3 Πώς κατατάσσει η Google τις ιστοσελίδες;

3.1 Μία Σύντομη Απάντηση

Τώρα στρεφόμαστε προς εκείνους τους συνδέσμους που βλέπουμε σε μία ιστοσελίδα αποτελεσμάτων αναζήτησης. Όχι τα αποτελέσματα από διαφημίσεις ή από χορηγούμενες αναζητήσεις αλλά από την πραγματική κατάταξη των ιστοσελίδων σε μηχανές αναζήτησης όπως η Google. Θα δούμε ότι, κάθε φορά που κάνετε αναζήτηση στο www.google.com, η Google λύνει μία πολύ μεγάλη γραμμική εξίσωση για την κατάταξη των ιστοσελίδων.

Η ιδέα της ενσωμάτωσης των συνδέσμων στο κείμενο χρονολογείται από τα μέσα του περασμένου αιώνα. Καθώς το Διαδίκτυο κλιμακώθηκε και με την εισαγωγή του παγκόσμιου ιστού το 1989, των προγραμμάτων περιήγησης το 1990 και των διαδικτυακών πυλών το 1994, αυτό το όραμα υλοποιήθηκε σε μία πρωτοφανή κλίμακα. Το δίκτυο των ιστοσελίδων είναι τεράστιο: κάπου μεταξύ των 40 και 60 δισεκατομμυρίων το 2008 σύμφωνα με διάφορες εκτιμήσεις. Και οι περισσότερες από αυτές είναι συνδεδεμένες μεταξύ τους σε μία γιγαντιαία συνιστώσα αυτού του δικτύου. Είναι επίσης αραιό: οι περισσότερες ιστοσελίδες έχουν μόνο μερικούς υπερσυνδέσμους που δείχνουν προς το εσωτερικό ή προς τα έξω. Η μηχανή αναζήτησης της Google οργανώνει αυτό το τεράστιο και πυκνό δίκτυο με το να κατατάσσει τις ιστοσελίδες.

Οι πιο σημαντικές ιστοσελίδες πρέπει να κατατάσσονται υψηλότερα. Αλλά πώς μπορείτε να ποσοτικοποιήσετε το *πόσο* σημαντική είναι μία ιστοσελίδα; Λοιπόν αν υπάρχουν πολλές άλλες σημαντικές ιστοσελίδες που δείχνουν προς μία ιστοσελίδα Α, πιθανότατα και η Α είναι σημαντική. Αυτό το επιχείρημα υποθέτει σιωπηρά δύο ιδέες:

- Οι ιστοσελίδες σχηματίζουν ένα δίκτυο, όπου μία ιστοσελίδα είναι ένας κόμβος και μία υπερσύνδεση είναι μία *κατευθυνόμενη* ζεύξη στο δίκτυο: Μία ιστοσελίδα Α μπορεί να δείχνει τη Β χωρίς η Β να δείχνει πίσω την Α.
- Μπορούμε να μετατρέψουμε τη φαινομενικά κυκλική λογική του «σημαντικές ιστοσελίδες που δείχνουν εσάς σημαίνει ότι και εσείς είστε σημαντικός» σε ένα σύνολο εξισώσεων που χαρακτηρίζουν την *ισορροπία* (μία ισορροπία σταθερού σημείου, όχι μία ισορροπία Nash βασισμένη στη θεωρία παιγνίων) βασιζομένοι σε έναν *αναδρομικό ορισμό* της «σημαντικότητας». Αυτό το μέτρο σημαντικότητας θα λειτουργήσει μετά ως μία προ-

σέγγιση της τελικής δοκιμής των μηχανών αναζήτησης: πόσο χρήσιμα βρίσκει ένας χρήστης τα αποτελέσματα μίας αναζήτησης.

Όπως αναφέρθηκε στο Κεφάλαιο 1, ένα δίκτυο αποτελείται και από *τοπολογία* και από *λειτουργία*. Η τοπολογία συχνά αναπαρίσταται από ένα γράφο και μερικούς πίνακες, πολλοί εκ των οποίων θα παρουσιαστούν στο κεφάλαιο αυτό και μερικοί ακόμη στα επόμενα κεφάλαια. Και θα υποθέσουμε, σε αυτό το κεφάλαιο, μερικά μοντέλα της λειτουργίας «αναζήτησης και πλοήγησης».

Ας υποθέσουμε ότι υπάρχουν *N* ιστοσελίδες. Κάθε ιστοσελίδα i έχει O_i **εξερχόμενους συνδέσμους** και I_i **εισερχόμενους συνδέσμους**. Δεν μπορούμε απλά να μετρήσουμε τον αριθμό των ιστοσελίδων που δείχνουν προς μία συγκεκριμένη ιστοσελίδα Α, διότι αυτός ο αριθμός, των **εισερχόμενων βαθμών** των κόμβων του γράφου των υπερσυνδέσμων, συχνά δεν αποδίδει το σωστό μέγεθος της σημαντικότητας.

