ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΜΣ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ

ΚΑΤΕΥΘΥΝΣΗ: ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Εργασία Μαθήματος:

Εξόρυξη και Ανάλυση Δεδομένων

Επίδραση των χαρακτηριστικών του δικτύου στη διάχυση της πληροφορίας

Δημήτριος Τσέλιος — **ΜΕ 2059**

Βασίλειος Ζώης — **ΜΕ 2044**

Επιβλέπουσα Καθηγήτρια:

Μαρία Χαλκίδη, Αναπληρώτρια Καθηγήτρια

ΠΕΙΡΑΙΑΣ

Ιούνιος 2021

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία έχει σαν στόχο τη μελέτη της επίδρασης των χαρακτηριστικών του δικτύου στη διάχυση της πληροφορίας. Στο πρώτο μέρος της εργασίας, χρησιμοποιήσαμε από τον ιστότοπο https://snap.stanford.edu/data/#socnets του Stanford Large Network Dataset Collection, το σύνολο δεδομένων ego-Facebook από το οποίο εξήγαμε ένα μη κατευθυνόμενο γράφο με κόμβους και ακμές. Υπολογίζουμε μετρικές οι οποίες αποτελούν μέτρα αξιολόγησης των γράφων. Επιπλέον, επιλέγονται οι κόμβοι με τις μεγαλύτερες τιμές για τις συγκεκριμένες μετρικές και τις σχολιάζουμε. Το δεύτερο μέρος της εργασίας, αφορά τη δημιουργία τυχαίων γράφων με διαφορετικά χαρακτηριστικά και τη μελέτη της διάχυσης της πληροφορίας, σύμφωνα με το Independent Cascade Model. Πιο συγκεκριμένα, μελετάμε το πώς μεταβάλλεται η διάχυση της πληροφορίας με τυχαία επιλογή αρχικών κόμβων σε σχέση με το degree, betweeness, clustering coefficient, closeness, eigenvector centrality των γράφων. Τέλος, μελετάμε το αν και πως επηρεάζει η αρχική επιλογή κόμβων τη διάχυση της πληροφορίας, επιλέγοντας αρχικούς κόμβους με βάση τις μεγαλύτερες τιμές για τις παραπάνω μετρικές.

Περιεχόμενα

1. ΕΙΣΑΓΩΓΗ

Τα κοινωνικά δίκτυα είναι οντότητες (άτομα, οργανισμοί) που συνδέονται μεταξύ τους μέσω μιας ή περισσότερων αλληλεξαρτήσεων. Αυτές οι σχέσεις αλληλεξάρτησης αντιπροσωπεύουν αξίες, φυσική επαφή, οικονομική επικοινωνία και συμμετοχή σε ομάδες. Τα κοινωνικά δίκτυα καθορίζουν τη συμπεριφορά των οντοτήτων και κατανοούν τις σχέσεις μεταξύ τους. Αυτό επιτυγχάνεται μέσω ενός συνόλου ανεπτυγμένων μεθόδων που μπορούν να αναλύσουν τη δομή του δικτύου και να κατανοήσουν δημοφιλή πρότυπα. Οι λόγοι για τους οποίους ξεχωρίζουν τα κοινωνικά δίκτυα είναι οι συνθήκες υπό τις οποίες οι χρήστες ενδέχεται να συνδέονται ή όχι, και το γεγονός ότι οι κόμβοι μπορούν να εμφανίζουν παρόμοια χαρακτηριστικά. Η ανάλυση των κοινωνικών δικτύων στοχεύει στη μελέτη των χαρακτηριστικών του δικτύου, όπως η συμπεριφορά των χρηστών και η σχέση μεταξύ των χρηστών, προκειμένου να εξαχθούν τα υπάρχοντα δομικά στοιχεία του δικτύου, να περιγραφεί και να μελετηθεί η ροή πληροφοριών και να ανακαλυφθεί ο αντίκτυπός της σε άτομα και διάφορα δίκτυα. Όσον αφορά το πώς τα κοινωνικά δίκτυα σχηματίζουν συνδέσμους, υπάρχουν ορισμένες παραδοχές που ισχύουν για τα κοινωνικά δίκτυα. Όταν οι συνθήκες είναι όμοιες, εάν δύο κόμβοι είναι τοπολογικά (γεωγραφικά) ο ένας κοντά στον άλλο, θεωρείται πιθανότερο να ενωθούν δύο κόμβοι.

