ρόλο 40 και η εικόνα είναι η εξής:

Id	**Eccentricity**	**closness centrality**	**harmonic closness centrality**	**betweenesscentrality**	**modularity_class**	**Primary-Role**
767	6	0.299295	0.324454	0	2	Role 40
2232	7	0.237691	0.26822	272496	7	Role 40
768	6	0.291764	0.324518	0	4	Role 40
769	6	0.291764	0.324518	0	4	Role 40
770	6	0.291764	0.324518	0	4	Role 40

Οι υπόλοιποι κόμβοι του ρόλου 40 έχουν ίδιες μετρικές με αυτές του κόμβου 770. Παρατηρούμε ότι με εξαίρεση τους δύο πρώτους κόμβους, όλοι οι υπόλοιποι έχουν ίδιες τιμές στα centralities.

Πραγματοποιούμε ίδια ανάλυση για τον ρόλο 29 και τον ρόλο 22:

Id	**Eccentricity**	**closness centrality**	**harmonic closnesscentrality**	**betweeness centrality**	**modularity_class**	**Primary-Role**
1525	6	0.356684	0.479442	1789042.55	3	Role 29
603	5	0.411899	0.528484	2290045.033	4	Role 29
710	6	0.345418	0.401501	530103.075	5	Role 29
714	6	0.344429	0.389002	469785.025	5	Role 29
289	6	0.299295	0.324454	0	2	Role 29

290	6	0.299295	0.324454	0	2	Role 29
291	6	0.299295	0.324454	0	2	Role 29
292	6	0.299295	0.324454	0	2	Role 29
293	6	0.299295	0.324454	0	2	Role 29
294	6	0.299295	0.324454	0	2	Role 29
295	6	0.299295	0.324454	0	2	Role 29

Id	**Eccentricity**	**closness centrality**	**harmonic closness centrality**	**betweeness centrality**	**modularity_class**	**Primary-Role**
2294	8	0.192057	0.204422	0	7	Role 22
2295	8	0.192057	0.204422	0	7	Role 22
2296	8	0.192057	0.204422	0	7	Role 22
2297	8	0.192057	0.204422	0	7	Role 22
2298	8	0.192057	0.204422	0	7	Role 22
2299	8	0.192057	0.204422	0	7	Role 22
2300	8	0.192057	0.204422	0	7	Role 22
2301	8	0.192057	0.204422	0	7	Role 22
2302	8	0.192057	0.204422	0	7	Role 22
2303	8	0.192057	0.204422	0	7	Role 22
2304	8	0.192057	0.204422	0	7	Role 22
2305	8	0.192057	0.204422	0	7	Role 22
2306	8	0.192057	0.204422	0	7	Role 22
2307	8	0.192057	0.204422	0	7	Role 22

Συμπεραίνουμε ότι υπάρχει σαφής συσχέτιση των μετρικών centralities με την ανάθεση ρόλου. Σε καμία περίπτωση όμως δεν έχουμε αντιστοιχία ένα προς έναν ή ταύτιση.

5.3 Δεύτερο πείραμα

5.3.1 Πληροφορίες για τα δεδομένα

Το δίκτυο που αντιπροσωπεύουν τα δεδομένα στο δεύτερο πείραμα αφορά ξανά ένα δίκτυο φιλίας στο Facebook. Σε αυτήν την περίπτωση η πηγή είναι η εξής: https://snap.stanford.edu/data/egonets-Facebook.html

Τα δεδομένα δίνονταν ως ένας συνδυασμός δέκα μικρότερων δικτύων. Δίδονταν και ένα σύνολο από διαθέσιμα χαρακτηριστικά τα οποία αποφασίσαμε να μη χρησιμοποιήσουμε, ώστε να κρατηθεί η πολυπλοκότητα της ανάλυσης σε λογικά επίπεδα

Πριν την επεξεργασία των δεδομένων τα στατιστικά που μας δίδονταν από την πηγή ήταν αυτά που εμφανίζονται στην επόμενη εικόνα.

Dataset statistics	
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

Στατιστικά πηγής

Με τη βοήθεια του Gephi υπολογίσαμε και εμείς μια σειρά από στατιστικά στοιχεία, τα οποία φαίνονται παρακάτω και συμφωνούν με αυτά της πηγής.

Πλήθος κόμβων	4039
Πλήθος ακμών	88234
Κατευθυνόμενος	Όχι
Μέσος βαθμός	43,691
Διάμετρος δικτύου	8
Modularity	0,835
Μέσο μήκος μονοπατιού	3,693

Στατιστικά στοιχεί δεδομένων δεύτερου πειράματος. Η εξαγωγή τους έγινε με το πρόγραμμα Gephi.

Σε αντίθεση με το πρώτο πείραμα, εδώ έχουμε ένα αρκετά πυκνό δίκτυο. Ο λόγος του πλήθους των ακμών προς τους κόμβους είναι αρκετά μεγάλος. Επιλέξαμε αυτόν τον γράφο για να δείξουμε τη δυσκολία ερμηνείας των αποτελεσμάτων από ανθρώπους, έστω και ειδικούς, αλλά και για να τονίσουμε τη μεγάλη χρησιμότητα του αλγορίθμου όταν αποτελέσματα ερμηνεύονται από κάποιο υπολογιστικό πρόγραμμα σε επόμενο στάδιο. Ο γράφος και σε αυτή την περίπτωση είναι μη κατευθυνόμενος, καθώς αντιπροσωπεύει φιλίες στο Facebook.