Ας δηλώσουμε τη «βαθμολογία σημαντικότητας» της κάθε ιστοσελίδας ως π_i. Εάν σημαντικές ιστοσελίδες δείχνουν στην ιστοσελίδα Α, ίσως η ιστοσελίδα Α θα πρέπει να είναι και αυτή πάρα πολύ σημαντική, δηλαδή, $\pi_A = \sum_{i \to A} \pi_i$, όπου το άθροισμα είναι πάνω σε όλες τις ιστοσελίδες που δείχνουν την Α. Αλλά αυτό δεν είναι σωστό, δεδομένου ότι ο κόμβος i μπορεί να δείχνει σε πολλούς άλλους κόμβους στο γράφο και αυτό σημαίνει ότι ο καθένας από αυτούς τους κόμβους λαμβάνει μόνο ένα μικρό μέρος από τη βαθμολογία σημαντικότητας του κόμβου i.

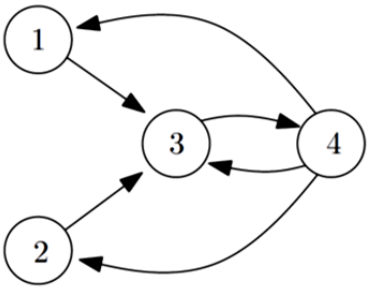
Ας υποθέσουμε ότι το μέτρο σημαντικότητας κάθε κόμβου είναι εξίσου κατανεμημένο σε όλες τις εξερχόμενες συνδέσμους/συνδέσεις. Για παράδειγμα, κάθε ένας από τους εξερχόμενους γείτονες του κόμβου i λαμβάνει βαθμολογία σημαντικότητας π_i/O_i. Τώρα κάθε βαθμολογία σημαντικότητας του κάθε κόμβου μπορεί επίσης να γραφεί ως το άθροισμα των βαθμολογιών σημαντικότητας που ελήφθησαν από το σύνολο των εισερχόμενων γειτόνων, κατηγοριοποιημένοι ως προς το j, π.χ., για τον κόμβο 1,

$$\sum_{j \to 1} \frac{\pi_j}{O_j}.$$

Εάν αυτό το άθροισμα είναι πράγματι π_1, έχουμε *συνέπεια* στις βαθμολογίες. Αλλά δεν είναι σαφές εάν μπορούμε εύκολα να υπολογίσουμε αυτές τις βαθμολογίες ή εάν σίγουρα υπάρχει πάντα ένα μοναδικό σύνολο συνεπών βαθμολογιών.

Αποδεικνύεται ότι με δύο τροποποιήσεις στη βασική ιδέα παραπάνω, υπάρχει πάντα ένα μοναδικό σύνολο συνεπών βαθμολογιών, που συμβολίζεται ως $\{\pi_i^*\}$. Αυτές οι βαθμολογίες καθορίζουν την κατάταξη των ιστοσελίδων: όσο υψηλότερη είναι η βαθμολογία τόσο υψηλότερη είναι η κατάταξη της ιστοσελίδας .

Για παράδειγμα, σκεφτείτε έναν πολύ μικρό γράφο με μόλις τέσσερις σελίδες και έξι υπερσυνδέσμους, όπως φαίνεται στο Σχήμα 3.1. Αυτός είναι ένας κατευθυνόμενος γράφος όπου κάθε κόμβος είναι μία ιστοσελίδα και κάθε σύνδεσμος μία υπερσύνδεση. Ένα συνεπές σύνολο των βαθμολογιών σημαντικότητας αποδεικνύεται ότι είναι το [0,125, 0,125, 0,375, 0,375]: οι ιστοσελίδες 3 και 4 είναι πιο σημαντικές από τις ιστοσελίδες 1 και 2. Σε αυτό το μικρό παράδειγμα, συμβαίνει επίσης το γεγονός ότι οι ιστοσελίδες 3 και 4, οι οποίες συνδέονται η μία με την άλλη, ωθούν και τις δύο σε υψηλότερη κατάταξη.



Σχήμα 3.1 *Ένα απλό παράδειγμα με τις βαθμολογίες σημαντικότητας τεσσάρων ιστοσελίδων και έξι συνδέσμων. Είναι ένας μικρός γράφος με μεγάλη συμμετρία, οδηγώντας σε έναν απλό υπολογισμό των βαθμολογιών σημαντικότητας των κόμβων.*

Διαισθητικά, οι βαθμολογίες αυτές βγάζουν νόημα. Πρώτον, λόγω της συμμετρίας του γράφου, οι ιστοσελίδες 1 και 2 πρέπει να έχουν την ίδια βαθμολογία σημαντικότητας. Μπορούμε να δούμε τις ιστοσελίδες 3 και 4 πρώτα, εφόσον αποτελούν μία ιστοσελίδα, τον υπερκόμβο 3+4. Δεδομένου ότι ο κόμβος 3+4 έχει δύο εισερχόμενες συνδέσεις και κάθε ένας από τους κόμβους 1 και 2 μόνο μία εισερχόμενη σύνδεση, ο κόμβος 3+4 πρέπει να έχει υψηλότερη βαθμολογία σημαντικότητας. Αφού ο κόμβος 3 δείχνει τον κόμβο 4 και αντίστροφα, οι βαθμολογίες σημαντικότητας αυτών των δύο κόμβων ενώνονται σε μία ίση κατανομή στην ισορροπία. Αυτή η συλλογιστική εξηγεί ποιοτικά τα πραγματικά αποτελέσματα που είδαμε.