2. Θεωρία Γράφων και Μετρικές Αξιολόγησης

Σύμφωνα με τη θεωρία γραφημάτων, ένας γράφος ή γράφημα G = (V, E) αποτελείται από ένα σύνολο σημείων ή κορυφών ή κόμβων που υποδηλώνονται με V και με E οι συνδέσεις μεταξύ ενός συνόλου στοιχείων V. Επομένως, το γράφημα είναι ένα διατεταγμένο ζεύγος G = (V, E) που αποτελείται από το σύνολο V κορυφών ή κόμβων και το σύνολο E άκρων ή γραμμών, δηλαδή, μια ακμή σχετίζεται με δύο κορυφές. Οι κορυφές που ανήκουν σε μια ακμή ονομάζονται άκρα της ακμής. Η κορυφή που ανήκει στην ακμή ονομάζεται άκρη της ακμής. Μια κορυφή μπορεί να υπάρχει στο γράφημα και να μην ανήκει στην ακμή.

Στα πλαίσια της εργασίας, χρησιμοποιήθηκε από τον ιστότοπο https://snap.stanford.edu/data/#socnets του Stanford Large Network Dataset Collection, το σύνολο δεδομένων ego-Facebook το οποίο περιέχει μη κατευθυνόμενους γράφους, με 4.039 κόμβους (Nodes) και 88.234 ακμές (Edges). Αναλύουμε τις βασικότερες μετρικές που χρησιμοποιούνται στην ανάλυση των κοινωνικών δικτύων με στόχο τη μελέτη και την αξιολόγηση τους. Με τη βοήθεια τους γίνεται ευκολότερη η κατανόηση των κοινωνικών χαρακτηριστικών και ιδιοτήτων των δικτύων και η μελέτη της δομής και της εξέλιξής τους με το πέρασμα του χρόνου. Ακόμη, είναι χρήσιμες ως μέτρα κοινωνικής επιρροής και μας βοηθούν να βγάλουμε χρήσιμα συμπεράσματα που αφορούν τη διάχυση πληροφοριών στα κοινωνικά δίκτυα.

Degree: Η συγκεκριμένη μετρική αφορά τη σημαντικότητα ενός κόμβου στο δίκτυο, καθώς όσο περισσότερους γειτονικούς κόμβους μπορεί να έχει, τόσο πιο ενεργός θα είναι στο δίκτυο.

Betweeness: Ανιχνεύει την επιρροή που έχει ένας κόμβος στη ροή των πληροφοριών σε ένα γράφημα. Χρησιμοποιείται συχνά για την εύρεση κόμβων που λειτουργούν ως γέφυρα από το ένα μέρος του γραφήματος στο άλλο.

Closeness: Το αμοιβαίο άθροισμα του μήκους των μικρότερων διαδρομών μεταξύ του κόμβου και όλων των άλλων κόμβων στο γράφημα. Έτσι, όσο πιο κεντρικός είναι ένας κόμβος, τόσο πιο κοντά είναι σε όλους τους άλλους κόμβους.

Clustering Coefficient: Ένας συντελεστής ομαδοποίησης που δείχνει το μέτρο του βαθμού στον οποίο οι κόμβοι σε ένα γράφημα τείνουν να συσσωρεύονται μαζί. Ιδιαίτερα στα κοινωνικά δίκτυα, οι κόμβοι τείνουν να δημιουργούν στενά δεμένες ομάδες, που χαρακτηρίζονται από σχετικά υψηλή πυκνότητα δεσμών.

Eigenvector Centrality: Είναι ένα μέτρο της επιρροής ενός κόμβου σε ένα δίκτυο. Οι σχετικές βαθμολογίες εκχωρούνται σε όλους τους κόμβους του δικτύου, με βάση την ιδέα ότι οι συνδέσεις με κόμβους υψηλής βαθμολογίας συμβάλλουν περισσότερο στην βαθμολογία του εν λόγω κόμβου από τις ίσες συνδέσεις με τους κόμβους χαμηλής βαθμολογίας. Η υψηλή βαθμολογία του eigenvector σημαίνει ότι ένας κόμβος συνδέεται με πολλούς κόμβους που οι ίδιοι έχουν υψηλές βαθμολογίες.

