8. Συστήματα Συστάσεων βασισμένα σε Γράφους

***Σύνοψη**: Στο όγδοο κεφάλαιο θα μελετήσουμε τα **βασισμένα σε γράφους συστήματα συστάσεων**. Επίσης, θα παρουσιάσουμε τα **μέτρα ομοιότητας βάσει τοπικών χαρακτηριστικών του γράφου** (π.χ. γειτονικοί κόμβοι του υπό εξέταση κόμβου, αριθμός των κοινών κόμβων κ.λπ.), καθώς και **βάσει των χαρακτηριστικών του συνολικού γράφου** (όπως το μήκος του μονοπατιού που συνδέει δύο κόμβους, ο αριθμός των διαφορετικών μονοπατιών που συνδέουν δύο κόμβους κ.λπ.). Ακόμη, εδραζόμενοι στα μέτρα ομοιότητας τοπικού χαρακτήρα θα αναλύσουμε αλγορίθμους όπως ο **Common Friends** και ο **Preferential Attachment**, ενώ θα πράξουμε το ίδιο και για αλγορίθμους όπως ο **Random Walk with Restart** (**Personalized PageRank**), ο **SimRank**, και ο **PathSim** επί μέτρων ομοιότητας που λαμβάνουν υπόψη τη συνολική όμως δομή του γράφου. Τέλος, θα εμβαθύνουμε στα **συστήματα συστάσεων που βασίζονται σε γράφους γνώσης**, όπου δηλαδή τα δεδομένα αποθηκεύονται με τη βοήθεια μια "οντολογίας" που επιτρέπει την εξαγωγή συμπερασμάτων ακολουθώντας απλούς κανόνες της **περιγραφικής λογικής** (**description logic**).*

***Προαπαιτούμενη γνώση**: Ο αναγνώστης καλό θα ήταν να έχει προηγουμένως μελετήσει το πρώτο και το δεύτερο κεφάλαιο.*

8.1 Εισαγωγή στους Ετερογενείς Γράφους: Βασικές γνώσεις

Σ' αυτήν την ενότητα θα παρουσιάσουμε την έννοια του *ετερογενούς δικτύου πληροφοριών* (*heterogeneous information network*) ορίζοντας ως τρέχον παράδειγμα ένα *σχήμα δικτύου* (*network schema*) μιας διαδικτυακής εφημερίδας, το οποίο και θα χρησιμοποιηθεί για την παρουσίαση του *αλγορίθμου αναζήτησης ομοιότητας με βάση έναν γράφο δεδομένων*, γνωστό ως *Personalized PageRank* (ή αλλιώς *Random Walk with Restart*).

8.1.1 Ετερογενές Δίκτυο Πληροφοριών

Ορισμός 8.1 ***Δίκτυο πληροφοριών:*** *[Yizhou et al., 2011] Ένα δίκτυο πληροφοριών ορίζεται ως ένας κατευθυνόμενος γράφος $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ με μια συνάρτηση αντιστοίχισης κόμβων (nodes) $\phi : \mathcal{V} \rightarrow \mathcal{Q}$ και μια συνάρτηση αντιστοίχισης ακμών (edges) $\psi : \mathcal{E} \rightarrow \mathcal{R}$ όπου κάθε κόμβος $v \in \mathcal{V}$ ανήκει σ' έναν συγκεκριμένο τύπο κόμβου $\phi(v) \in \mathcal{Q}$, και κάθε ακμή $e \in \mathcal{E}$ σ' έναν συγκεκριμένο τύπο ακμής $\psi(e) \in \mathcal{R}$.*

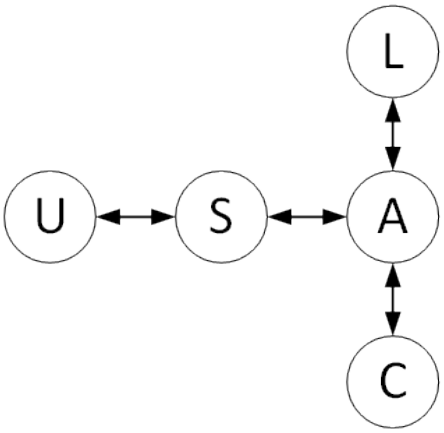
Σε αντίθεση με τον παραδοσιακό ορισμό του *δικτύου πληροφοριών*, στον Ορισμό 8.1 διακρίνουμε ξεκάθαρα τους διαφορετικούς τύπους κόμβου και τους διαφορετικούς τύπους ακμής του δικτύου. Όταν οι τύποι των κόμβων $|\mathcal{Q}| > 1$ ή οι τύποι των ακμών $|\mathcal{R}| > 1$ είναι περισσότεροι του ενός, τότε το δίκτυο ονομάζεται *Ετερογενές Δίκτυο Πληροφοριών* (**Heterogeneous Information Network**).

Παράδειγμα 8.1 Έστω ένας γράφος μιας διαδικτυακής εφημερίδας ως ένα *ετερογενές δίκτυο πληροφοριών* που περιέχει κόμβους από πέντε διαφορετικούς τύπους οντοτήτων $\mathcal{Q} = \{U, S, A, C, L\}$: χρήστες ($U$), συνεδρίες ($S$), άρθρα ($A$), κατηγορίες άρθρων ($C$) και γεωγραφικές τοποθεσίες που αφορούν στα άρθρα (L). Αναλυτικότερα, κάθε χρήστης $u : \phi(u) = U$ συνδέεται με μία ή περισσότερες ακμές με τις συνεδρίες $s : \phi(s) = S$. Κάθε συνεδρία s έχει έναν μοναδικό χρήστη u που συνδέεται μ' αυτήν και ένα ή περισσότερα άρθρα $a : \phi(u) = A$ που διαβάζει ο χρήστης εντός της συνεδρίας. Τέλος, κάθε άρθρο a μπορεί να εμφανίζεται σε μία ή περισσότερες συνεδρίες s, να ανήκει σε μία κατηγορία ειδήσεων $c : \phi(c) = C$ και να αντιστοιχίζεται με μία γεωγραφική τοποθεσία $l : \phi(l) = L$.

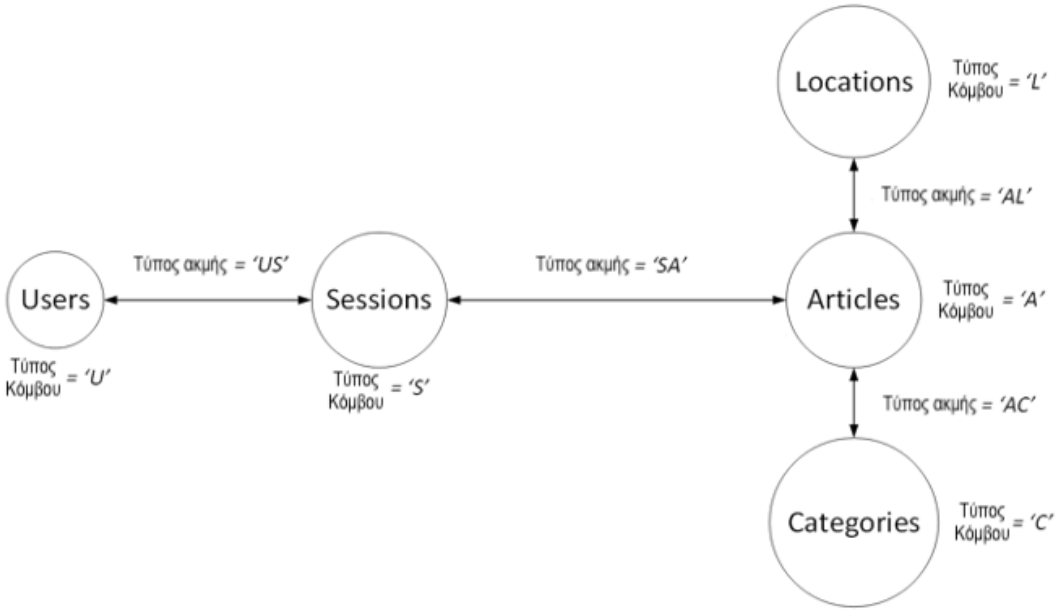
8.1.2 Σχήμα Δικτύου

Ορισμός 8.2 ***Σχήμα δικτύου:*** *[Yizhou et al., 2011] Το σχήμα δικτύου είναι ένα πρότυπο για ένα ετερογενές δίκτυο $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ με μια συνάρτηση αντιστοίχισης κόμβων $\phi : \mathcal{V} \rightarrow \mathcal{Q}$ και μια συνάρτηση αντιστοίχισης ακμών $\psi : \mathcal{E} \rightarrow \mathcal{R}$. Είναι, επίσης, ένας κατευθυνόμενος γράφος που ορίζεται πάνω σε διάφορους τύπους κόμβων $\mathcal{Q}$, με τις ακμές να απεικονίζουν τις σχέσεις (από το $\mathcal{R}$) μεταξύ των κόμβων, και συμβολίζεται ως $T_G = (\mathcal{Q}, \mathcal{R})$.*

Το *σχήμα δικτύου* (*network schema*) χρησιμεύει ως πρότυπο αναφοράς ενός δικτύου πληροφοριών και περιγράφει πόσοι τύποι κόμβων υπάρχουν στο δίκτυο πληροφοριών και ποιοι τύποι ακμών συνδέουν τους κόμβους μεταξύ τους. Το *συνοπτικό σχήμα δικτύου* για τη διαδικτυακή εφημερίδα του Παραδείγματος 8.1 εμφανίζεται στο Σχήμα 8.1. Επίσης, στο Σχήμα 8.2 παρουσιάζεται, βάσει των δεδομένων του Παραδείγματος 8.1, μια πιο λεπτομερής αναπαράσταση του *σχήματος δικτύου* με την οποία ορίζεται αναλυτικά ο τύπος των κόμβων και ο τύπος των ακμών του γράφου.



Σχήμα 8.1: *Συνοπτικό σχήμα δικτύου* της διαδικτυακής εφημερίδας του Παραδείγματος 8.1.



Σχήμα 8.2: *Λεπτομερές σχήμα δικτύου* της διαδικτυακής εφημερίδας του Παραδείγματος 8.1.

8.2 Σχετική Βιβλιογραφία

Στην παρούσα ενότητα εξετάζουμε τις σχετικές εργασίες για την πρόβλεψη νέων *συνδέσμων* (*links*) μεταξύ των *κόμβων* (*nodes*) ενός *γράφου*. Η ερευνητική περιοχή της πρόβλεψης συνδέσμων σε *μονομερή κοινωνικά δίκτυα* (δηλαδή αυτά που διαθέτουν μόνο έναν τύπο κόμβων, π.χ. δίκτυα φιλίας) προσπαθεί να κάνει μια εκτίμηση για το ποιες νέες αλληλεπιδράσεις μεταξύ των μελών ενός κοινωνικού δικτύου είναι πιθανό να συμβούν στο κοντινό μέλλον. Υπάρχουν δύο κύριες προσεγγίσεις [Liben-Nowell and Kleinberg, 2003] που χειρίζονται το πρόβλημα του *link prediction*, και οι οποίες περιγράφονται συνοπτικά παρακάτω, αλλά θα παρουσιασθούν αναλυτικά αργότερα.

Η πρώτη προσέγγιση βασίζεται σε *τοπικά χαρακτηριστικά* (*local attributes*) ενός δικτύου, εστιάζοντας κυρίως στην τοπική δομή σύνδεσης των κόμβων. Υπάρχει μια ποικιλία μέτρων *τοπικής ομοιότητας* [Liben-Nowell and Kleinberg, 2003], όπως ο *δείκτης ομοιότητας των Adamic και Adar* [Adamic and Adar, 2005], ο *αλγόριθμος Friend of a Friend (FOAF)* [Chen et al., 2009], ο *αλγόριθμος preferential attachement* [Liben-Nowell and Kleinberg, 2003] κ.ά.

Η δεύτερη προσέγγιση βασίζεται στα *συνολικά χαρακτηριστικά* (*global attributes*) ενός γράφου, διότι ανιχνεύει ολόκληρη τη δομή των μονοπατιών του γράφου. Υπάρχει και εδώ μια ποικιλία προσεγγίσεων, όπως ο *αλγόριθμος του τυχαίου περιπάτου με επανεκκίνηση* (*Random Walk with Restart* [Pan et al., 2004], ο *αλγόριθμος SimRank* [Jeh and Widom, 2002], ο *αλγόριθμος Katz* [kat], ο *αλγόριθμος PathSim* [Sun et al., 2011] κ.ά. Όλοι οι προαναφερθέντες αλγόριθμοι θα περιγραφούν αναλυτικά στις επόμενες υποενότητες.

Εκτός όμως από τους παραπάνω *αλγορίθμους πρόβλεψης συνδέσμων*, οι οποίοι εφαρμόζονται κυρίως σε *μονομερείς γράφους*, υπάρχουν και άλλες μέθοδοι που εκμεταλλεύονται επιπρόσθετες πηγές δεδομένων, όπως τα μηνύματα μεταξύ των χρηστών, τους συν-συγγραφείς μιας εργασίας και την *κοινή επισήμανση πληροφοριακών αντικειμένων* (*tag* σε φωτογραφίες, βίντεο κ.λπ.). Για παράδειγμα η ακόλουθη εργασία, [Guy et al., 2009], πρότεινε ένα νέο *widget* για την παροχή μιας λίστας με προτεινόμενους φίλους στον υπό εξέταση χρήστη. Η λίστα, λοιπόν, αυτή βασίστηκε σε συγκεντρωτικές πληροφορίες που συλλέχθηκαν από διάφορες πηγές από ολόκληρο τον οργανισμό της *IBM*. Τέλος, στην εργασία των Chen και συν-συγγραφέων [Chen et al., 2009] αξιολογήθηκαν τέσσερεις αλγόριθμοι συστάσεων (*Content Matching, Content-plus-Link, FOAF* και *SONAR*) που στοχεύουν στο να βοηθήσουν χρήστες να ανακαλύψουν νέους φίλους στο online κοινωνικό δίκτυο της *IBM*.

8.3 Local-based Similarity Measures

8.3.1 Shortest Path

Μέτρο ομοιότητας βάσει συντομότερης διαδρομής: Το *μέτρο του συντομότερου μονοπατιού* υπολογίζει τη μικρότερη απόσταση μεταξύ οποιουδήποτε ζεύγους χρηστών σ' ένα κοινωνικό δίκτυο. Επομένως, οι πιο σύντομες διαδρομές μονοπατιών μπορούν να προταθούν στον υπό εξέταση χρήστη v_x, και για να το πετύχουμε αυτό συνήθως χρησιμοποιούμε τον γνωστό *αλγόριθμο Dijkstra* ή τον αλγόριθμο της ακόλουθης μελέτης, [Fredman and Tarjan, 1987].

8.3.2 Common Neighbors

Ο δείκτης ομοιότητας κοινών γειτόνων: Ο *δείκτης ομοιότητας Common Neighbors (CN)* ή του *Φίλου ενός άλλου Φίλου [Friend of a Friend (FOAF)]* [Chen et al., 2009] βασίζεται στον αριθμό των φίλων που έχουν από κοινού οι δύο υπό εξέταση κόμβοι v_x και v_y σύμφωνα με την Εξίσωση 8.1:

$$score(v_x, v_y) := |\Gamma(v_x) \cap \Gamma(v_y)| \tag{8.1}$$

όπου $score(v_x, v_y)$ ο βαθμός συνάφειας των v_x και v_y, και $\Gamma(v_x), \Gamma(v_y)$ τα σύνολα των γειτονικών κόμβων των v_x και v_y αντίστοιχα. Άρα, το $|\Gamma(v_x) \cap \Gamma(v_y)|$ είναι ο αριθμός των από κοινού γειτονικών τους κόμβων. Οι υποψήφιοι φίλοι προτείνονται στον v_x κατά φθίνουσα σειρά βάσει της *βαθμολογίας* (*score*) τους.