Αλλά πώς μπορούμε να υπολογίσουμε τις ακριβείς βαθμολογίες; Σε αυτό το μικρό παράδειγμα ο υπολογισμός συνοψίζεται σε δύο απλές γραμμικές εξισώσεις. Έστω ότι η βαθμολογία του κόμβου 1 (και 2) είναι x και του κόμβου 3 (και 4) είναι y. Κοιτάζοντας τις εισερχόμενες συνδέσεις του κόμβου 1, βλέπουμε ότι υπάρχει μόνο μία τέτοια σχέση, που προέρχεται από τον κόμβο 4 που δείχνει τρεις κόμβους. Γνωρίζουμε λοιπόν ότι $x = y/3$. Δεδομένου ότι όλες οι βαθμολογίες θα πρέπει να αθροιστούν σε: $2x + 2y = 1$, έχουμε $x = 0{,}125$ και $y = 0{,}375$. Τώρα πώς μπορούμε να υπολογίσουμε αυτό το σύνολο συνεπών βαθμολογιών σε ένα τεράστιο, αραιό, γενικό διάγραμμα συνδεδεμένων υπερσυνδέσεων;

3.2 Μία Εκτενής Απάντηση

Σε κάθε μηχανή αναζήτησης, υπάρχουν δύο κύριες δραστηριότητες που δρουν συνεχώς στο παρασκήνιο: (α) η διάσχιση στο χώρο των συνδεδεμένων ιστοσελίδων του διαδικτύου για να συλλέξουν τις πληροφορίες της ιστοσελίδας, (β) η ευρετηρίαση αυτής της πληροφορίας σε συνοπτικές αναπαραστάσεις και η αποθήκευση των δεικτών της ευρετηρίασης. Όταν κάνετε αναζήτηση στο Google δίνεται το έναυσμα για μία διαδικασία κατάταξης που θα λαμβάνει υπόψη δύο βασικούς παράγοντες:

• Μία **βαθμολογία σχετικότητας**: Πόσο σχετικό με την αναζήτηση είναι το περιεχόμενο κάθε ιστοσελίδας;.

• Μία **βαθμολογία σημαντικότητας**: Πόσο σημαντική είναι η ιστοσελίδα;

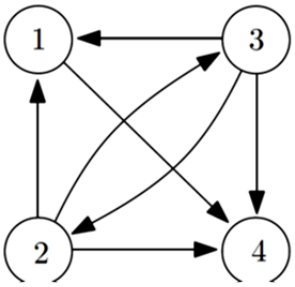
Είναι το άθροισμα της βαθμολογίας των δύο αυτών παραγόντων που καθορίζει την κατάταξη. Επικεντρωνόμαστε στη βαθμολογία σημαντικότητας, δεδομένου ότι συνήθως καθορίζει τη σειρά εμφάνισης στην κορυφή ενός μικρού αριθμού ιστοσελίδων σε οποιαδήποτε δημοφιλή αναζήτηση, και η οποία έχει τεράστιο αντίκτυπο στο πώς οι άνθρωποι λαμβάνουν πληροφορίες και πώς οι διαδικτυακές επιχειρήσεις δημιουργούν κίνηση στο διαδίκτυο.

Θα κατασκευάσουμε αρκετούς σχετικούς πίνακες: $\mathbf{H}$, $\widehat{\mathbf{H}}$ και $\mathbf{G}$, βήμα προς βήμα (αυτός ο πίνακας $\mathbf{G}$ δεν είναι ο πίνακας κέρδους του καναλιού του Κεφαλαίου 1. Δηλώνει τον πίνακα της Google σε αυτό το κεφάλαιο). Τελικά, θα πρέπει να υπολογίσουμε ένα ιδιοδιάνυσμα του $\mathbf{G}$ ως το διάνυσμα της βαθμολογίας σχετικότητας. Κάθε πίνακας είναι $N \times N$, όπου N είναι ο αριθμός των σχετικών ιστοσελίδων. Αυτοί είναι εξαιρετικά μεγάλοι πίνακες και θα συζητήσουμε την υπολογιστική αυτή πρόκληση της κλιμάκωσης στο Προχωρημένο Υλικό.