Με τη βοήθεια της Python και της βιβλιοθήκης NetworkX, πραγματοποιήθηκε μελέτη και ανάλυση των παραπάνω μετρικών στο γράφο. Υπολογίσθηκε η μέση τιμή για τις μετρικές closeness, clustering coefficient, betweenness, degree, eigenvector centrality και τα αποτελέσματά τους είναι τα παρακάτω.

Average Degree	43.691
Average Betweeness Centrality	0.0007
Average Eigenvector Centrality	0.0039
Average Clustering Coefficient	0.6055
Average Closeness Centrality	0.2762

Πίκανας 1: Μέση τιμή μετρικών αξιολόγησης του γράφου.

Από τον παραπάνω πίνακα μπορούμε να συμπεράνουμε για το γράφο που μελετάμε ότι κατά μέσο όρο ο κάθε κόμβος του γράφου, συνδέεται με περίπου 43 άλλους κόμβους. Επίσης, λόγω του χαμηλού μέσου betweeness centrality, προκύπτει ότι ο γράφος μας έχει συνοχή. Αυτό σημαίνει ότι δεν υπάρχουν πολλοί κόμβοι οι οποίοι να λειτουργούν ως γέφυρες και εν τέλει να έχουν μεγάλη επιρροή στην ροή των πληροφοριών. Επιπλέον, η μέση τιμή για το closeness είναι 0.2762, που σημαίνει ότι η μέση απόσταση μεταξύ των κόμβων είναι σχετικά μικρή, κάτι που σημαίνει ότι οι περισσότεροι κόμβοι του γράφου είναι κεντρικοί. Επί προσθέτως, παρατηρείται μεγάλη διαφορά μεταξύ του μέσου eigenvector centrality των κόμβων σε σχέση με τους δέκα μεγαλύτερους σε βαθμολογία. Αυτό σημαίνει ότι οι περισσότεροι κόμβοι δεν έχουν μεγάλη επιρροή στη διάχυση της πληροφορίας, όμως υπάρχουν μικρές σε αριθμό ομάδες κόμβων οι οποίες έχουν πολύ μεγάλη επιρροή. Τέλος, το μέσο clustering coefficient των κόμβων είναι σχετικά υψηλό, κάτι που σημαίνει ότι οι κόμβοι τείνουν να ομαδοποιούνται, δηλαδή να συσσωρεύονται μαζί και ως εκ τούτου να έχουμε σχετικά υψηλή πυκνότητα στο γράφο.

Στη συνέχεια, επιλέχθηκαν οι 10 κόμβοι με τις μεγαλύτερες τιμές για κάθε μια από τις παραπάνω μετρικές.

Top 10 nodes by degree:

('107'	1045)
('1684'	792)
('1912'	755)
('3437'	547)
('0'	347)
('2543'	294)
('2347'	291)
('1888'	254)
('1800'	245)
('1663'	235)

Πίκανας 2: Οι 10 κόμβοι με το μεγαλύτερο degree.

Top 10 nodes by betweeness:

('107'	0.4805180785560152)
('1684'	0.3377974497301992)
('3437'	0.23611535735892905)
('1912'	0.2292953395868782)
('1085'	0.14901509211665306)
('0'	0.14630592147442917)
('698'	0.11533045020560802)
('567'	0.09631033121856215)
('58'	0.08436020590796486)
('428'	0.06430906239323866)

Πίκανας 3: Οι 10 κόμβοι με το μεγαλύτερο betweeness.

Top 10 nodes by closeness:

('107'	0.45969945355191255)
('58'	0.3974018305284913)
('428'	0.3948371956585509)
('563'	0.3939127889961955)
('1684'	0.39360561458231796)
('171'	0.37049270575282134)

('348'	0.36991572004397216)
('483'	0.3698479575013739)
('414'	0.3695433330282786)
('376'	0.36655773420479304)

Πίκανας 4: Οι 10 κόμβοι με το μεγαλύτερο closeness.