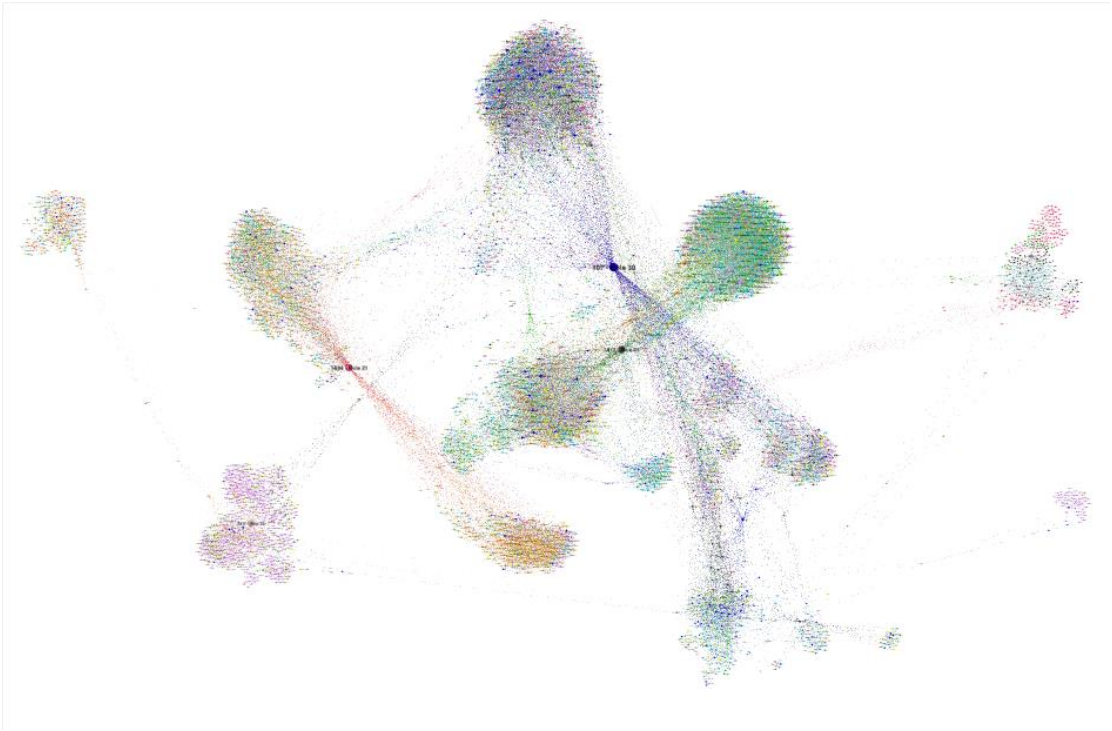
Όπως προαναφέρθηκε το πρόβλημα που αναμένουμε σε αυτό το πείραμα είναι η αδυναμία ερμηνίας των αποτελεσμάτων, άρα και η δυσκολία εξαγωγής συμπερασμάτων.

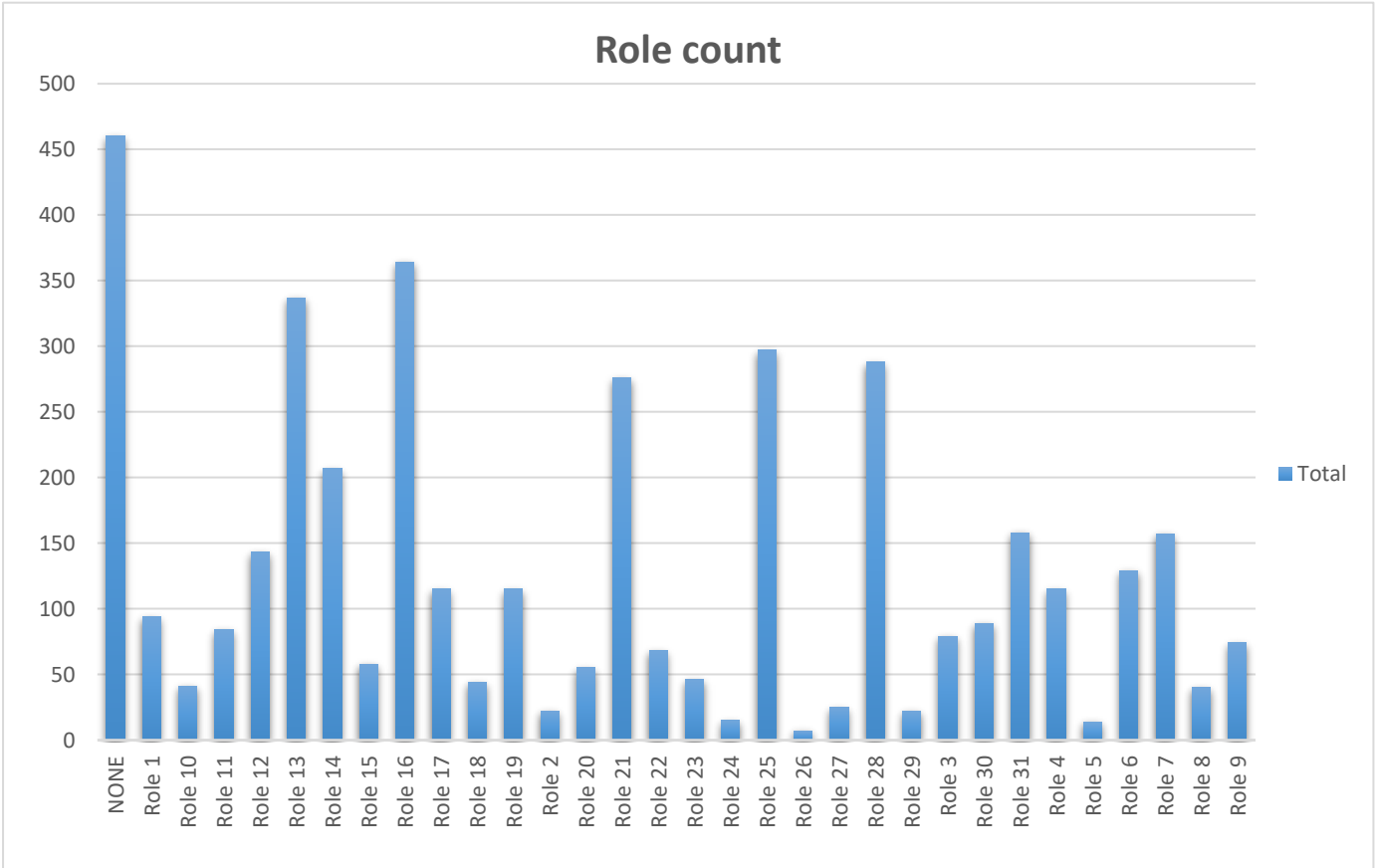
Τα αρχικά χαρακτηριστικά και οι τελεστές είναι ακριβώς ίδια με το προηγούμενο πείραμα, οπότε δεν θα υπάρξει ανάλυση για αυτά. Προχωράμε απευθείας στα αποτελέσματα.