3.2.1 Κατασκευάζοντας τον H

Ο πρώτος πίνακας που ορίζουμε είναι ο $\mathbf{H}$: το στοιχείο (i, j) του πίνακα είναι ίσο με $1/O_i$ αν υπάρχει μία υπερσύνδεση από την ιστοσελίδα i στην ιστοσελίδα j ή 0 διαφορετικά. Αυτός ο πίνακας περιγράφει την τοπολογία του δικτύου: ποιες ιστοσελίδες δείχνουν σε ποιες. Επίσης, μοιράζει ομοιόμορφα τη σημαντικότητα της κάθε ιστοσελίδας, μεταξύ των εξερχόμενων γειτόνων της ή των ιστοσελίδων που αυτή δείχνει.

Έστω π ένα διάνυσμα $N \times 1$ μίας στήλης, το οποίο δηλώνει τις βαθμολογίες σημαντικότητας των ιστοσελίδων N. Ξεκινάμε με την εικασία ότι η σχετική βαθμολογία του φορέα είναι 1, απλά ένα διάνυσμα με άσσους καθώς κάθε ιστοσελίδα είναι εξίσου σημαντική με κάθε άλλη. Έτσι έχουμε ένα αρχικό διάνυσμα $\pi[0] = \mathbf{1}$, όπου το 0 σημαίνει την αρχική κατάσταση.



Σχήμα 3.2 *Δίκτυο με διασυνδεδεμένες ιστοσελίδες με έναν αιωρούμενο κόμβο 4.*

Στη συνέχεια, πολλαπλασιάζουμε τον π^T από δεξιά με τον πίνακα **H**. (Ορίζουμε ένα διάνυσμα να είναι ένα διάνυσμα στήλης. Έτσι, όταν πολλαπλασιάζουμε ένα διάνυσμα στα δεξιά με έναν πίνακα, βάζουμε το σύμβολο του ανάστροφου T στο πάνω μέρος του συμβόλου του διανύσματος). Μπορείτε να αναπτύξετε αυτό τον πολλαπλασιασμό πινάκων και να δείτε ότι υπάρχει αυτή η διάδοση της βαθμολογίας σημαντικότητας από την τελευταία επανάληψη εξίσου μεταξύ των εξερχόμενων συνδέσεων και να επανυπολογιστεί η βαθμολογία σημαντικότητας της κάθε ιστοσελίδας σε αυτήν την επανάληψη με την άθροιση των βαθμολογιών σημαντικότητας των εισερχόμενων συνδέσεων. Για παράδειγμα, ο $\pi_1[2]$, (για την ιστοσελίδα 1 στη δεύτερη επανάληψη) μπορεί να εκφραστεί ως το ακόλουθο σταθμισμένο άθροισμα των βαθμολογίων σημαντικότητας από την πρώτη επανάληψη:

$$\pi_1[2] = \sum_{j \to 1} \frac{\pi_j[1]}{O_j}.$$

Δηλαδή, το εσωτερικό γινόμενο του διανύσματος π από την προηγούμενη επανάληψη και της πρώτης στήλης του H:

$$\pi_1[2] = (\pi[1])^T \text{ (στήλη 1 του } \mathbf{H}).$$

Παρόμοια,

$$\pi_i[2] = (\pi[1])^T \text{ (στήλη } i \text{ του } \mathbf{H}), \forall i.$$

Αν δεικτοδοτήσουμε τις επαναλήψεις με βάση το k, η ενημέρωση σε κάθε επανάληψη είναι απλά:

$$\pi^T[k] = \pi^T[k-1]\mathbf{H}. \tag{3.1}$$

Ακολουθήσαμε την (οπτικά λίγο αδέξια) σύμβαση στον τομέα της έρευνας που ορίζει τον **H** έτσι ώστε η ενημερωμένη έκδοση να είναι ένας πολλαπλασιασμός του διανύσματος γραμμής π^T με το **H** από τα δεξιά.

Δεδομένου ότι οι απόλυτες τιμές των εγγραφών του π δεν έχουν σημασία, αλλά μόνο η σειρά κατάταξης, μπορούμε επίσης να κανονικοποιήσουμε το αποτέλεσμα π, έτσι ώστε οι εγγραφές να προστίθενται στο 1.

Τώρα προκύπτει το ερώτημα: οι επαναλήψεις στην (3.1) συγκλίνουν; Δηλαδή, υπάρχει ένα αρκετά μεγάλο K, έτσι ώστε, για κάθε $k \geq K$, το διάνυσμα $\pi[k]$ να είναι αυθαίρετα κοντά στο $\pi[k-1]$ (χωρίς να έχει σημασία η αρχική επιλογή $\pi[0]$); Αν ναι, έχουμε έναν τρόπο να υπολογίζουμε ένα συνεπές διάνυσμα με βαθμολογίες, με όποια ακρίβεια επιθυμούμε;

Αλλά η απάντηση είναι «όχι ακόμα». Χρειαζόμαστε δύο προσαρμογές στο **H**.