Top 10 nodes by eigenvector centrality:

('1912'	0.09540696149067629)
('2266'	0.08698327767886553)
('2206'	0.08605239270584343)
('2233'	0.08517340912756598)
('2464'	0.08427877475676092)
('2142'	0.08419311897991796)
('2218'	0.08415573568055032)
('2078'	0.08413617041724979)
('2123'	0.08367141238206226)
('1993'	0.0835324284081597)

Πίκανας 5: Οι 10 κόμβοι με το μεγαλύτερο eigenvector centrality.

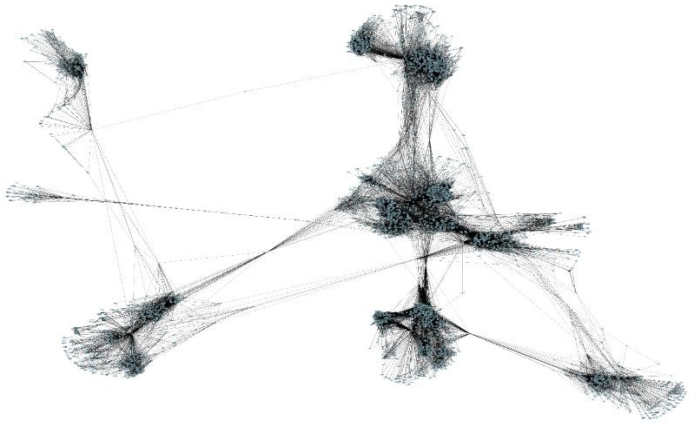
Top 10 nodes by clustering coefficient:

('32'	1.0)
('33'	1.0)
('35'	1.0)
('42'	1.0)
('44'	1.0)
('46'	1.0)
('47'	1.0)
('52'	1.0)
('63'	1.0)
('70'	1.0)

Πίκανας 6: Οι 10 κόμβοι με το μεγαλύτερο clustering coefficient.

Παρατηρείται ότι οι κόμβοι με τις μεγαλύτερες τιμές για τις μετρικές degree, betweeness, closeness είναι σχεδόν οι ίδιοι. Ειδικότερα, οι top 5 κόμβοι σε degree, είναι οι ίδιοι με το top 5 των κόμβων σε betweeness, ενώ ταυτόχρονα οι top 5 κόμβοι σε closeness είναι οι ίδιοι με τους top 10 του betweeness. Αυτό σημαίνει ότι πρόκειται για κεντρικούς κόμβους, με μεγάλη συνδεσιμότητα, μεγάλη επιρροή και έχουν μικρές αποστάσεις από τους άλλους κόμβους. Ακόμα, παρατηρούμε ότι οι κόμβοι με τις μεγαλύτερες τιμές για το clustering coefficient, βρίσκονται πολύ κοντά μεταξύ τους, κάτι που είναι λογικό καθώς στα κοινωνικά δίκτυα, οι κόμβοι τείνουν να δημιουργούν στενά δεμένες ομάδες, που χαρακτηρίζονται από σχετικά υψηλή πυκνότητα δεσμών. Συνεπώς η πιο ισχυρή ομαδοποίηση του γραφήματός μας, φαίνεται να είναι αυτή με τους παραπάνω κόμβους. Παρόμοια, παρατηρούμε ότι οι κόμβοι με τις μεγαλύτερες τιμές για το Eigenvector centrality, βρίσκονται κοντά μεταξύ τους, καθώς όπως επιβεβαιώνει και ο ορισμός, η υψηλή βαθμολογία του eigenvector σημαίνει ότι ένας κόμβος συνδέεται με πολλούς κόμβους που και οι ίδιοι έχουν υψηλές βαθμολογίες.

Οπτικοποίηση του γράφου από το σύνολο δεδομένων ego-Facebook:



Πίκανας 7: Ο Γράφος του κοινωνικού δικτύου από το σύνολο δεδομένων ego-Facebook.