5.3.2 Αποτελέσματα

Εντοπίστηκαν 31 ρόλοι μέσα από 89 χαρακτηριστικά. Και σε αυτήν την εκτέλεση κάθε κόμβος συμμετέχει σε πολλαπλούς ρόλους και κάθε ρόλος σε διαφορετικά χαρακτηριστικά. Εμείς κρατήσαμε μόνο τον κυρίαρχο ρόλο κάθε κόμβου, προκειμένου να παρέχουμε μια βασική ανάλυση. Σε αντίθετη περίπτωση, έστω και η στοιχειώδης ερμηνεία των αποτελεσμάτων θα ήταν αδύνατη από ανθρώπινο μάτι.

Στις επόμενες σελίδες παραθέτουμε το δίκτυο με χρωματισμένους τους ρόλους, την κατανομή των κυρίαρχων ρόλων, καθώς και μερικά ενδιαφέροντα στιγμιότυπα.

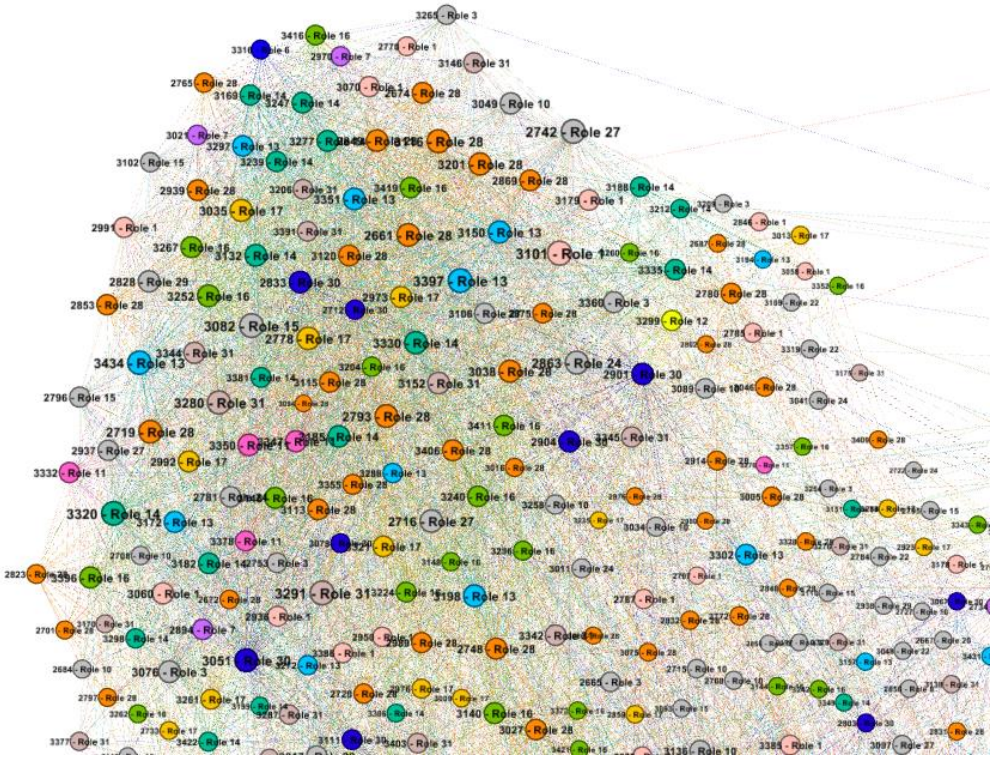



Role count
500
450
400
350
300
250
200
150
100
50
0
NONE
Role 1
Role 10
Role 11
Role 12
Role 13
Role 14
Role 15
Role 16
Role 17
Role 18
Role 19
Role 2
Role 20
Role 21
Role 22
Role 23
Role 24
Role 25
Role 26
Role 27
Role 28
Role 29
Role 3
Role 30
Role 31
Role 4
Role 5
Role 6
Role 7
Role 8
Role 9
Total

Από τις παραπάνω εικόνες δύο πράγματα μας κάνουν εντύπωση. Αρχικά δυσκολία στην ανάγνωση του γράφου και την παρατήρηση των ακμών και στη συνέχεια το μεγάλο πλήθος κόμβων που δεν έχουν ανατεθεί σε κανέναν ρόλο.

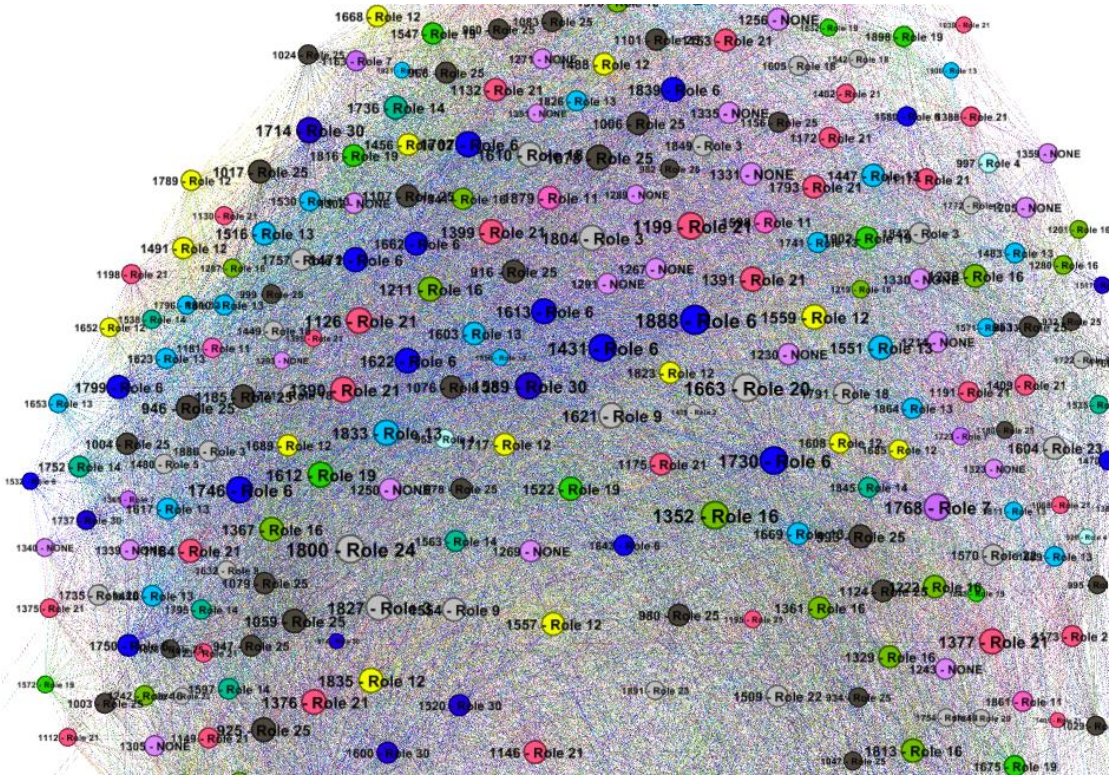
Ας δούμε μερικά στιγμιότυπα του γράφου:

Στιγμιότυπο Α



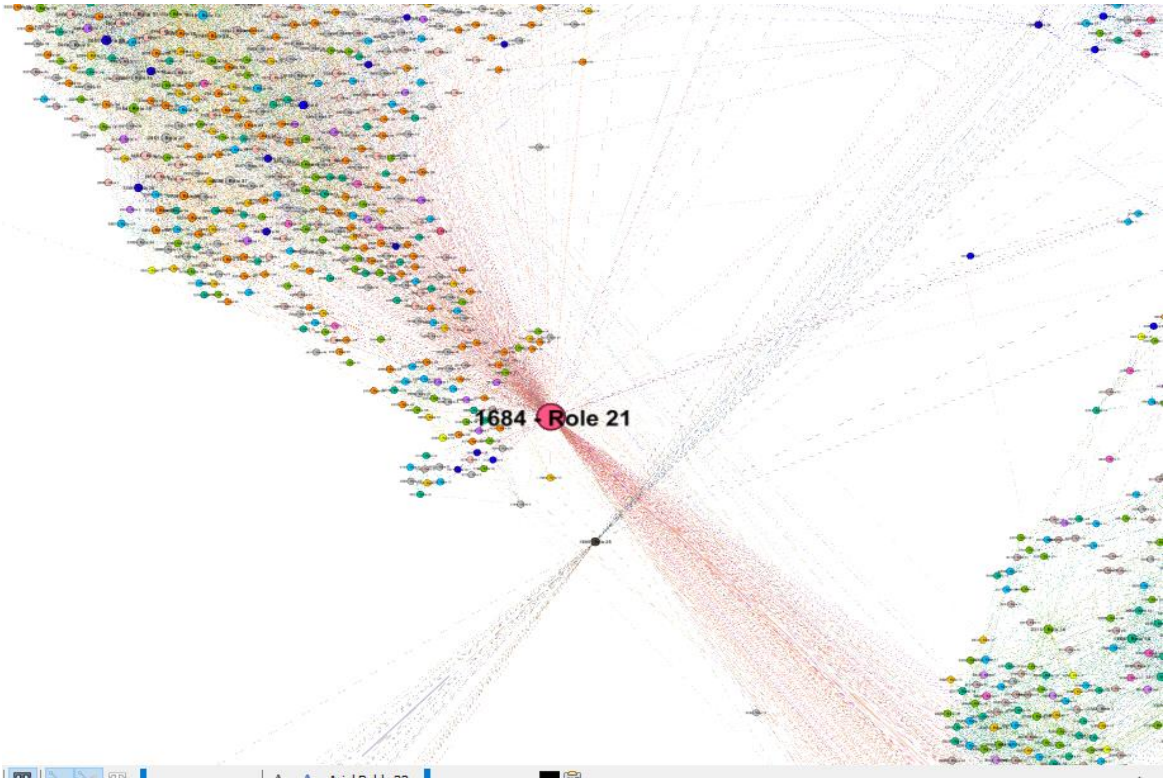
Αν και παρατηρούμε ένα τεράστιο πλήθος κόμβων, είναι αδύνατο να διακρίνουμε τη λειτουργικότητά τους. Το μέγεθος των κόμβων είναι ανάλογο του βαθμού τους. Μπορούμε λοιπόν να συμπεράνουμε με σχετική ασφάλεια, ότι ίδιοι ρόλοι (ίδια χρώματα), συνεπάγονται και παρόμοιο βαθμό.

Στιγμιότυπο Β



Και σε αυτήν την περίπτωση διαπιστώνουμε μια συσχέτιση χρώματος και μεγέθους, παρόλα αυτά δεν μπορούμε να βγάλουμε κάποιο άλλο συμπέρασμα για τους ρόλους. Να δώσουμε έμφαση μόνο στον ρόλο 21 (φούξια(, διότι θα μας χρειαστεί στο επόμενο στιγμιότυπο.

Στιγμιότυπο Γ



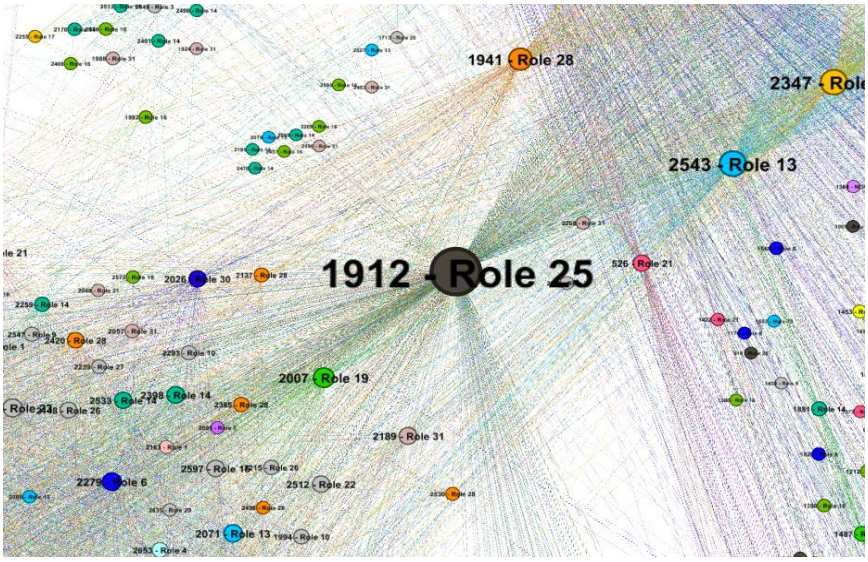
Παρατηρούμε έναν ακόμη κόμβο με ρόλο 21, ο οποίος όμως δείχνει πολύ πιο κεντρικός από αυτούς στο προηγούμενο στιγμιότυπο. Το σημείο αυτό είναι χρήσιμο ώστε να ελέγξουμε σε μια πιο ακραία κατάσταση την υπόθεσή που κάναμε για τα centralities στο πρώτο πείραμα.