3.2.2 Κατασκευάζοντας τον $\widehat{\boldsymbol{H}}$

Πρώτον, ορισμένες ιστοσελίδες δεν έχουν συνδέσμους προς άλλες ιστοσελίδες. Αυτοί είναι οι «αιωρούμενοι κόμβοι» στο γράφο υπερσυνδέσεων . Για παράδειγμα, στο Σχήμα 3.2, ο κόμβος 4 είναι ένας αιωρούμενος κόμβος και στη σειρά του ο πίνακας Η έχει όλο μηδενικά:

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Δεν υπάρχουν συνεπή αποτελέσματα. Για να το δούμε αυτό, γράφουμε το σύστημα γραμμικών εξισώσεων $\pi^{\mathrm{T}} = \pi^{T}\mathbf{H}$:

$$\begin{cases} \frac{1}{3}(\pi_2 + \pi_3) = \pi_1 \\ \frac{1}{3}\pi_3 = \pi_2 \\ \frac{1}{3}\pi_2 = \pi_3 \\ \pi_1 + \frac{1}{3}(\pi_2 + \pi_3) = \pi_4. \end{cases}$$

Η επίλυση αυτών των εξισώσεων δίνει $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0$, το οποίο παραβιάζει τον κανόνα κανονικοποίησης $\sum_i \pi_i = 1$.

Μία λύση είναι να αντικαταστήσουμε κάθε γραμμή με μηδενικά, όπως την τελευταία γραμμή του **H** παραπάνω, με μία σειρά από $1/N$. Διαισθητικά, αυτό που θέλει να πει είναι ότι ακόμη και αν μία ιστοσελίδα δεν οδηγεί σε καμία άλλη ιστοσελίδα, θα την αναγκάσει να διαδώσει τη σημασία της βαθμολογίας της εξίσου σε όλες τις ιστοσελίδες που αναλύονται.

Από μαθηματική άποψη, αυτό ισοδυναμεί με την προσθήκη του πίνακα $\frac{1}{N}w1^T$ στον **H**, όπου το **1** είναι απλά ένα διάνυσμα άσσων, και το *w* είναι ένα διάνυσμα δείκτης με την i-οστή είσοδο ίση με 1, αν η ιστοσελίδα i δεν δείχνει σε άλλες ιστοσελίδες (αιωρούμενος κόμβος) και 0 διαφορετικά (μη αιωρούμενος κόμβος). Αυτό είναι ένα εξωτερικό γινόμενο μεταξύ δύο N-διάστατων διανυσμάτων, τα οποία οδηγούν σε έναν πίνακα $N \times N$. Για παράδειγμα, αν $N = 2$ και $w = [1\ 0]^T$, έχουμε

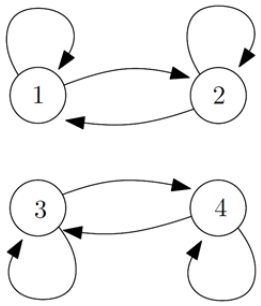
$$\frac{1}{2}\binom{1}{0}(1\ 1) = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 0 \end{pmatrix}.$$

Ο καινούριος πίνακας τον οποίο προσθέτουμε στον **H** είναι σαφώς απλός. Ακόμα και εάν είναι μεγάλος, $N \times N$, στην πραγματικότητα είναι το ίδιο διάνυσμα **w** επαναλαμβανόμενο N φορές. Τον ονομάζουμε πίνακα πρώτης τάξης. Ο πίνακας που προκύπτει,

$$\widehat{\mathbf{H}} = \mathbf{H} + \frac{1}{N}(w1^T)\ ,$$

έχει όλες τις καταχωρίσεις του μη αρνητικές και κάθε γραμμή έχει άθροισμα 1. Έτσι, μπορούμε να σκεφτούμε κάθε γραμμή ως φορέα πιθανότητας, με το στοιχείο (i, j) του $\widehat{\mathbf{H}}$ να υποδηλώνει την πιθανότητα ότι, αν είστε ήδη στην ιστοσελίδα *i*, θα επιλέξετε ένα σύνδεσμο και θα μεταβείτε στην ιστοσελίδα *j*.

Έτσι, η δομή του πίνακα λέει ότι θα είναι εξίσου πιθανό να επιλέξετε συνδέσμους που εμφανίζονται σε μία ιστοσελίδα και αν δεν υπάρχει κανένας υπερσύνδεσμος, θα είναι εξίσου πιθανό να επισκεφθείτε κάθε άλλη ιστοσελίδα. Η συμπεριφορά αυτή **ονομάζεται τυχαίος περίπατος στους γράφους** και μπορεί να μελετηθεί ως **αλυσίδες Markov** της θεωρίας πιθανοτήτων. Είναι προφανές ότι αυτό δεν είναι ακριβώς ένα μοντέλο της συμπεριφοράς περιήγησης, αλλά έχει επιτύχει μία πολύ αποτελεσματική ισορροπία μεταξύ της *απλότητας* του μοντέλου και της *χρησιμότητας* της προκύπτουσας κατάταξης των ιστοσελίδων. Θα δούμε ένα παρόμοιο μοντέλο επιρροής σε κοινωνικά δίκτυα στο Κεφάλαιο 8.