8.3.3 Jaccard index

Ο δείκτης Jaccard: Μοιάζει πολύ με τον *δείκτη ομοιότητας κοινών γειτόνων*. Πιο συγκεκριμένα, καταγράφει τη συνάφεια μεταξύ δύο κόμβων ενός γράφου μετρώντας τον βαθμό επικάλυψης των γειτονικών τους κόμβων, και ορίζεται ως ακολούθως:

$$score(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{|\Gamma(v_x) \cup \Gamma(v_y)|}$$

όπου στον αριθμητή της παραπάνω εξίσωσης υπολογίζουμε τον αριθμό των κοινών γειτόνων των δύο υπό εξέταση κόμβων, ενώ στον παρονομαστή υπολογίζουμε το σύνολο των γειτόνων τους.

8.3.4 Salton index

Ο δείκτης ομοιότητας Salton [Chowdhury, 2010]: Αυτός ο δείκτης μπορεί να θεωρηθεί ως βελτιωμένη εκδοχή του *δείκτη των κοινών γειτόνων*, διότι, εκτός από τον αριθμό κοινών γειτόνων, λαμβάνει υπόψη και τον *βαθμό* (*degree*) των υπό εξέταση κόμβων. Ορίζεται ως εξής:

$$score(v_x, v_y) = \frac{|\Gamma(v_x) \cap \Gamma(v_y)|}{\sqrt{d_{v_x} * d_{v_y}}}$$

όπου d_{v_x} και d_{v_y} ο βαθμός του v_x και v_y αντίστοιχα.

8.3.5 Adamic & Adar index

Δείκτης ομοιότητας Adamic & Adar: Οι Adamic και Adar [Adamic and Adar, 2005] πρότειναν ένα *μέτρο απόστασης* για να προσδιορίζουν πότε δύο ιστοσελίδες σχετίζονται μεταξύ τους. Υπολόγισαν δηλαδή χαρακτηριστικά των ιστοσελίδων και όρισαν την ομοιότητα μεταξύ δύο σελίδων x, y ως εξής: $\sum_z \frac{1}{\log(frequency(z))}$ όπου το z είναι ένα χαρακτηριστικό που μοιράζονται οι σελίδες x, y. Το σκεπτικό αυτό βελτιώνει την απλή καταμέτρηση των κοινών χαρακτηριστικών σταθμίζοντας περισσότερο τα σπανιότερα χαρακτηριστικά. Η ομοιότητα μεταξύ των κόμβων v_x και v_y μπορεί να υπολογιστεί από την Εξίσωση 8.2:

$$score(v_x, v_y) = \sum_{z \in \Gamma(v_x) \cap \Gamma(v_y)} \frac{1}{\log|\Gamma(z)|} \tag{8.2}$$

όπου $\Gamma(v_x), \Gamma(v_y)$ οι κοινοί γείτονες των v_x και v_y αντίστοιχα.

8.3.6 Preferential Attachement

Preferential Attachment: Η λογική εδώ είναι ότι η πιθανότητα μια νέα ακμή να συνδεθεί με τον υπό εξέταση κόμβο v_x είναι ανάλογη του βαθμού (degree) του. Μ' άλλα λόγια, η παραπάνω πιθανότητα είναι ανάλογη του αριθμού των γειτονικών του κόμβων. Ο Barabasi και οι συνεργάτες του [Barabasi et al., 2002] καθώς και ο Newman [Newman, 2001] απέδειξαν εμπειρικά ότι η πιθανότητα μια νέα ακμή να συνδεθεί με τους κόμβους v_x και v_y συσχετίζεται με το γινόμενο του βαθμού (degree) των υπό εξέταση κόμβων v_x και v_y, το οποίο ορίζεται στην Εξίσωση 8.3:

$$score(v_x, v_y) := |\Gamma(v_x) \cdot \Gamma(v_y)| \tag{8.3}$$

όπου $\Gamma(v_x), \Gamma(v_y)$ τα σύνολα των γειτονικών κόμβων των v_x και v_y αντίστοιχα.

8.3.7 FriendLink

Ο δείκτης ομοιότητας FriendLink: Ο *δείκτης ομοιότητας FriendLink* [Papadimitriou et al., 2011] χρησιμοποιείται για το *πρόβλημα της πρόβλεψης μελλοντικών συνδέσμων μεταξύ κόμβων ενός γράφου* (**Link Prediction problem**). Στα online κοινωνικά δίκτυα (π.χ. *Facebook*) οι χρήστες δηλώνουν ρητά τους φίλους τους έτσι ώστε να είναι σε θέση να μοιράζονται πληροφορίες μαζί τους (π.χ. φωτογραφίες, βίντεο κ.λπ.). Το *μέτρο ομοιότητας FriendLink* υποθέτει ότι τα άτομα σ' έναν γράφο φιλίας μπορούν να χρησιμοποιήσουν όλα τα μονοπάτια που τα συνδέουν κατά ποσοστιαία αναλογία εξαρτώμενη από τα μήκη των μονοπατιών αυτών. Έτσι, δύο άτομα που συνδέονται με πολλά μονοπάτια έχουν μεγαλύτερη πιθανότητα να γνωρίζονται μεταξύ τους αντιστρόφως ανάλογα με το μήκος των μονοπατιών αυτών (δηλαδή, όσο μικρότερο το μήκος μονοπατιού τόσο μεγαλύτερη η πιθανότητα συνάφειας). Η *ομοιότητα* $score(v_x, v_y)$ μεταξύ δύο κόμβων v_x και v_y ενός γράφου ορίζεται ως ο αριθμός των μονοπατιών διαφορετικού μήκους ℓ από τον v_x στον v_y:

$$score(v_x, v_y) = \sum_{i=2}^{\ell} \frac{1}{i-1} \cdot \frac{\left|paths^{i}_{v_x, v_y}\right|}{\prod_{j=2}^{i}(n-j)} \tag{8.4}$$

όπου:

- n ο αριθμός των κόμβων ενός γράφου G,
- ℓ το μέγιστο μήκος ενός μονοπατιού που συνδέει τους κόμβους v_x και v_y (εξαιρουμένων των *κυκλικών μονοπατιών*). Με τον όρο *κυκλικά μονοπάτια* εννοούμε τα μονοπάτια που είναι *κλειστά*. Ένας κόμβος μπορεί να εμφανίζεται μία φορά σ' ένα μονοπάτι, ενώ το μονοπάτι $v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_1 \rightarrow v_5$ είναι *κυκλικό/κλειστό*, επειδή το v_1 εμφανίζεται δύο φορές,
- $\frac{1}{i-1}$ ένας συντελεστής *μείωσης* της σημασίας των μεγάλου μήκους μονοπατιών, ο οποίος σταθμίζει τα μονοπάτια ανάλογα με το μήκος ℓ που έχει το καθένα. Άρα, ένα μονοπάτι μήκους 2 έχει τιμή ίση

με 1 ($\frac{1}{2-1}$ = 1), ενώ ένα μονοπάτι μήκους 3 έχει τιμή ίση με $\frac{1}{2}$ ($\frac{1}{3-1}$ = $\frac{1}{2}$) κ.λπ. Υπό αυτήν την έννοια, χρησιμοποιούμε κατάλληλα *βάρη* για να μειώσουμε τη σημασία των μακρύτερων μονοπατιών,

- $\left|paths^{\ell}_{v_x,v_y}\right|$ ο αριθμός όλων των μονοπατιών μήκους-ℓ από v_x σε v_y,
- $\prod_{j=2}^{i}(n-j)$ ο αριθμός όλων των πιθανών μονοπατιών μήκους ℓ από v_x σε v_y, εάν κάθε κόμβος στο γράφημα G ήταν συνδεδεμένος μ' όλους τους άλλους κόμβους. Χρησιμοποιώντας το κλάσμα $\frac{\left|paths^{\ell}_{v_x,v_y}\right|}{\prod_{j=2}^{i}(n-j)}$ το μέτρο ομοιότητας είναι κανονικοποιημένο και παίρνει τιμές στο [0,1].

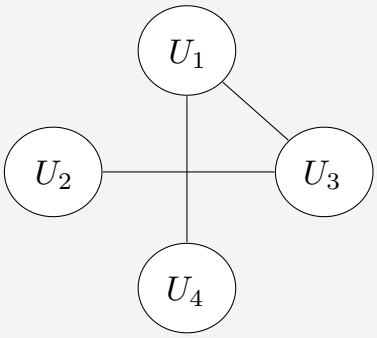
Σύμφωνα λοιπόν με το *μέτρο ομοιότητας FriendLink*, εάν δύο κόμβοι σχετίζονται πολύ μεταξύ τους, τότε αναμένουμε η τιμή $score(v_x, v_y)$ να είναι κοντά στο 1. Από την άλλη, εάν οι δύο κόμβοι δεν είναι συναφείς, αναμένουμε την τιμή $score(v_i, v_j)$ να είναι κοντά στο 0.

8.4 Global-based Similarity Algorithms

8.4.1 Katz Status Index

Για τον εντοπισμό των σημαντικότερων κόμβων σ' έναν γράφο, ο Leo Katz [Katz, 1953] συνδύασε την κοινωνιολογική θεωρία με τη γραμμική άλγεβρα προτείνοντας τον *Katz Status index*, ο οποίος αποτυπώνει το "prestige" ή αλλιώς το status ενός κόμβου μέσα στον γράφο. Εν συνεχεία, θα χρησιμοποιήσουμε ένα παράδειγμα ώστε να δείξουμε πώς μπορεί να τροποποιηθεί ο προαναφερθείς τελεστής, για να υπολογίζει τη *συνάφεια/ομοιότητα* μεταξύ των κόμβων ενός γράφου.

Παράδειγμα 8.2 Μας δίνεται ο *γράφος φιλίας* G του Σχήματος 8.3 και ζητείται να υπολογίσουμε τη *συνάφεια* των κόμβων του γράφου με τη βοήθεια του *Katz status index*. Όσον αφορά στα βασικά χαρακτηριστικά του γράφου G, αυτός αποτελείται από ένα σύνολο κόμβων V και ένα σύνολο ακμών E. Ακόμη, κάθε ακμή ορίζεται από ένα συγκεκριμένο ζευγάρι κόμβων του γράφου (v_i, v_j) όπου $v_i, v_j \in V$. Επίσης, υποθέτουμε ότι ο γράφος G είναι *μη κατευθυνόμενος* και *μη σταθμισμένος* (*non-directed* και *non-weighted*). Αυτό σημαίνει ότι οι ακμές του γράφου δεν έχουν βάρη, καθώς και ότι η σειρά των κόμβων σε μια ακμή δεν είναι σημαντική. Επομένως (v_i, v_j) και (v_j, v_i) υποδηλώνουν την ίδια ακμή στο G. Υποθέτουμε, επιπρόσθετα, ότι ο γράφος G δεν μπορεί να έχει πολλαπλές ακμές που να συνδέουν δύο κόμβους. Ως εκ τούτου, εάν δύο κόμβοι v_i, v_j συνδέονται με μια ακμή του E, τότε δεν μπορεί να υπάρχει άλλη ακμή στο E που να τους συνδέει, με δεδομένο ότι E είναι το σύνολο των ακμών του γράφου μας. Τέλος, υποθέτουμε ότι δεν μπορεί να υπάρχουν ακμές βρόχου στον γράφο G (δηλαδή ένας κόμβος δεν μπορεί να συνδέεται με τον εαυτό του μέσω μιας ακμής).



Σχήμα 8.3: Ένας γράφος φιλίας, όπου δηλαδή οι χρήστες συνδέονται μεταξύ τους με ακμές φιλίας.

Ένας γράφος $\mathcal{G}$ αναπαρίσταται συνήθως με τη βοήθεια ενός *πίνακα γειτνίασης* A (*Adjacency matrix*), στην περίπτωσή μας διαστάσεων $n \times n$ και όπου $n = |V|$ ο αριθμός των κόμβων του G. Επομένως ο *πίνακας γειτνίασης* έχει n γραμμές και n στήλες. Για έναν *μη σταθμισμένο* (*non-weighted*) και *μη πολλαπλό γράφο* (*non multi-graph*) οι τιμές του *πίνακα γειτνίασης* ορίζονται ως εξής:

$$A[v_i, v_j] = \begin{cases} 1, & (v_i, v_j) \in E \\ \\ 0, & (v_i, v_j) \notin E \end{cases}$$

Σημειώνουμε ότι ο *πίνακας γειτνίασης* ενός *μη κατευθυνόμενου και μη σταθμισμένου γράφου* G είναι ένας *συμμετρικός πίνακας* με τιμές 1 και 0 αντίστοιχα εφόσον δύο κόμβοι γειτονεύουν ή όχι. Επιπλέον, καθώς δεν υπάρχουν *ακμές βρόχου* (*self-connected edges*), η κύρια διαγώνιος του *πίνακα γειτνίασης* έχει μηδενικές τιμές. Συνεπώς, ο *πίνακας γειτνίασης του γράφου φιλίας* για τα δεδομένα του παραδείγματος 8.2 -ο οποίος εμφανίζεται στο Σχήμα 8.3- αναπαρίσταται ως *πίνακας γειτνίασης* στον Πίνακα 8.1:

	U_1	U_2	U_3	U_4
U_1	0	0	1	1
U_2	0	0	1	?
U_3	1	1	0	?
U_4	1	?	?	0

Πίνακας 8.1: *Πίνακας γειτνίασης* A για τα δεδομένα του παραδείγματος 8.2.

Καθώς θέλουμε να προβλέψουμε τις *ελλείπουσες τιμές* (βλ. το ?) του *πίνακα γειτνίασης* A, οι οποίες εμφανίζονται στον Πίνακα 8.1, μπορούμε να υποθέσουμε ότι αρχικά είναι ίσες με 0 (δηλαδή

δεν υπάρχουν ακμές μεταξύ των αντίστοιχων χρηστών). Είναι προφανές από το Σχήμα 8.3 και τον αντίστοιχο *πίνακα γειτνίασης* A που απεικονίζεται στον Πίνακα 8.1 ότι ο U_1 συνδέεται με τους U_3 και U_4, ενώ ο U_2 μόνο με τον U_3. Επίσης, οι U_1 και U_2 έχουν κοινό φίλο τον U_3, αφού είναι και οι δύο συνδεδεμένοι μ' αυτόν τον χρήστη. Ας υποθέσουμε λοιπόν για το τρέχον Παράδειγμα 8.2 ότι θέλουμε να προτείνουμε νέους φίλους στον χρήστη U_4. Αρχικά θα παρουσιάσουμε το θεωρητικό πλαίσιο και στη συνέχεια θα δώσουμε μια λύση για τα δεδομένα του Παραδείγματος 8.2.