3. Τυχαίοι Γράφοι και Διάχυση της Πληροφορίας

3.1 Independent Cascade Model

Το Independent Cascade μοντέλο ξεκινά με ένα αρχικό σύνολο ενεργών κόμβων A: η διαδικασία διάχυσης ξεδιπλώνεται σε διακριτά βήματα σύμφωνα με τον ακόλουθο τυχαιοποιημένο κανόνα: Όταν ο κόμβος V ενεργοποιηθεί στο βήμα t, δίνεται μια μοναδική ευκαιρία να ενεργοποιήσει κάθε τρέχοντα ανενεργό γείτονα W, που το πετυχαίνει με πιθανότητα P(V,W). Εάν ο W έχει πολλούς νέους ενεργοποιημένους γείτονες, οι προσπάθειές τους αλληλουχίζονται με αυθαίρετη σειρά. Εάν το V επιτύχει, τότε το W θα ενεργοποιηθεί στο βήμα t + 1. Είτε ο V πετύχει την ενεργοποίηση του W είτε όχι, δεν μπορούν να γίνουν περαιτέρω προσπάθειες ενεργοποίησης του W σε επόμενους γύρους. Η διαδικασία εκτελείται έως ότου δεν είναι δυνατές άλλες ενεργοποιήσεις.

3.2 Διάχυση της πληροφορίας με τυχαία επιλογή αρχικών κόμβων

Δημιουργήσαμε 4 τυχαίους γράφους με διαφορετικά χαρακτηριστικά (average degree, clustering coefficient, closeness, betweeness, eigenvector centrality). Πιο συγκεκριμένα, δημιουργήσαμε τους παρακάτω γράφους, σύμφωνα με το Independent Cascade Model όπου επιστρέφει ένα τυχαίο γράφημα Gn, p, επίσης γνωστό ως γράφημα Erdős-Rényi ή διωνυμικό γράφημα.
Το μοντέλο Gn, p επιλέγει καθένα από τα πιθανά άκρα με πιθανότητα p.
Παράμετροι:

- **n** (int) - Ο αριθμός των κόμβων.
- **p** (float) - Πιθανότητα δημιουργίας άκρων.
- **seed** -Τυχαία παραγωγή ενεργών κόμβων.

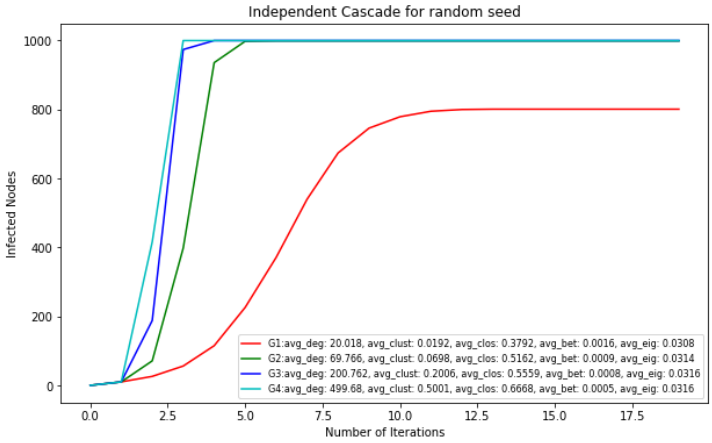
G1 = nx.erdos_renyi_graph(1000, 0.02)

G2 = nx.erdos_renyi_graph(1000, 0.05)

G3 = nx.erdos_renyi_graph(1000, 0.1)

G4 = nx.erdos_renyi_graph(1000, 0.5)

Υπολογίστηκαν οι τιμές των μετρικών για κάθε έναν από τους τυχαίους γράφους, με τα αποτελέσματα να φαίνονται στο παρακάτω γράφημα.