Σχήμα 3.3 *Ένα δίκτυο με υπερσυνδεδεμένες ιστοσελίδες με πολλαπλά συνεπή διανύσματα βαθμολογίας.*

3.2.3 Κατασκευάζοντας τον G

Αναφέραμε ότι υπάρχουν δύο θέματα με τον **H**. Το δεύτερο είναι ότι μπορεί να υπάρχουν *πολλά* συνεπή διανύσματα βαθμολογιών, όλα συμβατά με ένα δεδομένο πίνακα $\widehat{\mathbf{H}}$. Για παράδειγμα, για το γράφο στο Σχήμα 3.3, έχουμε

$$\mathbf{H} = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}.$$

Διαφορετικές επιλογές του $\pi[0]$ οδηγούν σε διαφορετικά αποτελέσματα π^*, τα οποία είναι όλα συνεπή. Για παράδειγμα, αν $\pi[0] = [1\ 0\ 0\ 0\ ]^T$, τότε $\pi^* = [0{,}5\ 0{,}5\ 0\ 0]^T$. Αν $\pi[0] = [0\ 0{,}3\ 0{,}7\ 0\ ]^T$, τότε $\pi^* = [0{,}15\ 0{,}15\ 0{,}35\ 0{,}35\ ]^T$.

Μία λύση στο πρόβλημα αυτό είναι να προσθέσουμε λίγη τυχαιοποίηση στην επαναληπτική διαδικασία και στον αναδρομικό ορισμό της σημαντικότητας. Διαισθητικά, μπορούμε να πούμε ότι υπάρχει μία πιθανότητα $(1-\theta)$ ότι θα μεταβείτε σε κάποια άλλη τυχαία ιστοσελίδα, χωρίς να επιλέξετε οποιαδήποτε από τις υπερσυνδέσεις στην τρέχουσα ιστοσελίδα.

Μαθηματικά, έχουμε προσθέσει μία ακόμη διάσταση $\frac{1}{N}\mathbf{1}\mathbf{1}^T$στον πίνακα, έναν πίνακα άσσων κλιμακούμενο από το $1/N$ (σαφώς έναν πίνακα τάξης 1), στο $\widehat{\boldsymbol{H}}$. Αλλά αυτή τη φορά είναι ένα σταθμισμένο άθροισμα με βάρος $\theta \in [0\ ,1]$. Το $(1-\theta)$ περιγράφει πόσο πιθανό είναι να μεταβείτε τυχαία σε κάποια άλλη ιστοσελίδα. Ο πίνακας που προκύπτει ονομάζεται **πίνακας Google**:

$$\mathbf{G} = \theta\widehat{\mathbf{H}} + (1-\theta)\frac{1}{\mathrm{N}}11^{\mathrm{T}}. \tag{3.2}$$

Τώρα μπορούμε να δείξουμε ότι, ανεξάρτητα από το διάνυσμα αρχικοποίησης π[0], η επαναληπτική διαδικασία:

$$\pi^T[\,k\,] = \pi^T[\,k-1\,]\mathbf{G} \tag{3.3}$$

θα συγκλίνει καθώς το $\kappa \to \infty$, και θα συγκλίνει σε ένα μοναδικό διάνυσμα π^* το οποίο αντιπροσωπεύει το συνεπές σύνολο των βαθμολογιών σημαντικότητας. Προφανώς, το π^* είναι το αριστερό ιδιοδιάνυσμα του **G** που αντιστοιχεί στην ιδιοτιμή 1:

$$\pi^{*T} = \pi^{*T}\mathbf{G}. \tag{3.4}$$

Κάποιος στη συνέχεια, μπορεί να κανονικοποιήσει το π^*: να πάρει το $\pi_i^*/\sum_j \pi_j^*$ με τη νέα τιμή του π_i^* και να κατατάξει τις εγγραφές σε φθίνουσα σειρά, πριν τα εμφανίσει στην ιστοσελίδα ως αποτελέσματα αναζήτησης με αυτήν τη σειρά. Ο πίνακας **G** είναι σχεδιασμένος έτσι ώστε να υπάρχει μία μοναδική λύση της (3.4) και η (3.3) να συγκλίνει με οποιεσδήποτε τιμές αρχικοποίησης.

Παρόλα αυτά, ο τρόπος υπολογισμού του π^*, λαμβάνοντας (την κανονικοποιημένη και διατεταγμένη έκδοσή του) του π^* ως τη βάση της κατάταξης, καλείται αλγόριθμος PageRank. Σε σύγκριση με τον DPC για τα ασύρματα δίκτυα στο Κεφάλαιο 1, ο πίνακας **G** στον Pagerank είναι πολύ μεγαλύτερος, αλλά μπορούμε να έχουμε έναν κεντρικό υπολογισμό.