Υπάρχουν, βέβαια, διάφορα μέτρα ομοιότητας που λαμβάνουν υπόψη τον *συνολικό γράφο* [Liben-Nowell and Kleinberg, 2003] (π.χ. *Katz Status index*, *αλγόριθμος RWR*, *αλγόριθμος SimRank* κ.λπ.) για την αποτύπωση της ομοιότητας των κόμβων σ' ένα δίκτυο, οι οποίοι εξαρτώνται από τον αριθμό των μονοπατιών που συνδέουν δύο κάθε φορά υπό εξέταση κόμβους. Εμείς εφαρμόζουμε τον *Katz Status index*, ο οποίος ορίζει το *σκορ ομοιότητας* μεταξύ δύο κόμβων V_x και V_y αθροίζοντας τα μονοπάτια διαφορετικού μήκους ℓ που συνδέουν τον V_x με τον V_y, όπως απεικονίζεται στην ακόλουθη Εξίσωση 8.5:

$$Katz_\beta = \sum_{\ell=1}^{\infty} \beta^\ell |paths^\ell_{V_x,V_y}| \tag{8.5}$$

όπου $paths^\ell_{V_x,V_y}$ το σύνολο όλων των μονοπατιών μήκους-ℓ από τον κόμβο V_x προς τον V_y, τα οποία υπολογίζονται από τον *πίνακα γειτνίασης A*. Ο *Katz Status index* εκμεταλλεύεται το γεγονός ότι η ύψωση του *πίνακα γειτνίασης* στη δύναμη n παράγει τον αριθμό των μήκους-n μονοπατιών που συνδέουν ένα ζεύγος κόμβων. Επίσης, στην Εξίσωση 8.5 υπάρχει ο *συντελεστής εξασθένισης* (ή αλλιώς, *απόσβεσης*) β, ο οποίος σταθμίζει τα διαφορετικού μήκους μονοπάτια ανάλογα με τη συνεισφορά τους στο τελικό *σκορ ομοιότητας/συνάφειας* των κόμβων. Ο συντελεστής λοιπόν αυτός μπορεί να πάρει τιμές όπως $\beta < 1/\lambda$, όπου λ είναι η μεγαλύτερη απόλυτη *ιδιοτιμή* του *πίνακα γειτνίασης* A [Katz, 1953, Foster and Muth, 2002]. Πολύ μικρές τιμές του β οδηγούν σε υψηλότερα βάρη των κοντινότερων μονοπατιών. Το γεγονός αυτό κάνει τη συνάρτηση συνολικής βαθμολογίας να εστιάζει περισσότερο στους *κοντινούς γείτονες του υπό εξέταση κόμβου*, οπότε την καθιστά και περισσότερο αποτελεσματική. Από την άλλη, πολύ χαμηλές τιμές του β έχουν ως αποτέλεσμα τα πολύ μεγάλα μονοπάτια να συνεισφέρουν πολύ λιγότερο στη συνολική *ομοιότητα/συνάφεια* μεταξύ δύο κόμβων του γράφου. Η ακόλουθη Εξίσωση 8.6, λοιπόν, αποτελεί την αναλυτική μορφή του *Katz index* όταν αυτός εφαρμόζεται στον *πίνακα γειτνίασης A*:

$$Katz(\mathbf{A};\beta) = \beta\mathbf{A} + \beta^2\mathbf{A}^2 + \beta^3\mathbf{A}^3 + \ldots = \sum_{\ell=1}^{\infty} \beta^\ell \mathbf{A}^\ell, \tag{8.6}$$

η οποία εκφράζεται επίσης και με τη βοήθεια των πινάκων γραμμικής άλγεβρας ως εξής:

$$\sum_{\ell=1}^{\infty} \beta^\ell \mathbf{A}^\ell = (I - \beta\mathbf{A})^{-1} - I \tag{8.7}$$

όπου ο *ταυτοτικός πίνακας* I είναι ένας τετραγωνικός πίνακας διαστάσεων $n \times n$, ο οποίος περιέχει μονάδες στην κύρια διαγώνιό του και έχει το ίδιο μέγεθος n με τον *πίνακα γειτνίασης A*. Ο δε

παράγοντας εξασθένισης β είναι μια παράμετρος που παίρνει τιμές που εξασφαλίζουν τη σύγκλιση της παραπάνω ακολουθίας και επιτρέπουν τον υπολογισμό του αντιστρόφου του *πίνακα γειτνίασης* A. Επιλέγουμε β ίσο με $1/K$ -όπως κάνουν ο Foster και ο Muth [Foster and Muth, 2002]- και όπου K είναι το μεγαλύτερο άθροισμα γραμμών ή στηλών του πίνακα A.

Στο τρέχον παράδειγμα, λοιπόν, ας υποθέσουμε ότι υπολογίζουμε την ομοιότητα μεταξύ U_4 με U_2 και U_3 αντίστοιχα. Εφαρμόζουμε επομένως τον *δείκτη κατάταστασης Katz* στον *γράφο φιλίας* G, προκειμένου να παρέχουμε συστάσεις με βάση τον υπολογισμό ενός *πίνακα ομοιότητας/συνάφειας* μεταξύ των κόμβων του γράφου. Πιο συγκεκριμένα, υπολογίζουμε τον *δείκτη Katz* εφαρμόζοντας την Εξίσωση 8.7 στον *πίνακα γειτνίασης* A του Πίνακα 8.1. Ο *δείκτης Katz* υπολογίζει την ομοιότητα μεταξύ δύο κόμβων λαμβάνοντας υπόψη μόνο μονοπάτια μήκους $\ell > 1$.

	U_1	U_2	U_3	U_4
U_1	0	0.1636	0.4909	0.4364
U_2	0.1636	0	0.4364	0.0545
U_3	0.4909	0.4364	0	0.1636
U_4	0.4364	**0.0545**	**0.1636**	0

Πίνακας 8.2: *Ο Πίνακας ομοιότητας/συνάφειας χρήστη-χρήστη* που υπολογίζουμε για τα δεδομένα του Παραδείγματος 8.2.

Παρατηρήστε ότι η ομοιότητα μεταξύ των U_4 και U_2 υπολογίζεται με βάση το μοναδικό μονοπάτι που τους συνδέει (4⟶1⟶3⟶2), όπως φαίνεται και στο Σχήμα 8.3. Σύμφωνα με τις προβλέψεις ομοιότητας χρηστών του Πίνακα 8.2, το μονοπάτι μήκους 3 βημάτων συνεισφέρει ένα *σκορ ομοιότητας* ίσο με 0.0545. Όσον αφορά στην ομοιότητα μεταξύ U_4 και U_3, υπάρχει μόνο ένα μονοπάτι μήκους 2 βημάτων (4⟶1⟶3) -όπως φαίνεται και στο Σχήμα 8.3- το οποίο αντιστοιχεί σε *σκορ ομοιότητας* ίσο με 0.1636.

Οι *τιμές ομοιότητας χρήστη-χρήστη* του Πίνακα 8.2 αποτυπώνουν τις μελλοντικές σχέσεις φιλίας που προβλέπουμε για το κοινωνικό δίκτυο του Παραδείγματος 8.2. Οι δε γραμμές του *πίνακα ομοιότητας* δείχνουν την "εγγύτητα" ή αλλιώς "συνάφεια" μεταξύ των χρηστών που προβλέπουμε προς σύσταση. Υπάρχει συνεπώς σαφής ένδειξη από τον παραπάνω *πίνακα ομοιότητας* ότι ο U_4 θα πρέπει να προταθεί ως φίλος στον U_3 έναντι του U_2, εφόσον έχει υψηλότερη *τιμή ομοιότητας*, $0.1636 > 0.0545$, μ' αυτόν.

8.4.2 SimRank για Μονομερείς Γράφους

Οι Jeh και Widom [Jeh and Widom, 2002] πρότειναν το 2002 ένα μέτρο *ομοιότητας/συνάφειας* των κόμβων ενός γράφου γνωστό ως *SimRank*, το οποίο μπορεί να χρησιμοποιηθεί για τη μέτρηση ομοιοτήτων είτε μεταξύ κόμβων ιδίου τύπου (π.χ. ομοιότητες μεταξύ χρηστών λαμβάνοντας υπόψη

τις σχέσεις φιλίας τους) είτε μεταξύ διαφορετικών τύπων κόμβων (π.χ. ομοιότητες μεταξύ χρηστών λαμβάνοντας υπόψη τα στοιχεία με τα οποία αυτοί αλληλεπιδρούν). Η λογική του *SimRank* είναι η ακόλουθη:

*Δύο κόμβοι ενός γράφου είναι **συναφείς** εφόσον αναφέρονται από **συναφείς** κόμβους.*

Παραδείγματος χάρη, δύο άρθρα είναι *συναφή* εάν διαβάζονται από *συναφείς* χρήστες.

Ορισμός 8.3 ***SimRank***. *Η ομοιότητα μεταξύ δύο κόμβων a και b ενός μονομερούς κατευθυνόμενου γράφου G ορίζεται ως εξής:*

$$s(a,b) = \begin{cases} 1, & \text{Αν } a = b; \\ \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)), & \text{Αν } a \neq b, \text{ και } (I(a) \neq \emptyset \text{ ή } I(b) \neq \emptyset); \\ 0, & \text{Αν } I(a) = \emptyset \text{ ή } I(b) = \emptyset, \end{cases} \tag{8.8}$$

*όπου $C \in [0,1]$ ο **συντελεστής εξασθένισης** ή αλλιώς, **απόσβεσης**, $I(a)$ οι εισερχόμενοι γειτονικοί κόμβοι του a, και $I(b)$ οι εισερχόμενοι γειτονικοί κόμβοι του b.*

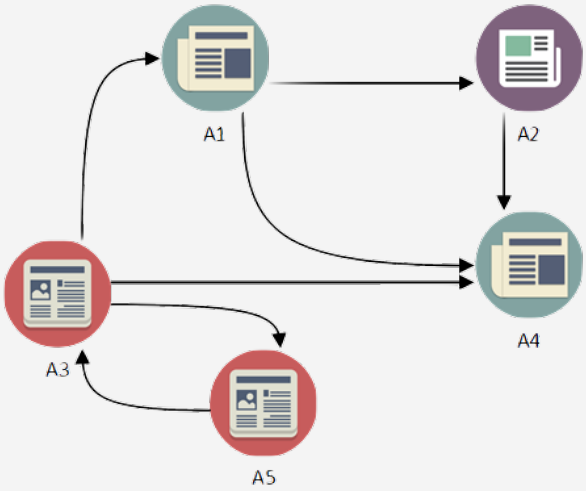
Με βάση την Εξίσωση 8.8 για τον υπολογισμό της *συνάφειας* $s(a,b)$ των κόμβων a και b, αθροίζουμε σωρευτικά την *ομοιότητα* $s(I_i(a), I_j(b))$ που προκύπτει από όλα τα δυνατά ζεύγη των *εισερχόμενων* (*in-coming*) γειτονικών κόμβων $(I_i(a), I_j(b))$ τους. Στη συνέχεια, διαιρούμε με τον συνολικό αριθμό των πιθανών ζευγαριών των *εισερχόμενων γειτονικών κόμβων* $|I(a)||I(b)|$, για να κανονικοποιήσουμε την υπολογισμένη σωρευτική ομοιότητα. Μ' άλλα λόγια, η *ομοιότητα* μεταξύ των κόμβων a και b είναι η *μέση ομοιότητα* μεταξύ των εισερχόμενων γειτόνων του a και των εισερχόμενων γειτόνων του b. Η Εξίσωση 8.8 υπολογίζεται επαναληπτικά και συνήθως συγκλίνει γρήγορα (μετά την τέταρτη ή πέμπτη επανάληψη).

Παράδειγμα 8.3 Μας δίνεται ο *μονομερής κατευθυνόμενος γράφος* του Σχήματος 8.4, οι κόμβοι του οποίου είναι ειδησεογραφικά άρθρα που παραπέμπουν το ένα στο άλλο. Ζητείται: (α) Να εφαρμόσουμε την Εξίσωση 8.8, ώστε να υπολογίσουμε την τιμή *συνάφειας* μεταξύ των άρθρων A4 και A5, όπως και μεταξύ των άρθρων A1 και A5 μετά την πρώτη μόνο επανάληψη του *αλγορίθμου SimRank*. (β) Να υπολογίσουμε τη *συνάφεια* μεταξύ όλων των άρθρων του γράφου εκτελώντας κατ' επανάληψη τον *αλγόριθμο SimRank* μέχρι αυτός να συγκλίνει. Οι τιμές εισόδου των παραμέτρων του *αλγορίθμου SimRank* είναι οι παρακάτω: *συντελεστής απόσβεσης* $C = 0.8$ και *κριτήριο σύγκλισης* $\varepsilon \leqslant 10^{-4}$.

Παρατηρήστε στο Σχήμα 8.4 ότι οι *εισερχόμενοι γειτονικοί κόμβοι* που αναφέρουν το υπό εξέταση άρθρο A4 είναι τα άρθρα A1, A2, και A3. Επίσης, ο *εισερχόμενος γειτονικός κόμβος* που αναφέρει το υπό εξέταση άρθρο A5 είναι το άρθρο A3. Σημειώνουμε εδώ ότι οι υπό εξέταση

κόμβοι A4 και A5 έχουν από κοινού το εισερχόμενο γειτονικό άρθρο A3 να τους αναφέρει. Βάσει των παραπάνω δεδομένων, με την Εξίσωση 8.8 μπορούμε αρχικά να υπολογίσουμε τη σωρευτική ομοιότητα που προκύπτει από όλους τους δυνατούς συνδυασμούς των *εισερχόμενων γειτονικών κόμβων* που αναφέρουν έστω έναν από τους δύο υπό εξέταση κόμβους:

$$\sum_{i=1}^{|I(A4)|}\sum_{j=1}^{|I(A5)|} s\Big(I_i(A4), {}_j(A5)\Big) = s(A1,A3) + s(A2,A3) + s(A3,A3) = 0+0+1 = 1.$$



Σχήμα 8.4: Ένας *μονομερής γράφος* με ειδησεογραφικά άρθρα.

Τέλος, εφαρμόζοντας τον *κανονικοποιημένο συντελεστή εξασθένισης* $\frac{C}{|(A4)||(A5)|}$ της Εξίσωσης 8.8 η *συνάφεια* μεταξύ των άρθρων A4 και A5 καταλήγει να είναι η ακόλουθη:

$$s(A4,A5) = \frac{0.8}{3*1} * 1 = 0.267.$$

Άρα, η *συνάφεια* μεταξύ των άρθρων $A4$ και $A5$ είναι ίση με 0.267. Παρατηρήστε στο Σχήμα 8.4 ότι τα άρθρα A1 και A5 έχουν μόνο έναν από κοινού *εισερχόμενο γειτονικό κόμβο* (το ειδησεογραφικό άρθρο A3) που τους αναφέρει. Επομένως η *συνάφειά* τους είναι η ακόλουθη:

$$s(A1,A5) = \frac{0.8}{1*1} * 1 = 0.8.$$

Συνεπώς, η *ομοιότητα* μεταξύ των άρθρων $A1$ με $A5$ είναι μεγαλύτερη απ' ό,τι αυτή μεταξύ των άρθρων A4 με A5 που είχαμε υπολογίσει προηγουμένως, επειδή το A4 έχει 3 *εισερχόμενους (in-coming) γειτονικούς κόμβους*, ενώ το A1 έχει μόνο 1 *εισερχόμενο γειτονικό κόμβο*.

Στη συνέχεια, επί των δεδομένων του Παραδείγματος 8.3 εφαρμόζουμε τον *αλγόριθμο SimRank*, για να υπολογίσουμε τη *συνάφεια* μεταξύ όλων των κόμβων του γράφου. Τα αποτελέσματα *συνάφειας* μεταξύ των άρθρων εμφανίζονται στον Πίνακα 8.3. Παρατηρώντας το

Σχήμα 8.4 μας δημιουργείται η αίσθηση ότι τα άρθρα $A3$ και $A5$ πρέπει να είναι πιο *συναφή*, επειδή και τα δύο "δείχνουν" (παραπέμπουν) το ένα το άλλο. Ωστόσο, ο *αλγόριθμος SimRank* δεν μπορεί να αποτυπώσει αυτήν την αίσθηση, επειδή λαμβάνει υπόψη μόνο τους *κοινούς εισερχόμενους γειτονικούς κόμβους*, που αναφέρουν τους δύο υπό εξέταση κόμβους. Έτσι, επειδή στο παράδειγμά μας τα άρθρα $A3$ και $A5$ δεν έχουν κανέναν *κοινό εισερχόμενο γειτονικό κόμβο*, οδηγούμαστε στον υπολογισμό μηδενικής *ομοιότητας* μεταξύ τους, όπως φαίνεται στον Πίνακα 8.3. Και αυτό είναι ένα μειονέκτημα του *αλγορίθμου SimRank*.

	A1	A2	A3	A4	A5
A1	1.000	0.000	0.000	0.437	0.800
A2	0.000	1.000	0.640	0.267	0.000
A3	0.000	0.640	1.000	0.213	0.000
A4	0.437	0.267	0.213	1.000	0.437
A5	0.800	0.000	0.000	0.437	1.000

Πίνακας 8.3: *Πίνακας ομοιότητας/συνάφειας* μεταξύ όλων των άρθρων του γράφου που υπολογίζουμε βάσει του *αλγορίθμου SimRank* επί των δεδομένων του Παραδείγματος 8.3.