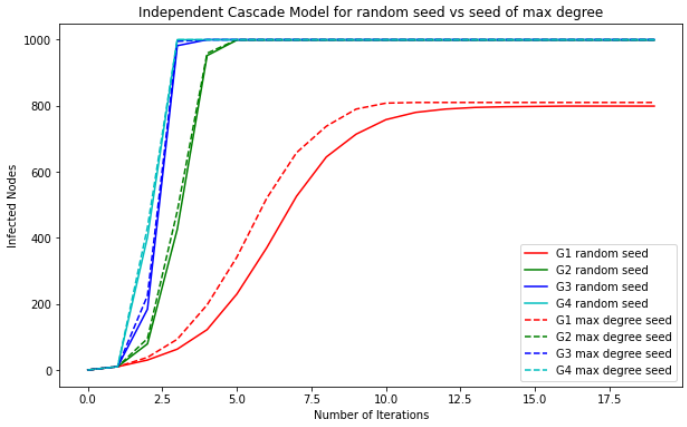


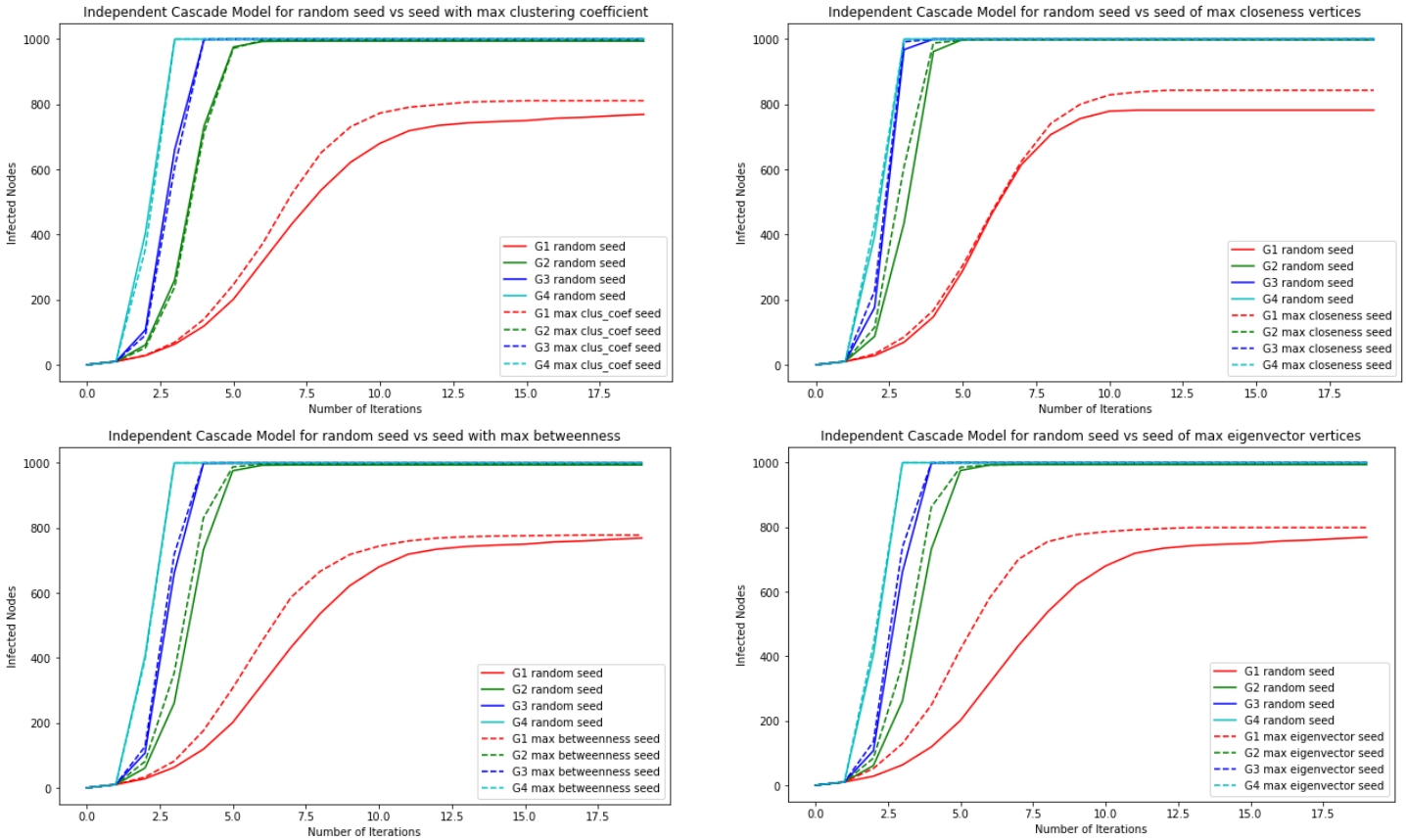
Πίκανας 8: Τυχαίοι γράφοι με διαφορετικές τιμές για κάθε μια από τις μετρικές και τυχαίο αριθμό ενεργών κόμβων.