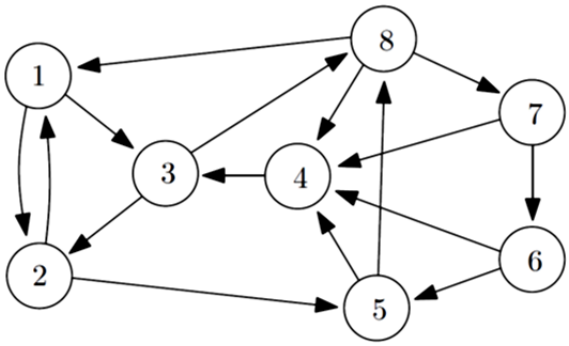
3.3. Παραδείγματα

Θεωρήστε το δίκτυο στο Σχήμα 3.4 με 8 κόμβους και 16 κατευθυνόμενες συνδέσεις. Έχουμε

$$\mathbf{H} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 0 & 1/3 & 0 \end{bmatrix}.$$

Εδώ $\widehat{\mathbf{H}} = \mathbf{H}$ αφού δεν υπάρχει κανένας αιωρούμενος κόμβος. Παίρνοντας $\theta = 0{,}85$, έχουμε

$$\mathbf{G} = \begin{bmatrix} 0{,}0188 & 0{,}4437 & 0{,}4437 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}0188 \\ 0{,}4437 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}4437 & 0{,}0188 & 0{,}0188 & 0{,}0188 \\ 0{,}0188 & 0{,}4437 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}4437 \\ 0{,}0188 & 0{,}0188 & 0{,}8688 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}0188 \\ 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}4437 & 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}4437 \\ 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}4437 & 0{,}4437 & 0{,}0188 & 0{,}0188 & 0{,}0188 \\ 0{,}0188 & 0{,}0188 & 0{,}0188 & 0{,}4437 & 0{,}0188 & 0{,}4437 & 0{,}0188 & 0{,}0188 \\ 0{,}3021 & 0{,}0188 & 0{,}0188 & 0{,}3021 & 0{,}0188 & 0{,}0188 & 0{,}3021 & 0{,}0188 \end{bmatrix}$$



Σχήμα 3.4 *Ένα παράδειγμα του αλγορίθμου PageRank με 8 σελίδες και 16 υπερσυνδέσεις. Η ιστοσελίδα 3 κατατάσσεται υψηλότερα παρόλο που η ιστοσελίδα 4 έχει το μεγαλύτερο βαθμό εισερχομένων συνδέσεων. Οι βαθμολογίες σημαντικότητας που υπολογίζονται από τον PageRank μπορεί να είναι αρκετά διαφορετικές από τους βαθμούς των κόμβων.*

Αρχικοποιώντας, $\pi[0] = [1/8\ 1/8 \cdots 1/8]^T$, η επανάληψη (3.3) δίνει

$$\pi[1] = [0{,}1073\ 0{,}1250\ 0{,}1781\ 0{,}2135\ 0{,}1250\ 0{,}0719\ 0{,}0542\ 0{,}1250\ ]^T$$

$$\pi[2] = [0{,}1073\ 0{,}1401\ 0{,}2459\ 0{,}1609\ 0{,}1024\ 0{,}0418\ 0{,}0542\ 0{,}1476\ ]^T$$

$$\pi[3] = [0{,}1201\ 0{,}1688\ 0{,}2011\ 0{,}1449\ 0{,}0960\ 0{,}0418\ 0{,}0606\ 0{,}1668\ ]^T$$

$$\pi[4] = [\ 0{,}1378\ 0{,}1552\ 0{,}1929\ 0{,}1503\ 0{,}1083\ 0{,}0445\ 0{,}0660\ 0{,}1450]^T$$

$$\pi[5] = [\ 0{,}1258\ 0{,}1593\ 0{,}2051\ 0{,}1528\ 0{,}1036\ 0{,}0468\ 0{,}0598\ 0{,}1468]^T$$

$$\pi[6] = [0{,}1280\ 0{,}1594\ 0{,}2021\ 0{,}1497\ 0{,}1063\ 0{,}0442\ 0{,}0603\ 0{,}1499\ ]^T$$

$$\vdots$$

και με ακρίβεια 4 δεκαδικών ψηφίων,
$\pi^* = [0{,}1286\ 0{,}1590\ 0{,}2015\ 0{,}1507\ 0{,}1053\ 0{,}0447\ 0{,}0610\ 0{,}1492\ ]^T$. Αυτό σημαίνει ότι η σειρά κατάταξης των ιστοσελίδων είναι: 3, 2, 4, 8, 1, 5, 7, 6.

Ο κόμβος με το μεγαλύτερο βαθμό εισερχομένων συνδέσεων, δηλαδή με τον μεγαλύτερο αριθμό συνδέσμων που δείχνουν προς ένα κόμβο, είναι ο κόμβος 4, ο οποίος δεν κατατάσσεται στην υψηλότερη θέση. Αυτό συμβαίνει επειδή η βαθμολογία σημαντικότητάς του έχει εξαπλωθεί κατά αποκλειστικότητα στον κόμβο 3 . Όπως θα δούμε και πάλι στο Κεφάλαιο 8, υπάρχουν πολλοί άλλοι χρήσιμοι τρόποι που μετρούν τη σημασία των κόμβων εκτός από το βαθμό εισερχομένων συνδέσεων.