8.4.3 Αλγόριθμος SimRank για Διμερείς Γράφους

Στους *διμερείς γράφους* έχουμε δύο διαφορετικούς τύπους κόμβων (π.χ. χρήστες και άρθρα). Η λογική του *SimRank* για *διμερείς γράφους* είναι παρόμοια και για τους *μονομερείς*. Δηλαδή:

*Δύο άρθρα είναι **συναφή** εφόσον έχουν διαβαστεί από **συναφείς** χρήστες.*
*Δύο χρήστες είναι **συναφείς** εφόσον διαβάζουν **συναφή** άρθρα.*

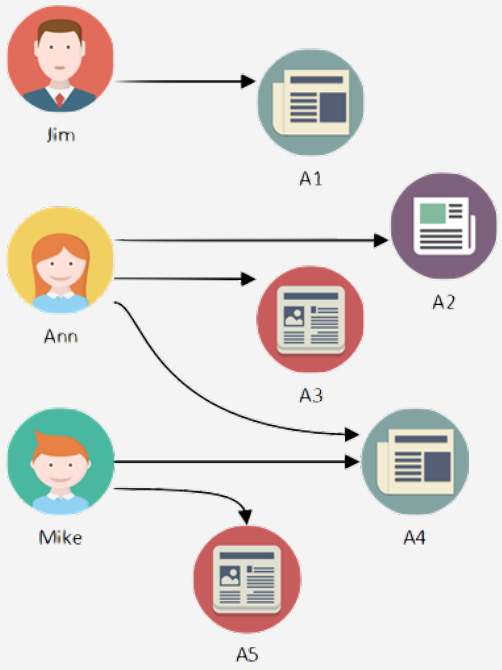
Η Εξίσωση 8.8, λοιπόν, βρίσκει την *ομοιότητα* μεταξύ των άρθρων (δηλαδή μόνο οι *εισερχόμενοι γειτονικοί κόμβοι* έχουν σημασία). Για δε τους *διμερείς γράφους χρηστών-άρθρων*, η μόνη διαφορά στη λογική είναι ότι οι *εισερχόμενοι γειτονικοί κόμβοι* δεν αντιπροσωπεύουν πλέον τις αναφορές από άλλα άρθρα, αλλά την αναγνωστική συμπεριφορά των χρηστών. Επίσης, για τον υπολογισμό της *ομοιότητας* μεταξύ των χρηστών (*ομοιότητα χρήστη-χρήστη*) οι *εισερχόμενοι γειτονικοί κόμβοι* αντικαθίστανται από τους *εξερχόμενους (out-going) γειτονικούς κόμβους*. Να σημειωθεί ότι αντί να ελέγχουμε πάντα τον τύπο των κόμβων (χρήστες ή στοιχεία), αρκεί να εξετάζουμε μόνο την ύπαρξη ενός *συνδέσμου* ή αλλιώς, *κατευθυνόμενης ακμής*. Επειδή λοιπόν κάθε σύνδεσμος στους *διμερείς γράφους χρηστών-άρθρων* συνδέει πάντα μόνο χρήστες με άρθρα, αυτό σημαίνει ότι κάθε σύνδεσμος μπορεί να αντιμετωπιστεί ως *εξερχόμενος* για έναν χρήστη και ως *εισερχόμενος* για ένα άρθρο.

Ορισμός 8.4 ***SimRank για διμερείς γράφους.*** *Η ομοιότητα μεταξύ δύο κόμβων e και f ενός διμερούς γράφου G ορίζεται ως εξής:*

$$s(e,f) = \begin{cases} 1, & \text{Αν } e = f \\ \frac{C}{|N(e)||N(f)|} \sum_{i=1}^{|N(e)|} \sum_{j=1}^{|N(f)|} s(N_i(e), N_j(f)), & \text{Αν } e \neq f \text{ και } (N(e) \neq \emptyset \text{ ή } N(f) \neq \emptyset) \\ 0, & \text{if } N(e) = \emptyset \text{ ή } N(f) = \emptyset, \end{cases} \quad (8.9)$$

όπου $N(e)$ και $N(f)$ τα σύνολα των γειτονικών κόμβων των e και f αντίστοιχα, και $C \in [0,1]$. Δηλαδή, εάν ένας κόμβος είναι τύπου "χρήστης", οι εξερχόμενοί του γείτονες είναι τα άρθρα με τα οποία έχει αυτός αλληλεπιδράσει. Και κατ' αναλογία, εάν ένας κόμβος είναι τύπου "άρθρο", τότε οι εισερχόμενοί του γείτονες είναι οι χρήστες που έχουν αλληλεπιδράσει μ' αυτό. Μ' άλλα λόγια, στους *διμερείς γράφους* δεν εξετάζουμε μόνο *εισερχόμενους γειτονικούς κόμβους (in-coming neighbor nodes)*, αλλά απλά την ύπαρξη *γειτονικών κόμβων (neighbor nodes)*, οι οποίοι μπορεί να είναι είτε *εισερχόμενοι (in-coming)* προς τον υπό εξέταση κόμβο είτε *εξερχόμενοι (out-going)* από τον υπό εξέταση κόμβο.

Παράδειγμα 8.4 Μας δίνεται ο *διμερής γράφος χρήστη-άρθρου* του Σχήματος 8.5, ο οποίος εμφανίζει τα άρθρα της διαδικτυακής εφημερίδας με τα οποία αλληλεπίδρασαν οι χρήστες.



Σχήμα 8.5: Παράδειγμα ενός διμερούς δικτύου χρήστη-άρθρου

Μας ζητείται να υπολογίσουμε τη *συνάφεια* μεταξύ όλων των άρθρων και μεταξύ όλων των χρηστών του *διμερούς γράφου*, εκτελώντας επαναληπτικά τον *αλγόριθμο SimRank* μέχρι αυτός

να συγκλίνει. Οι τιμές εισόδου των παραμέτρων του *αλγορίθμου SimRank* είναι οι παρακάτω: *συντελεστής απόσβεσης* $C = 0.8$ και *κριτήριο σύγκλισης* $\varepsilon \leqslant 10^{-4}$.
Εν συνεχεία, και επί των δεδομένων του Παραδείγματος 8.4, εφαρμόζουμε τον *αλγόριθμο SimRank*, για να υπολογίσουμε τη *συνάφεια* μεταξύ όλων των κόμβων του γράφου. Τα αποτελέσματα *συνάφειας* μεταξύ των άρθρων και μεταξύ των χρηστών εμφανίζονται στο Σχήμα 8.6.

	Jim	Ann	Mike	A1	A2	A3	A4	A5
Jim	1.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Ann	0.0	1.000	0.468	0.0	0.000	0.000	0.000	0.000
Mike	0.0	0.468	1.000	0.0	0.000	0.000	0.000	0.000
A1	0.0	0.000	0.000	1.0	0.000	0.000	0.000	0.000
A2	0.0	0.000	0.000	0.0	1.000	0.800	0.587	0.374
A3	0.0	0.000	0.000	0.0	0.800	1.000	0.587	0.374
A4	0.0	0.000	0.000	0.0	0.587	0.587	1.000	0.587
A5	0.0	0.000	0.000	0.0	0.374	0.374	0.587	1.000

Σχήμα 8.6: *Πίνακας ομοιότητας/συνάφειας* μεταξύ όλων των άρθρων και μεταξύ όλων των χρηστών του *διμερούς γράφου* που υπολογίζουμε βάσει του *αλγορίθμου SimRank* επί των δεδομένων του Παραδείγματος 8.4.

Παρατηρούμε στον πίνακα ομοιοτήτων του Σχήματος 8.6 ότι η *ομοιότητα* μεταξύ των άρθρων $A2$ και $A3$ είναι 0.8, επειδή διαβάζονται από τον ίδιο χρήστη, δηλαδή την Ann: $\frac{0.8}{1*1} * 1$. Επίσης, η Ann και ο Mike έχουν μεγαλύτερη *συνάφεια* μεταξύ τους συγκριτικά με το πόσο μοιάζουν και οι δύο με τον Jim, επειδή έχουν διαβάσει ένα κοινό άρθρο, το A4. Τέλος, το άρθρο A2 είναι *συναφές* με το άρθρο A5 επειδή και τα δύο τα έχουν διαβάσει *συναφείς* χρήστες, δηλαδή η Ann και ο Mike.

8.4.4 Αλγόριθμος Random Walk with Restart

Ο *αλγόριθμος Random Walk with Restart (RWR)* [Pan et al., 2004, Tong et al., 2006] είναι μια παραλλαγή του γνωστού *αλγορίθμου PageRank*. Ο *RWR* έχει ιδιότητες που μπορούν να αποτυπώσουν επαρκώς την *ομοιότητα* μεταξύ του υπό εξέταση κόμβου N και των υπολοίπων κόμβων ενός γράφου. Το κύριο πλεονέκτημα του *RWR* έναντι του *PageRank* είναι το χαρακτηριστικό της "τηλεμεταφοράς", το οποίο υποχρεώνει τον τυχαίο "περιπατητή" να ξεκινήσει εκ νέου τον "περίπατό" του από τον αρχικό κόμβο N, ενώ στον *PageRank* ο "περιπατητής" μεταπηδά σε οποιονδήποτε κόμβο του γράφου. Όπως είναι αναμενόμενο, ο *RWR* αποδίδει μεγαλύτερη *συνάφεια/ομοιότητα* στους κοντι-

νούς κόμβους του N. Έτσι, εάν δύο κόμβοι βρίσκονται κοντά ο ένας στον άλλο, η πιθανότητα να συνδεθούν μέσω μιας ακμής στο μέλλον είναι μεγαλύτερη. Στα κοινωνικά δίκτυα, ο *RWR* μπορεί να αποτυπώσει την έννοια της *συνάφειας* μεταξύ χρηστών που μοιράζονται μεγάλο αριθμό κοινών φίλων. Για έναν *διμερές γράφο χρήστη-στοιχείου* λοιπόν, όταν δύο χρήστες αλληλεπιδρούν με τα ίδια στοιχεία, τότε η συνολική πιθανότητα σύνδεσής τους (μέσω ενός κόμβου φιλίας) αυξάνεται.

Στην υποενότητα, λοιπόν, αυτή θα μελετήσουμε τον *αλγόριθμο RWR* κατά την εφαρμογή του σε *ετερογενείς γράφους* με διαφορετικούς τύπους κόμβων. Μπορούμε, ως εκ τούτου, να φανταστούμε έναν "τυχαίο περιπατητή" ο οποίος από οποιονδήποτε κόμβο v_i κινείται προς οποιονδήποτε γειτονικό του κόμβο με ίση πιθανότητα, ενώ σε κάθε βήμα επανάληψης με κάποια ορισμένη πιθανότητα μπορεί να επιστρέψει στον αρχικό κόμβο N. Μ' αυτόν τον τρόπο θέλουμε να προσδιορίσουμε το πού θα καταλήξει ο "τυχαίος περιπατητής" μετά από σύντομους "περιπάτους" που ξεκινούν από τον υπό εξέταση αρχικό κόμβο N. Άρα, θέλουμε να προβλέψουμε ποιοι κόμβοι του γράφου είναι πιο *συναφείς* με τον υπό εξέταση αρχικό κόμβο N. Αυτό το μέτρο ομοιότητας αναφέρεται ως **Personalized PageRank** (*P-PageRank*) [Page and Brin, 1998], **Topic-Sensitive PageRank** [Haveliwala, 2002] ή ως **Random Walk with Restart** (*RWR*) [Leskovec et al., 2019].

Μετά την εκτέλεση του *αλγορίθμου RWR* σ' έναν γράφο με n κόμβους, παίρνουμε ως αποτέλεσμα έναν πίνακα ομοιοτήτων $\mathbf{V}$ διαστάσεων $n \times n$, όπου κάθε στοιχείο v_{ij} αντιστοιχεί στον βαθμό *συνάφειας* ενός κόμβου i προς τον κόμβο j. Σ' αντίθεση με τον πίνακα ομοιότητας που υπολογίζει ο *αλγόριθμος SimRank*, ο πίνακας ομοιότητας που προκύπτει από τον *αλγόριθμο RWR* δεν είναι συμμετρικός. Αυτό σημαίνει ότι η *συνάφεια* ενός κόμβου i προς τον κόμβο j δεν είναι ίση με τη *συνάφεια* του κόμβου j προς τον κόμβο i. Συνεπώς, ο "τυχαίος περιπατητής" είναι πιο πιθανό να καταλήξει σ' έναν "πολύ καλά" συνδεδεμένο κόμβο από ό,τι σ' έναν "όχι καλά" συνδεδεμένο.

Έστω λοιπόν ότι αναπαριστούμε με τη *μεταβλητή* $\mathbf{M}$ τον *πίνακα πιθανοτήτων μετάβασης* (*transition probability matrix*) ενός *μονομερούς γράφου* G. Δηλαδή, το στοιχείο στη γραμμή i και στη στήλη j του $\mathbf{M}$ ισούται με $1/k$ όπου k είναι ο *βαθμός του κόμβου* j (*node's degree*), ενώ ο κόμβος i είναι ένας από τους γειτονικούς του κόμβους. Εάν $k = 0$, τότε το στοιχείο αυτό ισούται με 0. Τονίζουμε ότι ο *πίνακας πιθανοτήτων μετάβασης* $\mathbf{M}$ προκύπτει από τον πολλαπλασιασμό του *πίνακα γειτνίασης* $\mathbf{A}$ με τον αντίστροφό του πίνακα βαθμών των κόμβων $\mathbf{D^{-1}}$ του γράφου μας. Επιπρόσθετα, έστω ότι ορίζουμε με τη μεταβλητή β την πιθανότητα γι' έναν "περιπατητή" να συνεχίσει τυχαία τον "περίπατό" του, και κατά συνέπεια με $1 - \beta$ την πιθανότητα να διακόψει τον "περίπατό" του και να μεταπηδήσει στον αρχικό κόμβο Ν. Έστω, ακόμη, $\mathbf{e_N}$ ένα διάνυσμα στήλης με 1 στη γραμμή για τον κόμβο N, και 0 σε οποιαδήποτε άλλη θέση του. Τέλος, έστω ότι $\mathbf{v}$ είναι ένα διάνυσμα στήλης που αντικατοπτρίζει την πιθανότητα ο "περιπατητής" να βρίσκεται σ' έναν κόμβο σε μια συγκεκριμένη χρονική στιγμή, και $\mathbf{v}'$ είναι η πιθανότητα ο "περιπατητής" να βρίσκεται σε κάποιον άλλο κόμβο την αμέσως επόμενη χρονική στιγμή. Τότε το διάνυσμα $\mathbf{v}'$ σχετίζεται με το $\mathbf{v}$ ως εξής:

$$\mathbf{v}' = \beta \cdot \mathbf{M} \cdot \mathbf{v} + (1 - \beta) \cdot \mathbf{e_N} \tag{8.10}$$

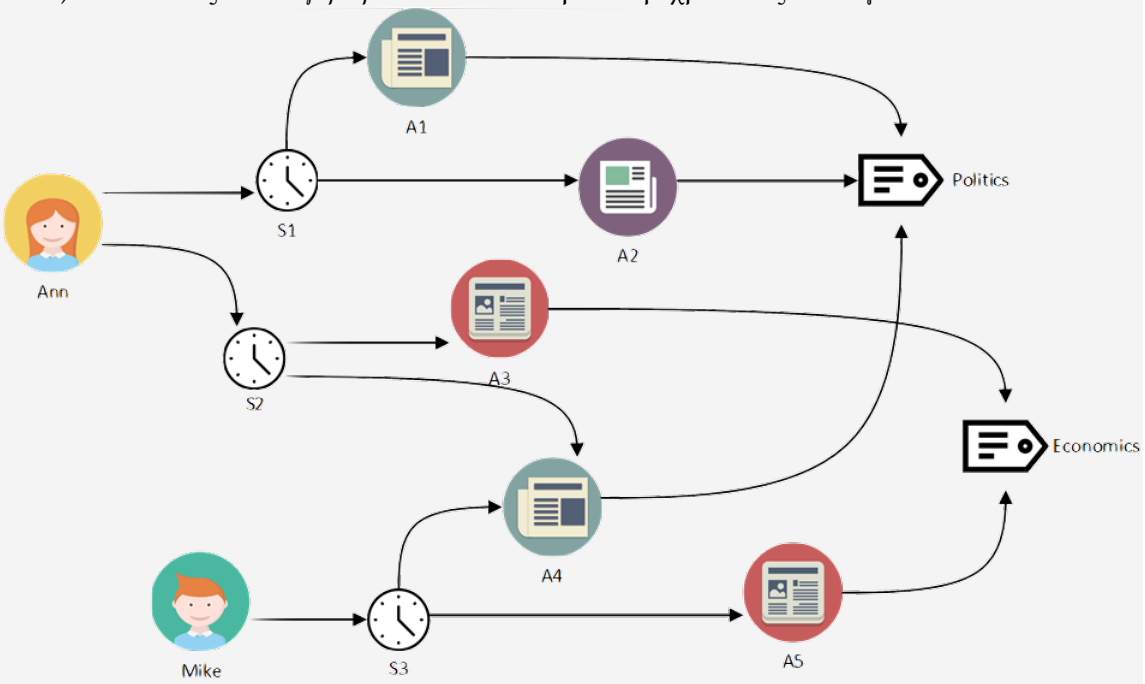
Έπειτα, ορίζουμε την Εξίσωση 8.11 η οποία μετασχηματίζει την Εξίσωση 8.10 ώστε να εφαρμοστεί σ' ολόκληρο τον γράφο G (και όχι μόνο σ' έναν κόμβο του γράφου), υπολογίζοντας κατά προσέγγιση τον πίνακα ομοιοτήτων μεταξύ των κόμβων του γράφου με τη μέθοδο του *power iteration*.