Μελετάμε λοιπόν κάποιους από τους παραπάνω κόμβους με τη βοήθεια του Gpehi:

Id	**Role**	**Degree**	**Eccentricity**	**closnesscentrality**	**harmonicclosnesscentrality**	**betweenescentrality**	**modularity_class**
1684	21	792	5	0.393606	0.491559	2753286.687	9
1191	21	110	6	0.318178	0.358672	167.443532	4
1376	21	182	6	0.338219	0.383808	6447.401493	4
1399	21	178	6	0.319361	0.366225	382.337241	4

Παρατηρούμε ότι ο κόμβος στην δεύτερη εικόνα έχει μικρή διαφορά στο closness centrality σε σχέση με αυτούς στη δεύτερη, αλλά οι τιμές τους στις άλλες μετρικές είναι τελείως διαφορετικές.

Στιγμιότυπο Δ



Παρατηρούμε έναν πολύ κεντρικό κόμβο στο γράφημα.

Τα κύρια χαρακτηριστικά του ρόλου 25 είναι τα παρακάτω:

xedu1-0-m0-1-s0-2-m0-3-m0-4-m0-5-m0	xedu1-0-m0-1-s0-2-m0-3-m0-4-m0	xedu1-0-m0-1-s0-2-m0-3-m0	xet1-0-m0-1-m0	xedu1-0-m0-1-m0

Η ανάλυση του gephi για τον κόμβο 1912 μαζί με άλλους με τον ίδιο ρόλο δίνει τα εξής:

Id	**Role**	**Degree**	**Eccentricity**	**closnesscentrality**	**harmonicclosnesscentrality**	**betweenesscentrality**	**modularity_class**
16	25	9	7	0.261241	0.284155	1.628466	0
27	25	5	7	0.261173	0.28366	0.25	0
29	25	13	7	0.261308	0.284651	11.374158	0
30	25	17	7	0.261376	0.285146	11.162084	0
1912	25	755	6	0.350947	0.447924	1868918.212	10

Δεν παρατηρούμε πειστική συσχέτιση στα centralities για τη συγκεκριμένη περίπτωση.

5.4 Συμπεράσματα

Ύστερα από την ανάλυση, το πρώτο συμπέρασμα που προκύπτει είναι η δυσκολία της εξαγωγής μετρήσιμων συμπερασμάτων. Είναι πρακτικά αδύνατο να λάβουμε υπόψη όλους τους ρόλους στους οποίους μπορεί να συμμετέχει ένας κόμβος, αφού και σε ένα μικρό γράφο των 3000 κόμβων εμφανίζονται πάνω από 30 ρόλοι. Ακόμα και αν καταφέρουμε να ξεπεράσουμε το εμπόδιο αυτό, μέσω μιας προσπάθειας οπτικοποίησης των πιο σημαντικών αποτελεσμάτων, η πυκνότητα του γράφου, αλλά και η δυσκολία αντιστοίχισης ρόλων και κατασκευαστικών χαρακτηριστικών, υπονομεύουν τη δυνατότητα εξαγωγής αξιόπιστων συμπερασμάτων.

Πιο συγκεκριμένα, προσπαθήσαμε να οπτικοποιήσουμε και να ερμηνεύσουμε τα σημαντικότερα δεδομένα των αποτελεσμάτων με τους εξής τρόπους:

- Χρωματισμός των ρόλων
- Θέση των ρόλων μέσα στον γράφο
- Προσπάθεια ανάλυσης των κατασκευαστικών χαρακτηριστικών κάθε ρόλου
- Προσπάθεια αντιστοίχισης μετρικών της κεντρικότητας ενός κόμβου με τον ρόλο του
- Αντιστοιχία βαθμού του κόμβου

Καμία από τις παραπάνω προσπάθειες δεν απέδειξε κάτι σαφές, αλλά έδωσε ενδείξεις για την ύπαρξη πρακτικού νοήματος κρυμμένη πίσω από τους ρόλους. Πιθανόν η συνεισφορά κάποιου ειδικού στα δίκτυα του facebook, αλλά και η ύπαρξη κοινωνιολογικής γνώσης γύρω από το τι αντιπροσωπεύει ο κάθε κόμβος, να βοηθούσε στην εξαγωγή περισσότερων συμπερασμάτων.

Σε αυτό το σημείο να αναφέρουμε ότι τρέξαμε τον αλγόριθμο και δεύτερη φορά πάνω στα ίδια δεδομένα. Αν και ο τελικός αριθμός των κόμβων προέκυψε ελαφρώς διαφορετικός, κάτι το οποίο δικαιολογείται λόγω των στρογγυλοποιήσεων που συμβαίνουν από τα μαθηματικά εργαλεία, η συνολική εικόνα των ρόλων του γράφου, παρέμεινε ίδια ύστερα από τις προσπάθειες οπτικοποίησης. Το γεγονός αυτό μας δείχνει, ότι παρά τις δυσκολίες ερμηνείας των αποτελεσμάτων, ο αλγόριθμος παρέχει μια επαρκή σταθερότητα, ώστε να θεωρήσουμε ότι δίνει χρήσιμες για τη συμπεριφορά ενός δικτύου πληροφορίες, οι οποίες όμως θα πρέπει να επεξεργαστούν από κάποιον άλλον αλγόριθμο ως προς την ερμηνεία τους. Δεν είναι τυχαίο το ότι κάτι αντίστοιχο συμβαίνει τελικά και στην πράξη.

Τέλος σημαντικό στοιχείο αποτελεί και ο χρόνος εκτέλεσης του αλγορίθμου. Στο πρώτο πείραμα χρειάστηκαν περίπου 20 λεπτά, ενώ στο δεύτερο 3 ώρες.

Η αδυναμία μας να αποδείξουμε τη χρησιμότητά του αλγορίθμου δεν αναιρεί την ύπαρξη πολλών πρακτικών εφαρμογών στις οποίες ο αλγόριθμος χρησιμοποιείται σαν ενδιάμεσο στάδιο και το μεγάλο ερευνητικό έργο γύρω από αυτόν. Σε κάποιον που επιθυμεί να κατανοήσει στην πράξη τα πλεονέκτηματά του αλγορίθμου θα προτείναμε την εφαρμογή του σε κάποιο μικρό και προφανές ως προς τους ρόλους δίκτυο, ώστε να μπορεί να επιβεβαιώσει την εγκυρότητά του.

Στο επόμενο και τελευταίο κεφάλαιο θα δείξουμε πρακτικές εφαρμογές του αλγορίθμου και πιθανές βελτιώσεις της διαδικασίας.