Όπως παρατηρείται και στο παραπάνω σχήμα, η διάχυση της πληροφορίας εξαρτάται από τον αριθμό των επαναλήψεων, καθώς επίσης και από τις τιμές που έχουν τα διαφορετικά χαρακτηριστικά τους (average degree, clustering coefficient, closeness, betweeness, eigenvector centrality).
Η διάχυση της πληροφορίας είναι πιο άμεση όταν έχουμε δίκτυα με μεγάλη συνδεσιμότητα, μεγάλη επιρροή και οι κόμβοι έχουν μικρές αποστάσεις από τους άλλους κόμβους. Συνεπώς, όσο μεγαλύτερες είναι οι τιμές των χαρακτηριστικών, τόσο μεγαλύτερη και σύντομη θα είναι η διαδικασία της διάχυσης της πληροφορίας.

3.3 Διάχυση της πληροφορίας με επιλογή συγκεκριμένων αρχικών κόμβων

Στη συνέχεια, παρουσιάζουμε πως επηρεάζει η αρχική επιλογή κόμβων στη διαδικασία διάχυσης. Αυτή τη φορά δεν επιλέξαμε τυχαίο αριθμό αρχικών κόμβων, αλλά επιλέξαμε αρχικούς κόμβους με βάση το μεγαλύτερο degree, clustering coefficient, betweenness, closeness, eigenvector centrality, όπως φαίνεται παρακάτω.





Πίκανας 9: Τυχαίοι γράφοι με διαφορετικές τιμές για κάθε μια από τις μετρικές και αριθμός ενεργών κόμβων σύμφωνα με τις μέγιστες τιμές για κάθε μια από τις μετρικές και σύγκριση με τους γράφους με τυχαίους ενεργούς κόμβους.

Το συμπέρασμα που προκύπτει από τα παραπάνω διαγράμματα είναι ότι για τυχαίους γράφους με μικρή πιθανότητα σύνδεσης, η επιλογή των αρχικών κόμβων με βάση τις μεγαλύτερες τιμές για τα χαρακτηριστικά degree, clustering coefficient, betweenness, eigenvector centrality, τις περισσότερες φορές φαίνεται να επηρεάζει θετικά τη διάχυση της πληροφορίας σε σχέση με την τυχαία επιλογή αρχικών κόμβων. Για να διαπιστώσουμε την παραπάνω παρατήρηση, επαναλάβαμε τη διαδικασία δημιουργίας τυχαίων γράφων. Τις περισσότερες φορές να επηρεάζει θετικά τη διάχυση της πληροφορίας επιβεβαιώνοντας τα παραπάνω, παρόλα αυτά, υπήρξαν λίγες φορές που λόγω της μικρής συνδεσιμότητας και της τυχαίας συνθήκης κάτω από την οποία δημιουργούνται οι τυχαίοι γράφοι και λόγω της τυχαίας πορείας που ακολουθεί η διαδικασία της διάχυσης της πληροφορίας, η τυχαία επιλογή κόμβων φάνηκε πιο αποτελεσματική. Ωστόσο, για τυχαίους γράφους με μεγαλύτερη συνδεσιμότητα, η επιλογή συγκεκριμένων αρχικών κόμβων δεν φαίνεται να επηρεάζει σημαντικά το χρόνο και την απόδοση της διάχυσης της πληροφορίας σε σχέση με την τυχαία επιλογή αρχικών κόμβων.

Βιβλιογραφία

Chi Wang, W. C. (2012, Νοέμβριος 1). Scalable influence maximization for independent cascade model in large-scale social networks. *https://doi.org/10.1007/s10618-012-0262-1*, σσ. 545-576.

Ghosh, S. (2009). *Network Theory: Analysis and Synthesis.* PHI Learning.

Jure Leskovec, A. R. (2014). *Mining of Massive Datasets (2nd Edition).* Cam-bridge University Press.

Newman, M. (2010). *Networks: An Introduction.* Oxford Scholarship Online.

Saito K., N. R. (2008). Prediction of Information Diffusion Probabilities for Independent Cascade Model. In: Lovrek I., Howlett R.J., Jain L.C. (eds) Knowledge-Based Intelligent Information and Engineering Systems. Στο N. R. Saito K., *Lecture Notes in Computer Science, vol 5179.* Springer, Berlin, Heidelberg.

Srinivasa, K. R. (2018). *Practical Social Network Analysis with Python.* Springer, Cham.