Ορισμός 8.5 ***Random Walk with Restart***. *Η μέθοδος RWR ορίζει έναν μη συμμετρικό και κανονικοποιημένο κατά γραμμές πίνακα συνάφειας* $\mathbf{V}'$ *όπου το άθροισμα των πιθανοτήτων ανά γραμμή ισούται με τη μονάδα. Ο πίνακας* $\mathbf{V}'$ *αποθηκεύει τις πιθανότητες μετάβασης μεταξύ δύο οποιονδήποτε κόμβων του γράφου και μπορεί να υπολογιστεί επαναληπτικά χρησιμοποιώντας τον ακόλουθο τύπο:*

$$\mathbf{V}' = \beta \cdot \mathbf{M} \cdot \mathbf{V} + (1 - \beta) \cdot \mathbf{I_n} \tag{8.11}$$

όπου $\mathbf{I_n}$ *ο ταυτοτικός πίνακας διαστάσεων* $n \times n$*, ενώ* $\mathbf{V}$ *και* $\mathbf{V}'$ *είναι οι κανονικοποιημένοι πίνακες πιθανοτήτων μετάβασης για δύο διαδοχικές χρονικές στιγμές.*

Παράδειγμα 8.5 Δίνεται ο *ετερογενής γράφος* του Σχήματος 8.7, ο οποίος αποτελείται από 4 διαφορετικούς τύπους κόμβων (χρήστες, συνεδρίες, άρθρα, κατηγορίες άρθρων), και μας ζητείται να υπολογίσουμε τη *συνάφεια* μεταξύ όλων των τύπων των κόμβων του γράφου (ανά τύπο) εκτελώντας τον *αλγόριθμο RWR* επαναληπτικά μέχρι αυτός να συγκλίνει.



Σχήμα 8.7: Ο *ετερογενής γράφος* μιας διαδικτυακής εφημερίδας.

Μετά την εφαρμογή του *αλγορίθμου RWR* στα δεδομένα του Παραδείγματος 8.5, τα αποτελέσματα εμφανίζονται στο Σχήμα 8.8. Το κύριο πλεονέκτημα του πίνακα ομοιοτήτων είναι ότι απο-

τελείται από υποπίνακες, οι οποίοι είναι χρήσιμοι για διάφορους σκοπούς. Ο υποπίνακας *Article-Article* (τρίτος στη διαγώνιο από πάνω αριστερά στο Σχήμα 8.8), λόγου χάρη, δείχνει τον *πίνακα πιθανοτήτων μετάβασης* από ένα άρθρο σ' ένα άλλο άρθρο, και μπορεί να εφαρμοστεί άμεσα για την *πρόβλεψη του επόμενου κλικ ενός χρήστη* βάσει του άρθρου με το οποίο αυτός αλληλεπιδρά την τρέχουσα χρονική στιγμή.

Ο υποπίνακας *User-Article Category* (δεξιά πάνω γωνία στο Σχήμα 8.8), με τη σειρά του, αποκαλύπτει τις γενικές προτιμήσεις των χρηστών, καθορίζοντας πόσο πιθανό είναι για τον χρήστη να διαβάσει ένα άρθρο της μίας ή της άλλης κατηγορίας άρθρων (π.χ Πολιτική ή Οικονομικά). Μια ενδιαφέρουσα και όχι τόσο εμπειρική παρατήρηση εμφανίζεται στον υποπίνακα *χρήστη-κατηγορίας άρθρων*: από το Σχήμα 8.7 λοιπόν βλέπουμε ότι ο Mike έχει διαβάσει ένα άρθρο από κάθε κατηγορία (Πολιτική και Οικονομία). Παρόλα αυτά, προκύπτει από τον *πίνακα συνάφειας* του Σχήματος 8.5 ότι το ενδιαφέρον του για τα Οικονομικά είναι πολύ μεγαλύτερο από ό,τι για την Πολιτική. Ο λόγος είναι ότι το άρθρο A4 που διάβασε ο Mike στη συνεδρία S3 εμφανίζεται στη συνεδρία S2 της Ann μαζί με το A3, το οποίο είναι ένα οικονομικό άρθρο, φέρνοντας έτσι το A4 πιο κοντά στην κατηγορία των Οικονομικών και, κατά συνέπεια, αυξάνοντας το ενδιαφέρον του Mike προς τα Οικονομικά όπως προβλέπουμε. Τέλος, σημειώνουμε ότι το Παράδειγμα 8.5 θα λυθεί και ως προγραμματιστική άσκηση με τη γλώσσα *Python* στο τέλος αυτού του κεφαλαίου.

	Ann	Mike	S1	S2	S3	A1	A2	A3	A4	A5	Politics	Economics
Ann	0.2838	0.0088	0.1629	0.1529	0.0336	0.0610	0.0610	0.0521	0.0670	0.0203	0.0673	0.0294
Mike	0.0175	0.2763	0.0199	0.0473	0.2869	0.0154	0.0154	0.0332	0.0990	0.0971	0.0393	0.0526
S1	0.1086	0.0066	0.3389	0.0668	0.0241	0.1201	0.1201	0.0239	0.0542	0.0125	0.1099	0.0141
S2	0.1019	0.0158	0.0668	0.3137	0.0584	0.0334	0.0334	0.1074	0.1150	0.0393	0.0567	0.0582
S3	0.0224	0.0956	0.0241	0.0584	0.3578	0.0198	0.0198	0.0421	0.1245	0.1219	0.0484	0.0652
A1	0.0610	0.0077	0.1802	0.0501	0.0296	0.2939	0.0939	0.0180	0.0669	0.0125	0.1736	0.0126
A2	0.0610	0.0077	0.1802	0.0501	0.0296	0.0939	0.2939	0.0180	0.0669	0.0125	0.1736	0.0126
A3	0.0521	0.0166	0.0358	0.1611	0.0631	0.0180	0.0180	0.3037	0.0680	0.0776	0.0331	0.1529
A4	0.0447	0.0330	0.0542	0.1150	0.1245	0.0446	0.0446	0.0453	0.2938	0.0479	0.1147	0.0377
A5	0.0203	0.0485	0.0188	0.0589	0.1829	0.0125	0.0125	0.0776	0.0718	0.3106	0.0298	0.1557
Politics	0.0449	0.0131	0.1099	0.0567	0.0484	0.1158	0.1158	0.0221	0.1147	0.0199	0.3225	0.0164
Economics	0.0294	0.0263	0.0212	0.0874	0.0977	0.0126	0.0126	0.1529	0.0566	0.1557	0.0245	0.3230

Σχήμα 8.8: *Πίνακας συνάφειας* μεταξύ των κόμβων του γράφου του Παραδείγματος 8.5.

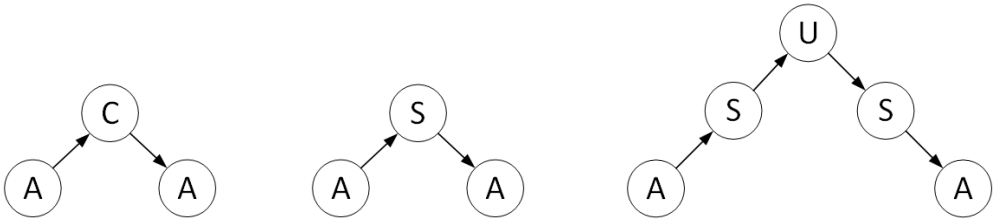
8.4.5 Αλγόριθμος PathSim

Ο Yizhou και οι λοιποί συν-συγγραφείς [Yizhou et al., 2011] πρότειναν μια νέα ιδέα για τη μέτρηση της *ομοιότητας* μεταξύ των κόμβων ενός δικτύου, η οποία βασίζεται στην ανάλυση των *μετα-*

μονοπατιών (*meta paths*) μέσω των οποίων συνδέονται οι κόμβοι. Σ' ένα ετερογενές δίκτυο δύο κόμβοι μπορεί να συνδέονται μέσω διαφορετικών μονοπατιών. Για παράδειγμα, δύο άρθρα μπορούν να συνδεθούν μέσω του μονοπατιού "άρθρο-κατηγορία-άρθρο" (*ομοιότητα βάσει περιεχομένου*), "άρθρο-συνεδρία-άρθρο" (*ομοιότητα βάσει συνεδρίας*) και "άρθρο-συνεδρία-χρήστης-συνεδρία-άρθρο" (*ομοιότητα βάσει συνεργατικού φιλτραρίσματος*). Χρησιμοποιώντας διαφορετικά μονοπάτια, παρατηρούνται διαφορετικού βαθμού και έντασης *ομοιότητες*. Αυτά τα μονοπάτια ονομάζονται *meta paths* όπως προαναφέραμε και ορίζονται τυπικά ως εξής:

Ορισμός 8.6 ***Meta Path****. [Yizhou et al., 2011] Ένα μετα-μονοπάτι $\mathcal{P}$ είναι ένα μονοπάτι που ορίζεται στον γράφο του σχήματος δικτύου $T_G = (\mathcal{Q}, \mathcal{R})$, και συμβολίζεται με τη μορφή $Q_1 \xrightarrow{R_1} Q_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} Q_{l+1}$, η οποία ορίζει μια σύνθετη σχέση $R = R_1 \circ R_2 \circ \ldots \circ R_l$ μεταξύ των διαφορετικών τύπων κόμβων Q_1 και Q_{l+1}, όπου το $\circ$ συμβολίζει τον τελεστή σύνθεσης των μεταξύ τους συσχετίσεων ή αλλιώς ακμών.*

Υπάρχουν πολλά διαφορετικά *μετα-μονοπάτια* που μπορούν να αναζητηθούν στον γράφο του Σχήματος 8.7, του οποίου το *σχήμα δικτύου* είναι USAC (User, Session, Article, και Category). Εάν ξεκινήσουμε από έναν κόμβο τύπου "άρθρο", μπορούμε τότε να κατασκευάσουμε και τα ακόλουθα *μετα-μονοπάτια* ACA, ASA, ASUSA, όπως φαίνεται στο Σχήμα 8.9. Κύριο πλεονέκτημα του θεωρητικού πλαισίου των *μετα-μονοπατιών* είναι η *επεξηγησιμότητα (explainability)* και, άρα, η δυνατότητα ελέγχου, ενώ επίσης παρέχει έναν ισχυρό μηχανισμό για τον χρήστη, ώστε να επιλέξει μια κατάλληλη σημασιολογία ως προς το τί θεωρεί ότι καθορίζει την *ομοιότητα* μεταξύ των κόμβων, διαλέγοντας ένα κατάλληλο *μετα-μονοπάτι*. Ένα *μέτρο ομοιότητας* λοιπόν που είναι σε θέση να αποτυπώσει τη σημασιολογία της *ομοιότητας* μεταξύ των κόμβων ενός δικτύου είναι το **PathSim**.



(α) *Meta Path*: ACA (β) *Meta Path*: ASA (γ) *Meta Path*: ASUSA

Σχήμα 8.9: Πιθανά *μετα-μονοπάτια* του δικτύου USAC.

Ορισμός 8.7 ***PathSim: Ένα μέτρο ομοιότητας βασισμένο σε Meta paths****. [Yizhou et al., 2011] Δοθέντος ενός συμμετρικού μετα-μονοπατιού P, το μέτρο ομοιότητας PathSim μεταξύ δύο κόμβων του ίδιου τύπου x και y είναι:*

$$s(x,y) = \frac{2 * |p_{x \rightsquigarrow y} : p_{x \rightsquigarrow y} \in P|}{|p_{x \rightsquigarrow x} : p_{x \rightsquigarrow x} \in P| + |p_{y \rightsquigarrow y} : p_{y \rightsquigarrow y} \in P|} \tag{8.12}$$

όπου $p_{x \rightsquigarrow y}$ ένα μονοπάτι που ξεκινά από τον κόμβο αφετηρίας x και καταλήγει στον κόμβο προορισμού y, $p_{x \rightsquigarrow x}$ ένα μονοπάτι που ξεκινά από τον κόμβο x και επιστρέφει πίσω στον ίδιο, και $p_{y \rightsquigarrow y}$ ένα μονοπάτι που ξεκινά από τον κόμβο y και επιστρέφει πίσω στον ίδιο.

Παράδειγμα 8.6 Δίνεται ο ετερογενής γράφος του Σχήματος 8.7 και μας ζητείται να υπολογίσουμε τις *ομοιότητες* μεταξύ των άρθρων χρησιμοποιώντας απλά *μετα-μονοπάτια*: ACA (*φιλτράρισμα βάσει περιεχομένου*), ASA (*φιλτράρισμα βάσει συνεδρίας*) και ASUSA (*συνεργατικό φιλτράρισμα*). Ο Πίνακας 8.4 δείχνει τις *ομοιότητες* του άρθρου A3 με τα υπόλοιπα άρθρα ως εξής: το A3 μοιράζεται μόνο την ίδια κατηγορία άρθρου με το A5, επομένως $\text{PathSim}_{\text{ACA}}(\text{A3}, \text{A5}) = \frac{2*1}{1+1} = 1$. Το δε A4 μοιράζεται *κοινές συνεδρίες* (*sessions*) με το A3 και το A5, επομένως $\text{PathSim}_{\text{ASA}}(\text{A3}, \text{A4}) = \frac{2*1}{1+2} \approx 0.666$. Ομοίως, επειδή στα άρθρα A3 και A1 έχει γίνει κλικ (άρα έχουν επιλεγεί) από τον ίδιο χρήστη, τότε $\text{PathSim}_{\text{ASUSA}}(\text{A3}, \text{A1}) = \frac{2*1}{1+1} = 1$. Από την άλλη, καθώς στα άρθρα A3 και A5 δεν έχει γίνει κλικ από τον ίδιο χρήστη, τότε $\text{PathSim}_{\text{ASUSA}}(\text{A3}, \text{A5}) = \frac{2*0}{1+1} = 0$.

	ACA	ASA	ASUSA
ομοιότητα (A3-A1)	0	0	1
ομοιότητα (A3-A2)	0	0	1
ομοιότητα (A3-A4)	0	0.666	0.666
ομοιότητα (A3-A5)	1	0	0

Πίνακας 8.4: Υπολογισμένη *συνάφεια* μεταξύ του άρθρου A3 και των υπολοίπων άρθρων με τη χρήση διαφορετικών *μετα-μονοπατιών* βάσει του αλγορίθμου.

Διαφορετικά από τα μέτρα ομοιότητας *SimRank* και *RWR*, τα οποία "μεροληπτούν" ως προς τους πιο "καλά" συνδεδεμένους κόμβους στον γράφο, το *PathSim* λαμβάνει υπόψη την *ορατότητα* (*visibility*) των κόμβων, φέρνοντας πιο κοντά τους κόμβους που μοιράζονται παρόμοια *ορατότητα*. Για παράδειγμα, εάν δύο ερευνητές έχουν δημοσιεύσει 10 εργασίες ο καθένας και ένας τρίτος ερευνητής έχει δημοσιεύσει 200 εργασίες, το *SimRank* και το *RWR* θα έκαναν τους δύο προαναφερθέντες και συγκριτικώς μη έμπειρους ερευνητές να μοιάζουν πολύ με τον πιο έμπειρο τρίτο, και άρα και να μοιάζουν λιγότερο μεταξύ τους. Από την άλλη, το *PathSim* θα μείωνε την *ομοιότητά* τους με τον τρίτο και έμπειρο ερευνητή και θα τους έφερνε (τους μη έμπειρους) κοντά τον έναν στον άλλον, θεωρώντας τον μικρό αριθμό δημοσιεύσεών τους ως έναν ακόμη παράγοντα της *ομοιότητας* τους (δηλαδή, οι μη έμπειροι ερευνητές, μοιάζουν περισσότερο μεταξύ τους παρά με τους έμπειρους ερευνητές).