Κεφάλαιο 6 - Πρακτικές εφαρμογές και βελτιώσεις

6.1 Εισαγωγή

Ύστερα από την εξήγηση του θεωρητικού μέρους, την ανάλυση του αλγόριθμου και την εκτέλεση του, θεωρούμε σκόπιμο να δείξουμε πιθανές βελτιώσεις του αλγορίθμου, αλλά και πρακτικές εφαρμογές στις οποίες ο Refex και ο Rolx αποτελούν το κύριο εργαλείο τους. Δεν θα αναφέρουμε λεπτομέρειες, καθώς δεν αποτελεί αντικείμενο αυτής της εργασίας, με μια εξαίρεση για τα recommender systems, τα οποία θα αναλυθούν περισσότερο για λόγους πληρότητας.

6.2 Βελτιώσεις

Στην υποενότητα αυτή θα αναφέρουμε βελτιώσεις της προσέγγισης που αναλύσαμε.

6.2.1 Αριθμός ρόλων

Και στα δύο σετ εντοπίσαμε έναν πολύ μεγάλο αριθμό ρόλων το οποίο τελικά δυσκόλεψε πολύ την ανάλυση μας. Θα ήταν πιο χρήσιμη η εφαρμογή του αλγορίθμου σε ομογενείς γράφους, ώστε να περιοριστούν οι ρόλοι, ή σε μικρότερα egonets. Μία διαφορετική προσέγγιση είναι με τη βοήθεια κάποιου ειδικού για την συγκεκριμένη ομάδα δικτύων, να καθοριστεί από πριν το πλήθος των δικτύων. Σκοπός εδώ είναι να αποφύγουμε την τυχαιότητα και την ανακρίβεια λόγω στρογγυλοποιήσεων της MDL.

6.2.2 Όγκος δεδομένων

Η εφαρμογή δεν είναι εύκολα εφαρμόσιμη για πάρα πολύ μεγάλο πλήθος ακμών, εξαιτίας της χρονικής πολυπλοκότητας αλλά και της δυσκολίας στην ερμηνεία των αποτελεσμάτων.

Για την αντιμετώπιση του πρώτου προβλήματος υπάρχει μια διαφορετική προσέγγιση στη διευκόλυνση εξαγωγής ρόλων. Οι κόμβοι πριν την εξαγωγή των ρόλων διαμερίζονται σε κατασκευαστικά ισοδύναμα σύνολα μέσα από μια σειρά αλγορίθμων που φροντίζουν οι κόμβοι κάθε συνόλου να έχουν μεταξύ τους όμοια μια συγκεκριμένη μαθηματική μετρική. Η διαδικασία αυτή μπορεί να παραλληλοποιηθεί μέσω της μεθόδου map – reduce. Όλη η προσέγγιση οδηγεί σε ταχύτερη εξαγωγή ρόλων. Για περισσότερες πληροφορίες ανατρέξτε στη βιβλιογραφία.

Map – reduce

Το map – reduce αποτελεί ένα προγραμματιστικό μοντέλο για την επεξεργασία big – data με παράλληλο τρόπο.

Αποτελείται από μία διαδικασία Map() η οποία φιλτράρει, ταξινομεί και οδηγεί τα δεδομένα σε ουρές και μία διαδικασία reduce() σε κάθε ουρά η οποία πραγματοποιεί την επεξεργασία των δεομένων.

6.3 Πρακτικές εφαρμογές

6.3.1 Πρόβλεψη συμπεριφοράς και συνδέσμων

Φανταστείτε να εκτελούμε τον αλγόριθμο ανά συγκεκριμένα χρονικά διαστήματα. Έστω ότι τον τρέξαμε t φορές, θα έχουμε τότε t πίνακες G_{nxr}. Μπορούμε λοιπόν στη συνέχεια να υπολογίσουμε έναν πίνακα μετάβασης T από το G_{t-1} στο G_t, όπου $G_{t-1} \times T = G_t$ και να τον χρησιμοποιήσουμε για την εξαγωγή της πιθανότητας ενός κόμβου να μεταβεί από κάποιον ρόλο σε έναν άλλο. Επεκτείνοντας τη λογική αυτή για πολλά στιγμιότυπα μπορεί να υπολογιστεί ένας πίνακας μετάβασης T με την χρήση πολλών στιγμιότυπων, για παράδειγμα:

$$\begin{bmatrix} Gt-1 \\ Gt-2 \\ . \\ Gk-1 \end{bmatrix}_{T=} \begin{bmatrix} Gt \\ Gt-1 \\ . \\ Gk \end{bmatrix}$$

Η διαφορετικά με μία συνάρτηση μα βάρη, ώστε παλιότερα στιγμιότυπα να έχουν μικρότερη επιρροή στην πρόβλεψη. Όταν γνωρίζουμε τη μετάβαση, αλλά και τις ιδιότητες κάθε ρόλου, μπορούμε να προβλέψουμε συμπεριφορά, αλλά και πιθανούς συνδέσμους που ίσως δημιουργηθούν.

6.3.2 Εντοπισμός ανωμαλιών δικτύου και αναγνώριση επιθέσεων.

Μέσα από την τακτική εφαρμογή του αλγορίθμου σε δίκτυο συνδέσεων IP, μπορούμε να αναγνωρίσουμε το μοτίβο μετάβασης για συγκεκριμένες ώρες τις ημέρας. Σε περίπτωση που εμφανιστεί κάτι πολύ διαφορετικό από τα αναμενόμενα, σημαίνει ότι κάτι ασυνήθιστο συμβαίνει στο δίκτυο. Για παράδειγμα μπορεί να αναγνωριστεί η κακή λειτουργία κάποιου switch, ή μία επίθεση DDOS.

6.3.3 Ερευνητικά πακέτα λογισμικού ανάλυσης δικτύου

Οι Rolx και Refex βρίσκονται στα πακέτα ανάλυσης δικτύου του Stanford(snap) και του MIT.

6.3.4 Recommender Systems

Βάση της πρόβλεψης μετάβασης των κόμβων μπορούμε σε ένα πελατο-κεντρικό δίκτυο, να προτείνουμε στους πελάτες προϊόντα τα οποία έχουν αγοράσει πελάτες που ανήκουν στον ρόλο που πρόκειται να μεταβεί ο χρήστης.

Σε μια άλλη κοινωνιολογική προσέγγιση, ο αλγόριθμος μπορεί να χρησιμοποιηθεί για την πρόταση συνεργασιών, ή φίλων, ή δραστηριοτήτων.