8.4.6 Επεξήγηση Συστάσεων βάσει μετα-μονοπατιών

Μία σημαντική λειτουργία ενός αλγορίθμου συστάσεων, εκτός από το να να παράγει τις συστάσεις, είναι να μπορεί επίσης να τις τεκμηριώνει, ώστε να γίνεται -με "διάφανο" τρόπο- αντιλη-

πτός στον χρήστη ο λόγος που οδήγησε στη σύσταση ενός στοιχείου. Λόγου χάρη, ο *αλγόριθμος PathSim* έχει τη δυνατότητα να επεξηγεί τις συστάσεις του βάσει του αριθμού των στιγμιοτύπων ενός *μετα-μονοπατιού* που οδήγησε σε μια σύσταση.

Παράδειγμα 8.7 Στο παράδειγμα του Σχήματος 8.7, για την παροχή συστάσεων βάσει παρόμοιων άρθρων, πρώτα βρίσκουμε άρθρα παρόμοια μ' εκείνο στο οποίο ο χρήστης μόλις έκανε κλικ, και στη συνέχεια τα κατατάσσουμε με βάση την *ομοιότητά* τους με το υπό εξέταση άρθρο με το οποίο αλληλεπιδρά ο χρήστης προτείνοντάς του μια λίστα με τα κορυφαία-N άρθρα. Εναλλακτικά, για την πρόβλεψη του επόμενου άρθρου (που πρόκειται να συστηθεί), μπορεί να ληφθεί υπόψη ολόκληρη η τελευταία *συνεδρία του χρήστη* (*user session*) για να προσδιοριστούν οι *βραχυχρόνιες* (*short-term*) προτιμήσεις του, ώστε να του προταθούν τα ανάλογα για τη δεδομένη χρονική στιγμή άρθρα. Κατ' αυτόν τον τρόπο, οι *ομοιότητες* μεταξύ άρθρων μπορούν να προσδιοριστούν εκτελώντας τον *PathSim* σε *μετα-μονοπάτια* ASA, AUA, ACA και ALA. Για παράδειγμα, σχετικά με το *μετα-μονοπάτι* ASA, στο Σχήμα 8.10 δείχνουμε πώς τα άρθρα κατατάσσονται σε μια λίστα με τα 5 κορυφαία προτεινόμενα άρθρα.

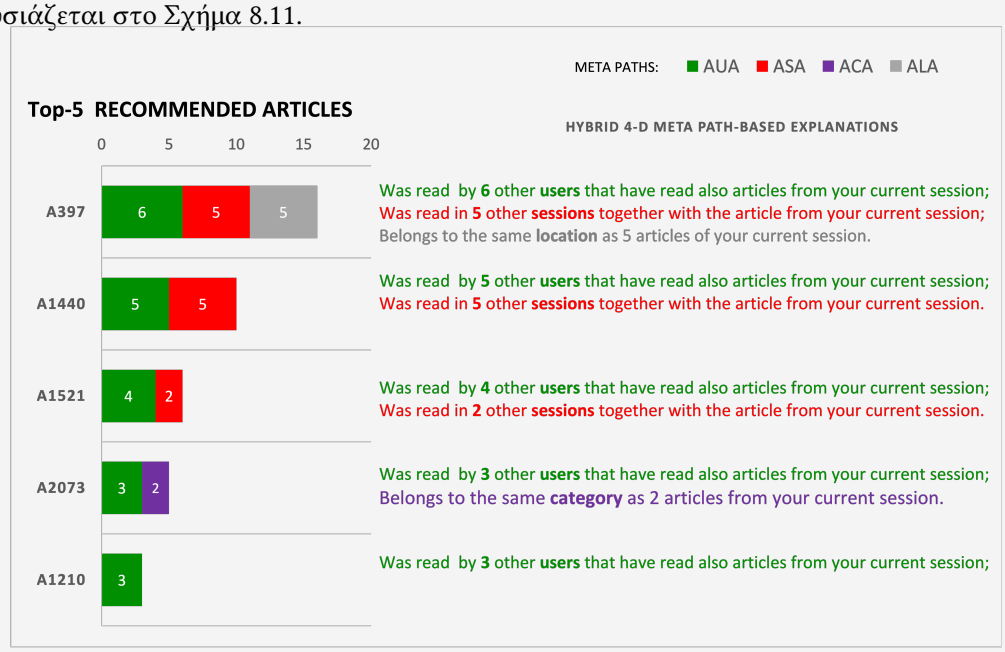
Σχήμα 8.10: Οπτική απεικόνιση συστάσεων άρθρων βάσει του *μετα-μονοπατιού* ASA τα οποία συνοδεύονται και από κατάλληλη επεξήγηση.

Όπως λοιπόν φαίνεται στο Σχήμα 8.10, παρέχουμε συστάσεις που αντικατοπτρίζουν το συνολικό ενδιαφέρον του χρήστη κατά τη διάρκεια μίας *συνεδρίας* (*user session*) μαζί με τις κατάλληλες επεξηγήσεις (δηλ. γιατί ένα άρθρο προτείνεται στον χρήστη). Αυτός είναι ένας εμπειρικός και φιλικός προς τον χρήστη τρόπος επεξήγησης της σχέσης μεταξύ του προτεινόμενου άρθρου και του άρθρου που χρησιμοποιείται για την επεξήγησή του, π.χ. *Αυτά τα δύο άρθρα έχουν διαβαστεί μαζί σε 10 διαφορετικές συνεδρίες χρηστών.*

Επιπρόσθετα, για την *παροχή συστάσεων βάσει παρόμοιων χρηστών* (*User-based CF*), αρχικά εντοπίζουμε χρήστες παρόμοιους με τον υπό εξέταση χρήστη και εν συνεχεία προτείνουμε τα πιο συχνά άρθρα μέσα στη γειτονιά των παρόμοιων με αυτόν χρηστών. Τέλος, εάν στόχος μας είναι οι *συστάσεις βάσει παρόμοιων συνεδριών* (*session-based recommendations*), τότε εντοπίζουμε πρώτα τις συνεδρίες που είναι παρόμοιες με τη συνεδρία του τρέχοντος χρήστη και έπειτα προτείνουμε τα πιο συχνά άρθρα μέσα στη γειτονιά των παρόμοιων συνεδριών.

Κατόπιν, θα δούμε πώς μπορούμε να συνδυάσουμε διάφορα *μετα-μονοπάτια* για να παρέχουμε "υβριδικές" (πολυδιάστατες) επεξηγήσεις μαζί με τις συστάσεις. Θα χρησιμοποιήσουμε λοιπόν τέσσερα *μετα-μονοπάτια* για την υποστήριξη ενός προτεινόμενου άρθρου: (AUA: όταν δύο άρθρα έχουν διαβαστεί *από τον ίδιο χρήστη*, ASA: όταν δύο άρθρα έχουν διαβαστεί *εντός της ίδιας συνεδρίας*, ACA: όταν δύο άρθρα *ανήκουν στην ίδια κατηγορία άρθρων*, και ALA: όταν δύο άρθρα *αναφέρονται στην ίδια γεωγραφική περιοχή*).

Παράδειγμα 8.8 Στο τρέχον παράδειγμά μας, συνδυάζοντας τα τέσσερα *μετα-μονοπάτια* [AUA, ASA, ACA, ALA], καταλήγουμε σε μια υβριδική 4-διάστατη επεξήγηση, η οποία παρουσιάζεται στο Σχήμα 8.11.



Σχήμα 8.11: Υβριδική 4-διάστατη επεξήγηση βάσει *μετα-μονοπατιών*.

Όπως φαίνεται εκεί, τα 5 κορυφαία προτεινόμενα άρθρα κατατάσσονται με βάση τον συνολικό αριθμό στιγμιοτύπων *μετα-μονοπατιών*. Παρατηρήστε ότι χρησιμοποιούμε διαφορετικά χρώματα στις οριζόντιες μπάρες (horizontal mixed bars) για τα διαφορετικά στυλ επεξηγήσεων που χρησιμοποιούνται (χρήστης, συνεδρία, και χαρακτηριστικό που υποδηλώνει είτε κατηγορία άρθρου είτε γεωγραφική περιοχή στην οποία αφορά το άρθρο), καθώς και το προτεινόμενο άρθρο.

8.5 Σημασιολογικοί Γράφοι Γνώσης (Knowledge Graphs)

Η *RDF (Resource Description Framework)* είναι μια γλώσσα *αναπαράστασης γνώσης* που επιτρέπει την περιγραφή *πόρων* (*resources*) διαθέσιμων στον πραγματικό κόσμο, όπως λόγου χάρη ένας αρθρογράφος, ένα άρθρο του, ένας αναγνώστης του κ.λπ. Η περιγραφή των πόρων γίνεται με τον ορισμό σχέσεων μεταξύ τους.

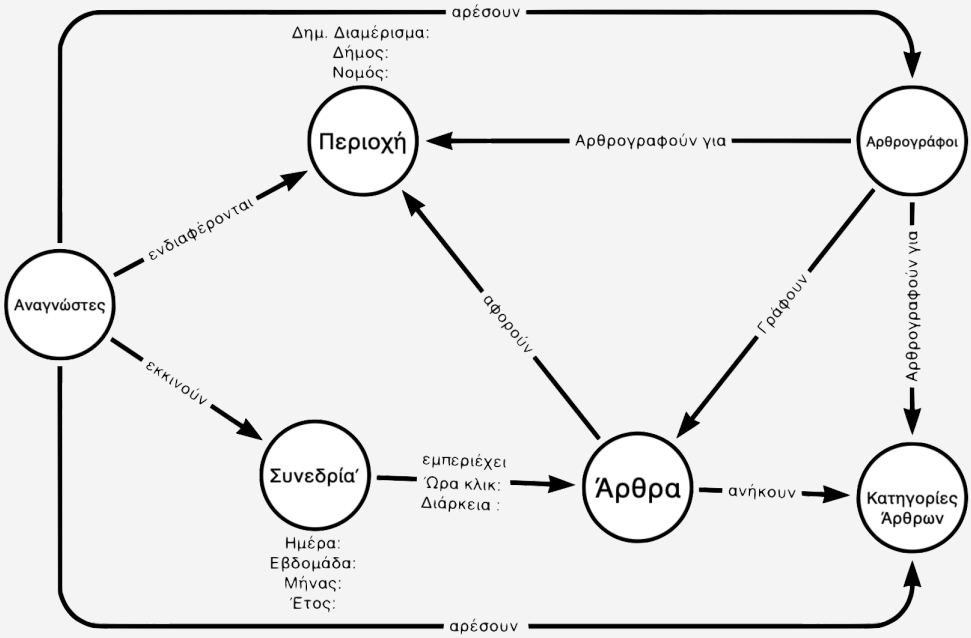
Οι "προτάσεις" (*statements*) ή αλλιώς, *τριπλέτες* (*triplets*) είναι το βασικότερο στοιχείο στην *RDF* και χρησιμοποιούνται για να δημιουργήσουμε ισχυρισμούς για έναν πόρο. Οι *τριπλέτες RDF* μπορούν να αποτελούνται από την οντότητα που περιγράφεται, τις ιδιότητες της οντότητας που περιγράφεται και μία τιμή για καθεμία από τις ιδιότητές της. Εναλλακτικά, οι προαναφερθείσες *τριπλέτες RDF* μπορούν να περιλαμβάνουν τρία διαφορετικά πράγματα: ένα "υποκείμενο", ένα "κατηγόρημα" και ένα "αντικείμενο". Τονίζεται ότι το "υποκείμενο" και το "αντικείμενο" είναι *πόροι*, ενώ το "κατηγόρημα" είναι η *προσδιορισμένη μεταξύ τους σχέση* (*relationship*).

Ένας *γράφος γνώσης*, λοιπόν, μας βοηθά να αναπαραστήσουμε μία *τριπλέτα RDF* με *γράφημα* και μέσω *κόμβων (nodes)* και *κατευθυνόμενων ακμών (directed edges)* οι οποίες περιέχουν *ετικέτες* (*labels*) για να προσδιορίσουν τη σχέση που συνδέει δύο κόμβους. Οι ακμές σε μία *τριπλέτα RDF* έχουν πάντα κατεύθυνση από το "υποκείμενο" στο "αντικείμενο", ενώ η *ετικέτα της ακμής* δηλώνει το "κατηγόρημα". Ένα παράδειγμα ενός γράφου γνώσης για μια online εφημερίδα απεικονίζεται στο Σχήμα 8.12.

Παράδειγμα 8.9 Στο Σχήμα 8.12 εμφανίζεται ο *Γράφος Γνώσης (Knowledge Graph)* για μια online εφημερίδα που συνίσταται από αρθρογράφους, άρθρα, κατηγορίες άρθρων, περιοχές και αναγνώστες που αλληλεπιδρούν με τα άρθρα κατά τη διάρκεια της επίσκεψής τους στην ηλεκτρονική εφημερίδα. Αναλυτικότερα, οι αρθρογράφοι γράφουν για συγκεκριμένες γεωγραφικές περιοχές (π.χ. Θεσσαλονίκη, Αθήνα κ.λπ.). Η δε γεωγραφική διάσταση έχει επιμέρους επίπεδα ανάλυσης: Δημοτικό Διαμέρισμα, Δήμος, Νομός. Επίσης, η διάσταση του χρόνου χωρίζεται σε διαφορετικά επίπεδα ανάλυσης: Ημέρα, Εβδομάδα, Μήνας, Έτος. Τέλος, οι αναγνώστες δηλώνουν ρητά ότι τους αρέσουν συγκεκριμένα άρθρα, αρθρογράφοι και κατηγορίες άρθρων. Έστω, επομένως, ότι ο ιδιοκτήτης της online εφημερίδας θέλει να αναπτυχθεί για λογαριασμό του ένας *γράφος γνώσης* που να συσσωρεύει γνώση για τις σχέσεις μεταξύ αναγνωστών και άρθρων, καθώς και για άλλες "οντότητες" (αρθρογράφοι, κατηγορίες άρθρων, περιοχές κ.λπ.) με τις οποίες αυτοί αλληλεπιδρούν, για να προτείνει εξατομικευμένα άρθρα στους αναγνώστες βάσει της προηγούμενης χρονικά αλληλεπίδρασής τους με τα άρθρα και των λοιπών συσχετίσεων του γράφου γνώσης. Όπως ήδη αναφέραμε, η γενική δομή των *τριπλετών RDF* είναι η ακόλουθη:

Statement (resource, property relationship, resource)

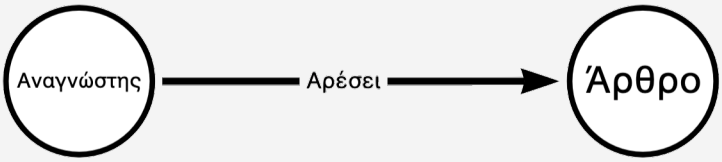
Πρόταση (πόρος, ιδιότητα σχέσης, πόρος)



Σχήμα 8.12: Παράδειγμα μιας "οντολογίας" (*RDF scheme*) που αφορά σε μια online εφημερίδα.

Να σημειωθεί εδώ ότι τη δομή των *τριπλετών RDF* θα τη δείτε επίσης στη βιβλιογραφία, αλλά και με τη μορφή που εμφανίζεται στο Σχήμα 8.13. Τέλος, ένα παράδειγμα μιας *τριπλέτας RDF* για την παραπάνω περίπτωση της online εφημερίδας εμφανίζεται στο Σχήμα 8.14.

Σχήμα 8.13: Γράφος μιας *πρότασης* ή *τριπλέτας RDF*.



Σχήμα 8.14: Παράδειγμα *πρότασης* ή *τριπλέτας RDF* από την online εφημερίδα.

Συνεπώς, βάσει του *γράφου γνώσης* του Σχήματος 8.12 μπορούμε να προτείνουμε άρθρα των ίδιων κατηγοριών με αυτά που έχει αλληλεπιδράσει ο υπό εξέταση χρήστης (εδώ, αναγνώστης) σε προηγούμενες επισκέψεις του στη διαδικτυακή εφημερίδα. Έστω, επομένως, ότι ο

υπό εξέταση αναγνώστης αλληλεπίδρασε στο παρελθόν με άρθρα των κατηγοριών "Politics" και "Sports". Σ' αυτήν την περίπτωση, το σύστημα συστάσεων που βασίζεται στον *γράφο γνώσης* του Σχήματος 8.12 βρίσκει τις προηγούμενες *επισκέψεις* (*συνεδρίες*) του αναγνώστη στη διαδικτυακή εφημερίδα καθώς και το ποια άρθρα αυτός έχει διαβάσει σ' αυτές. Στη συνέχεια, ανακτά την κατηγορία στην οποία ανήκουν αυτά τα άρθρα, και του προτείνει άλλα άρθρα που ανήκουν στις ίδιες κατηγορίες, και με τα οποία ο αναγνώστης δεν έχει αλληλεπιδράσει στο παρελθόν.

Ένα δεύτερο παράδειγμα σύστασης που βασίζεται στον *γράφο γνώσης* του Σχήματος 8.12 είναι το ακόλουθο: προτείνουμε στον υπό εξέταση αναγνώστη/χρήστη αρθρογράφους που έχουν γράψει κάποιο άρθρο για μια γεωγραφική περιοχή που τον ενδιαφέρει. Έστω λοιπόν ότι ο υπό εξέταση αναγνώστης έχει αλληλεπιδράσει με άρθρα που αφορούν στη γεωγραφική περιοχή της πόλης του Καρλοβασίου. Εάν, τώρα, παρατηρήσουμε την ιεραρχία της "οντότητας" *Περιοχή* του Σχήματος 8.12, τότε θα δούμε ότι αυτό αναλύεται σε 3 επίπεδα (Δημ. Διαμέρισμα, Δήμος, Νομός). Συνεπώς, καθώς το σύστημα συστάσεων λαμβάνει υπόψη ένα υψηλότερο (ευρύτερο) επίπεδο αφαίρεσης (π.χ. τον Νομό), μπορεί να εξαγάγει το συμπέρασμα ότι ο υπό εξέταση αναγνώστης θα ενδιαφέρεται για εκείνους τους αρθρογράφους της εφημερίδας που αναφέρονται συνολικά στον Νομού Σάμου. Επομένως, το σύστημα συστάσεων θα προτείνει στον υπό εξέταση αναγνώστη μια ανάλογη λίστα αφαιρώντας ενδεχομένως εκείνους με άρθρα των οποίων αυτός έχει ήδη αλληλεπιδράσει στο παρελθόν.

8.6 Ερωτήσεις Κεφαλαίου

1. Αναφέρατε τοπικά μέτρα *ομοιότητας* μεταξύ των κόμβων ενός γράφου. Ορίστε για το καθένα τον τρόπο υπολογισμού του.

2. Περιγράψτε συνοπτικά τον *αλγόριθμο Katz Status index* για τη μέτρηση *ομοιότητας* μεταξύ των κόμβων ενός γράφου.

3. Περιγράψτε συνοπτικά τον *αλγόριθμο SimRank* για τη μέτρηση *ομοιότητας* μεταξύ των κόμβων ενός γράφου.

4. Περιγράψτε συνοπτικά τον *αλγόριθμο Random Walk with Restart* για τη μέτρηση *ομοιότητας* μεταξύ των κόμβων ενός γράφου.

5. Περιγράψτε συνοπτικά τον *αλγόριθμο PathSim* για τη μέτρηση *ομοιότητας* μεταξύ των κόμβων ενός γράφου.

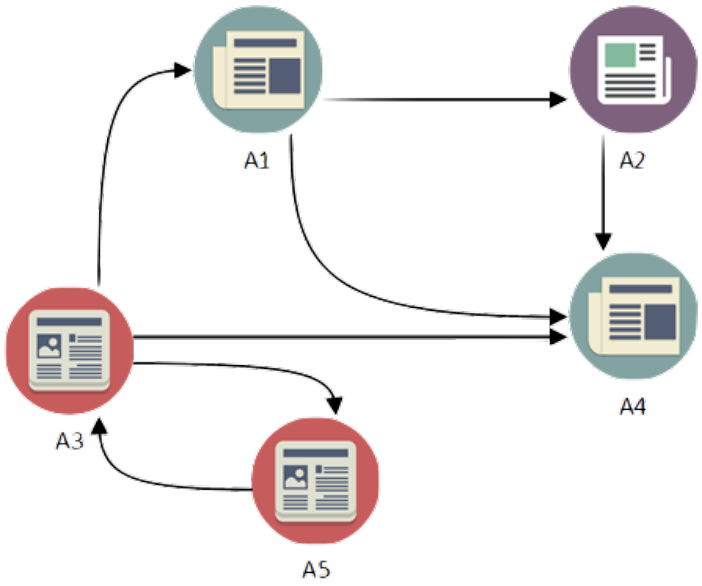
6. Εφαρμόστε στον γράφο του Σχήματος 8.15 τον ακόλουθο απλό τύπο *PageRank* χωρίς "τηλεμεταφορά":

$$\mathbf{v}' = \mathbf{M} \cdot \mathbf{v}$$

όπου $\mathbf{M}$ ο πίνακας πιθανοτήτων μετάβασης από έναν κόμβο στους άλλους και $\mathbf{v}$ διάνυσμα που εκφράζει τη σημαντικότητα του κάθε κόμβου στον γράφο βάσει της προσβασιμότητάς του. Το σκεπτικό πίσω από τον *PageRank*, θυμίζουμε, βασίζεται στο γεγονός ότι ένας τυχαίος "περιπατητής" ξεκινά από τυχαίο κόμβο και κάθε φορά "μεταβαίνει" τυχαία σ' έναν διπλανό κόμβο μέχρι να ολοκληρώσει το "ταξίδι" του. Στο τέλος αυτού του "ταξιδιού" το ζητούμενο είναι να προσδιορίσουμε ποιοι είναι οι κόμβοι που επισκέφθηκε πιο συχνά.

(α) Αρχικοποιήστε κάθε τιμή του διανύσματος αρχικών πιθανοτήτων ως ακολούθως: $\mathbf{v_0} = \mathbf{e}/n$, όπου n ο αριθμός των κόμβων του γράφου, και $\mathbf{e}$ ένα διάνυσμα μήκους n με όλα τα στοιχεία του να ισούνται με 1 (μονάδα). Ποιες θα είναι οι τιμές του αρχικού διανύσματος $\mathbf{v_0}$;

(β) Υπολογίστε τον *πίνακα πιθανοτήτων μετάβασης* $\mathbf{M}$ από έναν κόμβο στους άλλους.



Σχήμα 8.15: Ένας *μονομερής γράφος* με ειδησεογραφικά άρθρα.

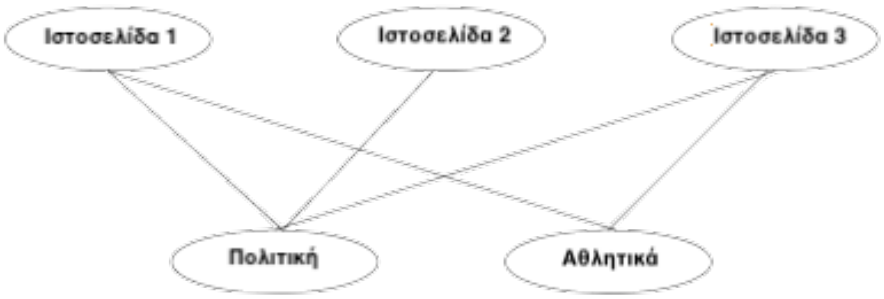
(γ) Εφαρμόστε τον παραπάνω απλό τύπο του *PageRank* και εκτελέστε τον κατ' επανάληψη για τρία βήματα ή μέχρι να συγκλίνει το διάνυσμα $\mathbf{v}'$ (με 1e-4 παράγοντα σύγκλισης). Ποιες είναι οι τελικές τιμές του διανύσματος *PageRank* $\mathbf{v}'$;

7. Εφαρμόστε τον αλγόριθμο *PageRank* με "τηλεμεταφορά" στον γράφο του Σχήματος 8.15, όπου σε κάθε βήμα του αλγορίθμου ο "περιπατητής" θα ακολουθεί ένα "τυχαίο μονοπάτι" με πιθανότητα β, ενώ με πιθανότητα $(1 - \beta)$ θα μπορεί να μεταπηδήσει σ' έναν τυχαίο κόμβο του γράφου, χρησιμοποιώντας την ακόλουθη Εξίσωση:

$$\mathbf{v}' = \beta \times \mathbf{M} \times \mathbf{v} + (1 - \beta) \times \mathbf{e}/n$$

όπου n ο αριθμός των κόμβων του δικτύου και $\mathbf{e}$ διάνυσμα μήκους n μ' όλα τα στοιχεία του να ισούνται με 1 (μονάδα). Εφαρμόστε τον τύπο του *PageRank* με "τηλεμεταφορά" για τρεις επαναλήψεις. Πιο αποτέλεσμα παίρνετε; Χρησιμοποιήστε $\beta = 0.8$.

8. Δίνεται ο γράφος του παρακάτω Σχήματος:



Για την εύρεση *ομοιοτήτων* μεταξύ των κόμβων στον παραπάνω γράφο, είναι αναγκαίο να βρούμε τις *ομοιότητες* μεταξύ του υπό εξέταση i και όλων των υπόλοιπων κόμβων. Έστω, λοιπόν, ότι βάζουμε έναν "τυχαίο περιπατητή" να κάνει κάποιους "τυχαίους περιπάτους" γύρω από τον κόμβο i εφαρμόζοντας μια λογική παρόμοια με του *PageRank* με "τηλεμεταφορά". Πιο συγκεκριμένα, ο "περιπατητής" συνεχίζει τον "περίπατο" με πιθανότητα β και μεταπηδά πάντα στον ίδιο *αρχικό* υπό εξέταση κόμβο i με πιθανότητα $(1 - \beta)$. Έτσι, ο τύπος για τον υπολογισμό του διανύσματος *συνάφειας* των κόμβων με τον υπό εξέταση κόμβο i είναι ο παρακάτω:

$$\mathbf{v}' = \beta \times \mathbf{M} \times \mathbf{v} + (1 - \beta) \times \mathbf{e_i}$$

όπου $\mathbf{e_i}$ διάνυσμα μήκους ίσου με τον αριθμό των κόμβων του γράφου όπου όλα τα στοιχεία του ισούνται με 0 (μηδέν), εκτός από τον κόμβο-στόχο i του οποίου η τιμή ισούται με 1 (μονάδα). Μ' αυτόν τον τρόπο υπολογίζουμε τα διανύσματα *συνάφειας* για κάθε κόμβο του δικτύου έτσι ώστε στο τέλος να έχουμε έναν (nxn) τετραγωνικό πίνακα ως ακολούθως:

$$\mathbf{V}' = \beta \times \mathbf{M} \times \mathbf{V} + (1 - \beta) \times \mathbf{I_n}$$

όπου $\mathbf{V}$ (*nxn*) πίνακας αρχικοποιημένος από τον *ταυτοτικό πίνακα* $\mathbf{I_n}$.

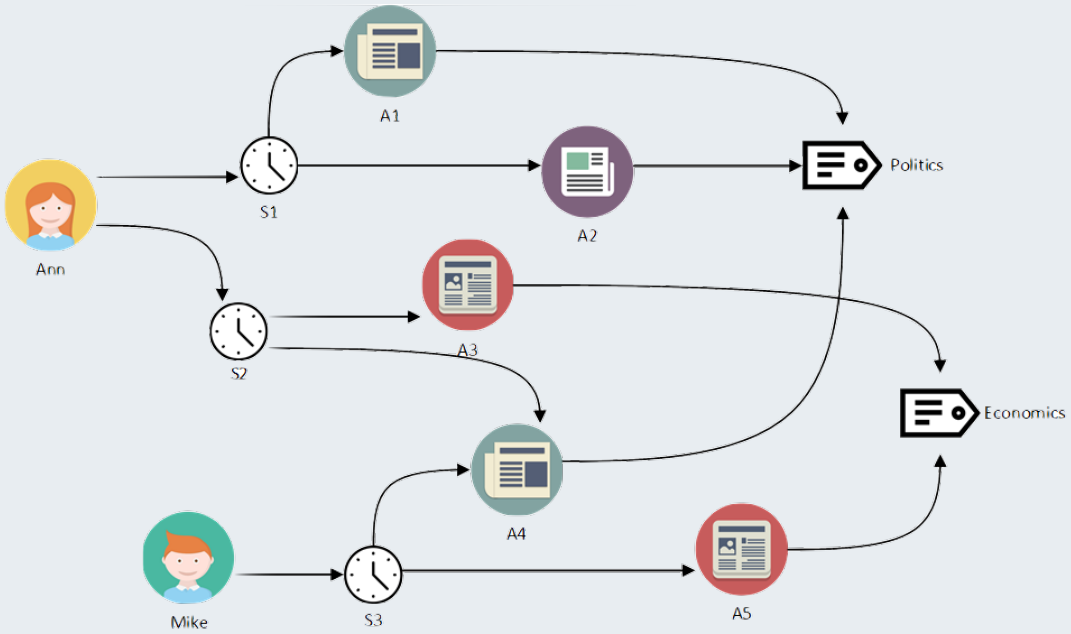
(α) Υπολογίστε τον πίνακα πιθανοτήτων μετάβασης $\mathbf{M}$.

(β) Εφαρμόστε έναν *τυχαίο περίπατο με επανεκκίνηση* (*Random walk with restart*) χρησιμοποιώντας $\beta = 0.8$ και $\epsilon = 1e - 4$. Ποιες είναι οι τελικές τιμές του πίνακα $\mathbf{V}'$ μετά από τρεις επαναλήψεις; Μπορούμε να τον ονομάσουμε *πίνακα ομοιοτήτων*;

8.7 Ασκήσεις Προγραμματισμού

8.7.1 Σύσταση Άρθρων σε ένα Ετερογενές Δίκτυο Πληροφοριών

Δίνεται ο παρακάτω γράφος που αποτελείται από 4 διαφορετικούς τύπους κόμβων (χρήστες, συνεδρίες, άρθρα, κατηγορίες άρθρων):



Ζητείται:

1. Να δημιουργήσετε έναν *ετερογενές γράφο* με τους τέσσερεις διαφορετικούς τύπους κόμβων και να τον παρουσιάσετε οπτικά.

2. Να γίνει εκτέλεση του *αλγορίθμου RWR* (*τυχαίος περίπατος με επανεκκίνηση*) σ' ολόκληρο το *ετερογενές δίκτυο*.