Για λόγους πληρότητας της εργασίας, θα πραγματοποιηθεί μια σκιαγράφηση των recommender systems, έτσι όπως αναλύονται στην εργασία των M. Eirinaki, J. Gao, I. Varlamis και K. Tserpes.

Recommender Systems for Large-Scale Social Networks: A review of challenges and solutions

Τα recommender systems ανήκουν στα επιστημονικά πεδία της υπολογιστικής νοημοσύνης και της διαχείρισης γνώσης. Η ικανότητα δημιουργίας γνώσης και νοημοσύνης μέσα από την ανάλυση δεδομένων,

εφαρμόζεται εδώ και αρκετά χρόνια στις επιχειρήσεις, τη βιομηχανία, την επιστήμη και τα κοινωνικά δίκτυα. Ξεχωρίζει η σύνδεση των recommender systems με τα κοινωνικά δίκτυα, καθώς η κοινωνική επιρροή θεωρείται ιδιαίτερα σημαντική για το μάρκετινγκ προϊόντων, αλλά και τα κοινωνικά δίκτυα προωθούν τη βελτίωση της εμπειρίας των χρηστών τους, μέσα από την προβολή στοχευμένου και προσωποποιημένου περιεχομένου σε αυτούς.

Μπορούμε να διακρίνουμε δύο βασικές προσεγγίσεις στην δημιουργία recommender systems, το φιλτράρισμα με βάση το περιεχόμενο (content-based filtering - CB) και το συνεργατικό φιλτράρισμα (collaborative filtering - CF).

Γενικά στα συστήματα αυτά, κάθε χρήστης αντιπροσωπεύεται από ένα προφίλ, στο οποίο περιλαμβάνονται όλα τα αντικείμενα τα οποία έχει αξιολογήσει ή αγοράσει. Στο CB, νέα αντικείμενα προτείνονται στον χρήστη, με βάση τα αντικείμενα που υπάρχουν ήδη στο προφίλ, ενώ στο CF ο αλγόριθμος λειτουργεί ανεξάρτητα από τα αντικείμενα, αφού χρησιμοποιούνται οι αξιολογήσεις και οι αγορές του προφίλ, ώστε να εντοπιστούν άλλοι χρήστες με παρόμοια χαρακτηριστικά και να χρησιμοποιηθούν οι επιλογές των δεύτερων, ως προτάσεις στους πρώτους.

Με την εύκολη πλέον πρόσβαση σε πληθώρα δεδομένων, είναι φανερό ότι δομές, όπως ένας διημερής γράφος από χρήστες και αντικείμενα, δεν είναι πια ικανές να αναπαραστήσουν όλη τη διαθέσιμη πληροφορία, όπως το περιεχόμενο, το γενικότερο πλαίσιο, την κοινωνική πληροφορία και τα μεταδεδομένα. Έτσι οι συγγραφείς της μελέτης αναλύουν προσπάθειες που έχουν γίνει για τον εμπλουτισμό της αναπαράστασης. Έχουμε λοιπόν συστήματα που προσπαθούν να προσεγγίσουν το γενικότερο πλαίσιο ενός χρήστη, συστήματα που προσαρμόζονται στη χρονική εξέλιξη, συστήματα βασισμένα στην τοποθεσία και συστήματα προσαρμοσμένα στην κοινότητα.

Τα συστήματα προσέγγισης του γενικού πλαισίου (context aware recommender systems – CARS) εντοπίζουν το υπόβαθρο του χρήστη, μέσα από το χρόνο, την τοποθεσία του και το σκοπό του. Αντλούν πληροφορίες από πολλές πηγές μέσα στο κοινωνικό δίκτυο έτσι ώστε να προσεγγίσουν καλύτερα τον χρήστη και να λύσουν προβλήματα όπως το ξεκίνημα των προτάσεων αλλά και την επεκτασιμότητά τους. Λαμβάνουν επίσης υπόψη την πιθανότητα αλλαγής του υπόβαθρου του χρήστη με την πάροδο του χρόνου.

Έτσι ερχόμαστε στα συστήματα χρονικής προσέγγισης (time aware recommender systems – TARS), τα οποία επικεντρώνονται στην πρόβλεψη της χρονικής εξέλιξης των επιθυμιών του χρήστη, στην παρουσία γενικότερων τάσεων και την φήμη των προϊόντων η οποία συνεχώς μεταβάλλεται. Σε ένα κοινωνικό δίκτυο με πλήθος προϊόντων και χρηστών, πολλά χαρακτηριστικά αλλάζουν ταυτόχρονα και μπορεί το ένα να

επηρεάζει το άλλο. Το γεγονός αυτό αντιμετωπίζεται με την εφαρμογή κυλιόμενων χρονικών παραθύρων, αλλά και συναρτήσεων φθοράς.

Τα συστήματα που βασίζονται στην τοποθεσία (location aware recommender systems – LARS) βρίσκουν πρακτική εφαρμογή στην βιομηχανία των ταξιδιών και του του τουρισμού, χρησιμοποιώντας για τις προτάσεις τους τοπικές αξιολογήσεις. Τέτοια παραδείγματα αποτελούν το Foursquare και το TripAdvisor. Τα συστήματα αυτά αποτελούνται συνήθως από δύο μέρη, ένα σύστημα μοντελοποίησης όπου μαθαίνει τα ενδιαφέροντα κάθε χρήστη και τις τοπικές προτιμήσεις κάθε περιοχής και ένα δεύτερο σύστημα που συνδυάζει την παραπάνω γνώσεις και παράγει προτάσεις.

Τέλος τα συστήματα προσαρμοσμένα στην κοινότητα εκμεταλλεύονται κοινωνικές σχέσεις προκειμένου να παρέχουν πιο αξιόπιστες προτάσεις. Βασισμένα στην άποψη ότι οι προτιμήσεις των χρηστών επηρεάζονται περισσότερο από αυτές των φίλων τους, παρά από αυτές αγνώστων, προσαρμόζουν στους αλγόριθμους των προτάσεων την έννοια της εμπιστοσύνης και της μεταβατικότητας των προτιμήσεων. Με τον τρόπο αυτό, επιλύεται το πρόβλημα της απαρχής των προτάσεων, βρίσκοντας ομοιότητες μεταξύ των χρηστών. Σαν γενικότερο κανόνα, τα συστήματα αυτά διαμερίζουν τους χρήστες σε κοινωνικές ομάδες με την βοήθεια των διαθέσιμων δεδομένων και στη συνέχεια χωρίζουν το αρχικό πρόβλημα της πρότασης σε πολλά μικρότερα.

Ως εξέλιξη όλων των παραπάνω, το πρόβλημα μπορεί να μεταφερθεί στην πρόταση πακέτων με προϊόντα, αντί για μεμονωμένα αντικείμενα, όπως η πρόταση αγοράς ενός συνόλου μαθημάτων, ή ενός ολοκληρωμένου γεύματος αποτελούμενο από πολλά πιάτα. Μια ακόμα εξέλιξη των συστημάτων αυτών, αποτελεί η πρόταση φίλων και η πρόβλεψη συνδέσμων, η οποία έχει αναλυθεί εκτενώς στην παρούσα εργασία. Τέλος τα τελευταία χρόνια έχει κάνει την εμφάνιση του ένας νέος τρόπος εξαγωγής προτάσεων, βασισμένος σε μεμονωμένες συνεδρίες (sessions) κάθε χρήστη. Εφαρμόζονται επαναλαμβανόμενα νευρωνικά δίκτυα για την μοντελοποίηση των δεδομένων από τα ‘κλικ’ σε μια συνεδρία και την εκμάθηση της συμπεριφοράς του χρήστη.

Κλείνοντας θα δώσουμε έμφαση στα ανοιχτά προβλήματα των συστημάτων αυτών, σύμφωνα με την έρευνα των συγγραφέων. Αυτά είναι η ποικιλία των δεδομένων και η αδυναμία αποτελεσματικής επεξεργασίας τους, η μεταβλητότητα των δεδομένων, εξαιτίας της συνεχούς μεταβολής των προτιμήσεων των χρηστών και ο τεράστιος όγκος των διαθέσιμων δεδομένων. Για την ικανοποίηση των αναγκών αυτών, το πεδίο της έρευνας στρέφεται στη δημιουργία νέων αλγορίθμων και μαθηματικών εργαλείων, αλλά και στην προσπάθεια παραλληλοποίησης μέρους αυτών.

Με την παραπάνω ανάλυση θέλαμε να δείξουμε την πορεία της έρευνας από την επίλυση μικρών μεμονωμένων προβλημάτων, στη δημιουργία τεράστιων πολύπλοκων συστημάτων, χωρίς όμως να αναλωθούμε σε εξισώσεις και μαθηματικά. Προσπαθήσαμε επίσης να αναδείξουμε περισσότερο το πεδίο εφαρμογής της παρούσας εργασίας σε πραγματικά καθημερινά συστήματα και τη χρησιμότητα που μπορεί να έχει η μελέτη και εκμάθηση τέτοιων διαδικασιών και εργαλείων στον έξω κόσμο και την αλληλεπίδραση των ανθρώπων.

Επίλογος

Στην εργασία αυτή ακολουθήσαμε έναν παραγωγικό τρόπο εξήγησης και ανάλυσης. Αφού αρχικά ορίσαμε το πρόβλημα της πρόβλεψης της συμπεριφοράς ενός γράφου και το συνδέσαμε με την πρόβλεψη συνδέσμου, προσπαθήσαμε να δούμε επιγραμματικά τρόπους επίτευξης του στόχου. Επικεντρωθήκαμε σε ένα από τα μοντέλα περιγραφής της συμπεριφοράς του δικτύου και το αναλύσαμε σε βάθος ερμηνευτικά, μαθηματικά και αλγοριθμικά. Εφαρμόσαμε στην πράξη το μοντέλο σε δύο σύνολα δεδομένων και προσπαθήσαμε να εξάγουμε συμπεράσματα. Αν και το τελικό αποτέλεσμα ήταν διφορούμενο, η όλη διαδικασία συνέβαλε σημαντικά στην κατανόηση της διαδικασίας ανάλυσης ενός γράφου και στη σκιαγράφηση περιγραφικών μοντέλων συμπεριφοράς για κοινωνικά δίκτυα.

Βιβλιογραφία

1. *Ryan A. Rossi, Jennifer Neville, Brian Gallagher. Modeling Dynamic Behavior in Large Evolving Graphs.*
2. *Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Sugato Basu, Leman Akoglu, Danai Koutra, Christos Faloutsos, Lei Li. Rolx: Stractural Role Extraction & Mining in Large Graphs.*
3. *Keith Henderson, Brian Gallagher, Tina Eliassi-Rad, Hanghang Tong, Leman Akoglu, Christos Faloutsos, Lei Li. It's who you know: Graph Mining Using Recursive Structural Features.*
4. *Purnamrita Sarkar, Deepayan Chakrabarti, Michael I. Jordan. Nonparametric Link Prediction in Dynamic Networks.*
5. *Inderjit S. Dhillon, Suvrit Sra. Generalized Nonnegative Matrix Approximations with Bregman Divergences.*
6. *Peter Grunwald. Introducing the Minimum Description Length.*
7. *Kostantinos Semertzidis, Kostas Lillis, Evaggelia Potoura. TimeReach: Historical Reachability Queries on Evolving Graphs.*
8. *Patrik Vinay Gupte, Balaraman Ravindran. Scalable Positional Analysis for Studying Evolution of Nodes in Networks*
9. *Ryan A. Rossi, Luke K. McDowell, David W. Aha, Jennifer Neville. Transforming Graph Data for Statistical Relational Learning*
10. *Ryan A. Rossi, Nesreen K. Ahmed. Role Discovery in Networks*
11. *David Liben-Nowell, Jon Kleinberg. The Link prediction Problem for Social Networks*
12. *Yizhou Sun, Jiawei Han, Charu C. Aggrawal, Nitesh V. Chawla. When Will It Happen? – Relationship Predictions in Heterogeneous Information Networks*

13. *Magdalini Eirinaki, Malamati Louta, Iraklis Varlamis. A Trust-Aware System for Personalized User Recommendations in Social Networks*
14. *Tomasz Tylenda, Ralitsa Angelova, Srikanta Bedathur. Towards Time-aware Link prediction in Evolving Social Graphs*
15. *Magdalini Eirinaki, Jerry Gao, Iraklis Varlamis, Konstantinos Tserpes. Recommender Systems for Large-Scale Social Networks: A review of challenges and solutions*

Datasets

[1] http://networkrepository.com/socfb-nips-ego.php
[2] https://snap.stanford.edu/data/egonets-Facebook.html

Source Code Based on

[1] https://github.com/randomsurfer/refex