3. Να αναλύσετε τους υποπίνακες AA (*Article-Article*), UA (*User-Article*) και UC (*User-Category*) απαντώντας στις ακόλουθες ερωτήσεις:

   (α) Για τον υποπίνακα AA:

   - Ποιο είναι το πιο *συναφές* άρθρο στο άρθρο A1; Γιατί;
   - Ποιο άρθρο μοιάζει περισσότερο με το άρθρο A2: το A3 ή το A4; Γιατί;

   (β) Για τον υποπίνακα UA:

   - Εάν θα θέλαμε να ξαναπροτείνουμε στην Ann κάποιο άρθρο το οποίο ενδεχομένως έχει ήδη διαβάσει, ποιο θα ήταν αυτό; Επίσης, ποιο άρθρο θα ξαναπροτείναμε στον Mike;

- Γιατί τα άρθρα A1 και A2 είναι πιο πιθανό να ξαναδιαβαστούν από την Ann από ό,τι το άρθρο A3;

(γ) Για τον υποπίνακα UC :

- Γιατί η κατηγορία Οικονομικά ενδιαφέρει περισσότερο τον Mike από ό,τι η κατηγορία Πολιτική;

Λύσεις:

1. Ακολουθεί ο πίνακας αποτελεσμάτων μετά την εκτέλεση του αλγορίθμου *τυχαίου περιπάτου με επανεκκίνηση* (Random Walk with Restart):

	Ann	Mike	S1	S2	S3	A1	A2	A3	A4	A5	Politics	Economics
Ann	0.2838	0.0088	0.1629	0.1529	0.0336	0.0610	0.0610	0.0521	0.0670	0.0203	0.0673	0.0294
Mike	0.0175	0.2763	0.0199	0.0473	0.2869	0.0154	0.0154	0.0332	0.0990	0.0971	0.0393	0.0526
S1	0.1086	0.0066	0.3389	0.0668	0.0241	0.1201	0.1201	0.0239	0.0542	0.0125	0.1099	0.0141
S2	0.1019	0.0158	0.0668	0.3137	0.0584	0.0334	0.0334	0.1074	0.1150	0.0393	0.0567	0.0582
S3	0.0224	0.0956	0.0241	0.0584	0.3578	0.0198	0.0198	0.0421	0.1245	0.1219	0.0484	0.0652
A1	0.0610	0.0077	0.1802	0.0501	0.0296	0.2939	0.0939	0.0180	0.0669	0.0125	0.1736	0.0126
A2	0.0610	0.0077	0.1802	0.0501	0.0296	0.0939	0.2939	0.0180	0.0669	0.0125	0.1736	0.0126
A3	0.0521	0.0166	0.0358	0.1611	0.0631	0.0180	0.0180	0.3037	0.0680	0.0776	0.0331	0.1529
A4	0.0447	0.0330	0.0542	0.1150	0.1245	0.0446	0.0446	0.0453	0.2938	0.0479	0.1147	0.0377
A5	0.0203	0.0485	0.0188	0.0589	0.1829	0.0125	0.0125	0.0776	0.0718	0.3106	0.0298	0.1557
Politics	0.0449	0.0131	0.1099	0.0567	0.0484	0.1158	0.1158	0.0221	0.1147	0.0199	0.3225	0.0164
Economics	0.0294	0.0263	0.0212	0.0874	0.0977	0.0126	0.0126	0.1529	0.0566	0.1557	0.0245	0.3230

2. Ακολουθεί λεπτομερής σχολιασμός των αποτελεσμάτων για τους υποπίνακες AA, UA και UC:

(α) Για τον υποπίνακα AA:

- Το άρθρο A2 είναι το πιο συναφές στο A1, επειδή έχουν γίνει κλικ σ' αυτό από τον ίδιο χρήστη κατά την ίδια συνεδρία και ανήκουν στην ίδια κατηγορία άρθρων.
- Το άρθρο A4 μοιάζει περισσότερο με το A2, επειδή έχουν γίνει κλικ σ' αυτό από τον ίδιο χρήστη και ανήκουν στην ίδια κατηγορία άρθρων, ενώ στα A3 και A2 έχουν γίνει μόνο κλικ από τον ίδιο χρήστη.

(β) Για τον υποπίνακα UA:

- Θα προτείναμε τόσο στην Ann όσο και στον Mike να ξαναδιαβάσουν το άρθρο A4, επειδή είναι το πιο δημοφιλές άρθρο στον γράφο.
- Μπορούμε να παρατηρήσουμε ότι η Ann δείχνει μεγαλύτερο ενδιαφέρον για τα πολιτικά άρθρα από ό,τι για τα οικονομικά (3 από τα 4 άρθρα που έχει διαβάσει ανήκουν στην Πολιτική). Γι' αυτό και η *συνάφεια* των A1 και A2 είναι μεγαλύτερη απ' αυτήν του A3.

(γ) Για τον υποπίνακα UC:

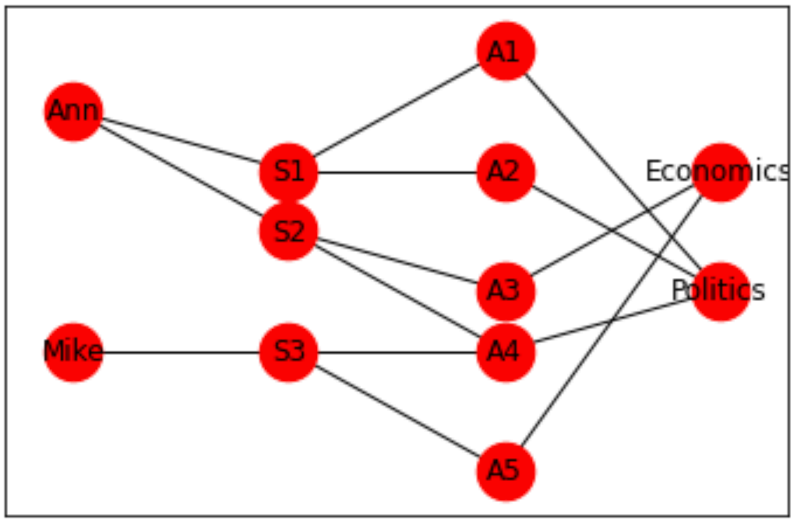
- Αν και ο Mike έχει διαβάσει ένα άρθρο Πολιτικής και ένα Οικονομίας, μπορούμε να παρατηρήσουμε ότι οι βαθμολογίες τους δεν είναι ίδιες, επειδή το άρθρο A4 που αφορά στην Πολιτική -όπως μόλις προαναφέραμε- εμφανίστηκε σε μία συνεδρία με το άρθρο A3 που ανήκει στην Οικονομία και που διάβασε η Ann, γεγονός που φέρνει τις δύο κατηγορίες πιο κοντά τη μία στην άλλη. Ως εκ τούτου, έρχεται το άρθρο A4 λίγο πιο κοντά στα Οικονομικά, γεγονός που με τη σειρά του σημαίνει ότι ο Mike προτιμά περισσότερο τα άρθρα Οικονομικών από αυτά της Πολιτικής.

Ακολουθεί ο κώδικας *Python* που υλοποιεί την προγραμματιστική άσκηση:

```
#Import Libraries
import networkx as nx
import pandas as pd
import numpy as np
from collections import defaultdict
import copy

#Δημιουργία του ετερογενούς γράφου
G = nx.Graph() # undirected  graph
G.add_nodes_from(['Ann','Mike'], entity='U')
G.add_nodes_from(['S1','S2','S3'], entity='S')
G.add_nodes_from(['A1','A2','A3','A4','A5'], entity='A')
G.add_nodes_from(['Politics','Economics'], entity='C')

G.add_edges_from([('Ann','S1'), ('Ann','S2'), ('Mike','S3')])

G.add_edges_from([('S1','A1'), ('S1','A2'),
                  ('S2','A3'), ('S2','A4'),
                  ('S3','A4'), ('S3','A5')])

G.add_edges_from([('A1','Politics'), ('A2','Politics'), ('A4','
    Politics'), ('A3','Economics'), ('A5','Economics')])

positions = {'Ann':[0,5],'Mike':[0,1],
             'S1':[2,4],'S2':[2,3],'S3':[2,1],
```

```
                'A1':[4,6], 'A2':[4,4], 'A3':[4,2], 'A4':[4,1], 'A5'
    :[4,-1],
                'Politics':[6,2], 'Economics':[6,4]}

nx.draw_networkx(G, node_size=600, with_labels=True, node_color=
    'red',pos=positions)
```

Ακολουθεί η οπτική αναπαράσταση του γράφου μας:




```
# Δημιουργία του πίνακα γειτνίασης A
A = nx.adjacency_matrix(G)
A.todense()
matrix([[0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0],
        [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0],
        [1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0],
        [1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0],
        [0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0],
        [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0],
        [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0],
        [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1],
        [0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0],
        [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1],
        [0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0],
        [0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0]])

```

```
# Δημιουργία του πίνακα M πιθανοτήτων μετάβασης
A.todense()
# Υπολογισμός του πίνακα βαθμών των κόμβων του γράφου
# d = np.array(list(G.degree().values()))
d = np.array([len(nbrs) for n,nbrs in G.adj.items()])

# Υπολογισμός του διανύσματος μετάβασης από έναν κόμβο του γράφου σ' όλους τους προσκείμενους κόμβους σ' αυτόν
p = 1./d

M = A.multiply(p)
d
#array([2, 1, 3, 3, 3, 2, 2, 2, 3, 2, 3, 2])

#Αρχικοποίηση των παραμέτρων του αλγορίθμου RWR
beta = 0.8
max_iter = 100
eps = 1e-4

n = nx.number_of_nodes(G)
I = np.identity(n)

# Υλοποίηση του αλγορίθμου RWR
V = I
for i in range(max_iter):
    V_new = beta*M*V + (1-beta)*I

    if (abs(V-V_new) < eps).all():
        print("Converged after %d iteration" % (i))
        break

    V = V_new
V.round(4)

#Σύγκλιση τιμών αλγορίθμου μετά από 34 επαναλήψεις
array([[0.2838, 0.0178, 0.1083, 0.1016, 0.0222, 0.0612, 0.0612,
```

```
    0.0524, 0.045 , 0.0206, 0.0446, 0.0291],
52      [0.0089, 0.2764, 0.0065, 0.0156, 0.0955, 0.0078, 0.0078,
    0.0167, 0.0331, 0.0487, 0.013 , 0.0261],
53      [0.1625, 0.0196, 0.3393, 0.0672, 0.0245, 0.1797, 0.1797,
    0.0355, 0.0538, 0.0184, 0.1103, 0.0216],
54      [0.1525, 0.0469, 0.0672, 0.3141, 0.0588, 0.0497, 0.0497,
    0.1607, 0.1146, 0.0585, 0.0571, 0.0878],
55      [0.0332, 0.2865, 0.0245, 0.0588, 0.3583, 0.0292, 0.0292,
    0.0627,0.1241, 0.1825, 0.0488, 0.0981],
56      [0.0612, 0.0157, 0.1198, 0.0331, 0.0195, 0.2942, 0.0942,
    0.0182, 0.0449, 0.0128, 0.1155, 0.0124],
57      [0.0612, 0.0157, 0.1198, 0.0331, 0.0195, 0.0942, 0.2942,
    0.0182, 0.0449, 0.0128, 0.1155, 0.0124],
58      [0.0524, 0.0335, 0.0236, 0.1071, 0.0418, 0.0182, 0.0182,
    0.304 , 0.0456, 0.0778, 0.0218, 0.1527],
59      [0.0674, 0.0994, 0.0538, 0.1146, 0.1241, 0.0673, 0.0673,
    0.0684, 0.2942, 0.0722, 0.1143, 0.0561],
60      [0.0206, 0.0974, 0.0123, 0.039 , 0.1216, 0.0128, 0.0128,
    0.0778, 0.0481, 0.3109, 0.0196, 0.1554],
61      [0.0669, 0.0389, 0.1103, 0.0571, 0.0488, 0.1732, 0.1732,
    0.0327, 0.1143, 0.0294, 0.3229, 0.0249],
62      [0.0291, 0.0523, 0.0144, 0.0585, 0.0654, 0.0124, 0.0124,
    0.1527, 0.0374, 0.1554, 0.0166, 0.3233]])
```

Σημείωση: Για ευκολότερη και καλύτερη ανάλυση του παραπάνω πίνακα αποτελεσμάτων μπορούμε να τον μετατρέψουμε σε μορφή *πλαισίου δεδομένων* (*data frame*) χρησιμοποιώντας τον ακόλουθο κώδικα *Python*:

```
#Αντιστοίχιση ενός κόμβου με τον αύξοντα αριθμό του
n_dict = dict(zip(G.nodes(), range(n)))
#Δημιουργήστε ένα λεξικό στη βάση του πίνακα με ανάθεση των ονομάτων των
    κόμβων στα κλειδιά
result = defaultdict(list)
for n1 in G.nodes():
  result[n1] = defaultdict(int)
  for n2 in G.nodes():
      result[n1][n2] = V[n_dict[n2]][n_dict[n1]]
```

```
#Αποθήκευση του αποτελέσματος στο αντικείμενο data frame
result_df = pd.DataFrame(result).round(4)
# Καθορισμός των στηλών και των γραμμών στο πλαίσιο δεδομένων
order = ['Ann','Mike','S1','S2','S3','A1','A2','A3','A4','A5','
    Politics','Economics']
result_df = result_df[order]
result_df.reindex(order)
```

	Ann	Mike	S1	S2	S3	A1	A2	A3	A4	A5	Politics	Economics
Ann	0.2838	0.0088	0.1629	0.1529	0.0336	0.0610	0.0610	0.0521	0.0670	0.0203	0.0673	0.0294
Mike	0.0175	0.2763	0.0199	0.0473	0.2869	0.0154	0.0154	0.0332	0.0990	0.0971	0.0393	0.0526
S1	0.1086	0.0066	0.3389	0.0668	0.0241	0.1201	0.1201	0.0239	0.0542	0.0125	0.1099	0.0141
S2	0.1019	0.0158	0.0668	0.3137	0.0584	0.0334	0.0334	0.1074	0.1150	0.0393	0.0567	0.0582
S3	0.0224	0.0956	0.0241	0.0584	0.3578	0.0198	0.0198	0.0421	0.1245	0.1219	0.0484	0.0652
A1	0.0610	0.0077	0.1802	0.0501	0.0296	0.2939	0.0939	0.0180	0.0669	0.0125	0.1736	0.0126
A2	0.0610	0.0077	0.1802	0.0501	0.0296	0.0939	0.2939	0.0180	0.0669	0.0125	0.1736	0.0126
A3	0.0521	0.0166	0.0358	0.1611	0.0631	0.0180	0.0180	0.3037	0.0680	0.0776	0.0331	0.1529
A4	0.0447	0.0330	0.0542	0.1150	0.1245	0.0446	0.0446	0.0453	0.2938	0.0479	0.1147	0.0377
A5	0.0203	0.0485	0.0188	0.0589	0.1829	0.0125	0.0125	0.0776	0.0718	0.3106	0.0298	0.1557
Politics	0.0449	0.0131	0.1099	0.0567	0.0484	0.1158	0.1158	0.0221	0.1147	0.0199	0.3225	0.0164
Economics	0.0294	0.0263	0.0212	0.0874	0.0977	0.0126	0.0126	0.1529	0.0566	0.1557	0.0245	0.3230

Βιβλιογραφία

L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27(3):187--203, 2005.

A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590--614, 2002.

J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI)*, pages 201--210, Boston, MA, 2009.

G. G. Chowdhury. *Introduction to modern information retrieval.* Facet publishing, 2010.

K. C. Foster and S. Q. Muth. A Faster Katz Status Score Algorithm. *Computational & Mathematical Organization Theory*, 7(4):275--285, 2002.

M. Fredman and R. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, 34:596--615, 1987.

I. Guy, I. Ronen, and E. Wilcox. Do you know?: recommending people to invite into your social network. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI)*, pages 77--86, Sanibel Island, FL, 2009.

T. Haveliwala. Topic sensitive page rank. In *World Wide Web Conference(WWW)*, Honolulu, Hawaii, USA, 2002.

G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2002)*, pages 538--543, Edmonton, Canada, 2002.

L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39--43, Mar. 1953. ISSN 0033-3123. 10.1007/BF02289026.

J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of massive data sets*. Cambridge university press, 2019.

D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM)*, pages 556--559, New Orleans, LO, 2003.

M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2), 2001.

L. Page and S. Brin. The pagerank citation ranking: Bringing order to the web. In *World Wide Web Conference(WWW)*, 1998.

J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 653--658, Seattle, WA, 2004.

A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. Friendlink: Link prediction in social networks via bounded local path traversal. In *Proceedings of the 3rd Conference on Computational Aspects of Social Networks (CASON)*, pages 66--71, Salamanca, Spain, 2011.

Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992--1003, 2011.

H. Tong, C. Faloutsos, and J. Pan. Fast random walk with restart and its applications. In *ICDM '06: Proceedings of the 6th International Conference on Data Mining*, pages 613--622. IEEE Computer Society, 2006. ISBN 0-7695-2701-9. http://dx.doi.org/10.1109/ICDM.2006.70.

S. Yizhou, H. Jiawei, Y. Xifeng, Y. Philip S., and W. Tianyi. Pathsim: Meta path-based top-k similarity search in heterogenuos information networks. In *Proceedings of the VLDB Endowment (VLDB'2011)*, Seattle, Washigton, 2011.