3.4 Προχωρημένο Υλικό

3.4.1 Γενικευμένο PageRank και κάποιες βασικές ιδιότητες

Ο πίνακας Google **G** μπορεί να γενικευθεί, αν τα συστατικά τυχαιοποίησης είναι πιο εκλεπτυσμένα. Πρώτα, αντί του πίνακα $\frac{1}{N}\mathbf{1}\mathbf{1}^T$, μπορούμε να προσθέσουμε τον πίνακα $\mathbf{1}\mathbf{v}^T$ (πάλι, το εξωτερικό γινόμενο δύο διανυσμάτων), όπου η **v** μπορεί να είναι οποιαδήποτε κατανομή πιθανότητας. Βέβαια, το $\frac{1}{N}\mathbf{1}\mathbf{1}^T$ είναι μία ειδική περίπτωση αυτού.

Μπορούμε επίσης να γενικεύσουμε τον τρόπο μεταχείρισης των αιωρούμενων κόμβων: αντί της προσθήκης του $\frac{1}{N}w1^T$ στον **H** , όπου **w** είναι το διάνυσμα δείκτης των αιωρούμενων κόμβων, μπορούμε να προσθέσουμε το $w\mathbf{v}^T$. Και πάλι, το $\frac{1}{N}\mathbf{1}$ προκύπτει ως μία ειδική περίπτωση του **v**.

Τώρα, η εξίσωση ενημέρωσης της Google μπορεί να γραφεί στη μεγάλη μορφή (χωρίς να χρησιμοποιηθεί ο συμβολισμός συντομογραφίας **G**) ως συνάρτηση ενός δεδομένου πίνακα συνδεδεμένων ιστοσελίδων **H**, του διανύσματος **w** που δείχνει τις αιωρούμενες ιστοσελίδες και των δύο αλγοριθμικών παραμέτρων του βαθμού θ και του διανύσματος **v**:

$$\pi^T\mathbf{G} = \theta\pi^T\mathbf{H} + \pi^T(\theta\mathbf{w} + (1-\theta)\mathbf{1})\mathbf{v}^T. \qquad (3.5)$$

Θα πρέπει να επιβεβαιώσετε ότι η παραπάνω εξίσωση είναι πράγματι η ίδια με την (3.3).
Υπάρχουν πολλές απόψεις για την περαιτέρω ερμηνεία της (3.3) και τη σύνδεσή της με τη θεωρία πινάκων, τη θεωρία της αλυσίδας Markov και τη θεωρία γραμμικών συστημάτων. Για παράδειγμα:

- το π^* είναι το αριστερό ιδιοδιάνυσμα που αντιστοιχεί στην κυρίαρχη ιδιοτιμή ενός θετικού πίνακα,
- αντιπροσωπεύει τη λεγόμενη στάσιμη κατανομή μίας αλυσίδας Markov της οποίας οι πιθανότητες μετάβασης είναι στον **G**, και
- αντιπροσωπεύει την ισορροπία ενός μοντέλου οικονομικής ανάπτυξης σύμφωνα με τον **G** (περισσότερα για αυτή την άποψη αργότερα σε αυτή την ενότητα).

Οι κύριες προκλήσεις κατά τη λειτουργία της φαινομενικά απλής ενημερωμένης συνάρτησης (3.3) είναι η κλίμακα και η ταχύτητα: υπάρχουν δισεκατομμύρια ιστοσελίδες και η Google θα πρέπει να επιστρέψει τα αποτελέσματα σχεδόν αμέσως.

Όμως, η ισχυρή μέθοδος (3.3) προσφέρει πολλά αριθμητικά πλεονεκτήματα σε σύγκριση με έναν άμεσο υπολογισμό του κυρίαρχου ιδιοδιανύσματος **G**. Πρώτον, η (3.3) μπορεί να εκτελεστεί με τον πολλαπλασιασμό ενός διανύσματος με το άθροισμα του **H** και δύο πρώτης τάξης πινάκων. Αυτό είναι αριθμητικά απλό: το **H** είναι πολύ μεγάλο αλλά και πολύ αραιό: κάθε ιστοσελίδα συνδέεται συνήθως με λίγες άλλες ιστοσελίδες, έτσι σχεδόν όλες οι εγγραφές του **H** είναι μηδέν. Πολλαπλασιάζοντας με πίνακες πρώτης τάξης είναι επίσης εύκολο. Επιπλέον, σε κάθε επανάληψη, χρειάζεται μόνο να αποθηκεύσουμε το τρέχον διάνυσμα π.

Αν και δεν έχουμε ποσοτικοποιήσει την ταχύτητα της σύγκλισης, είναι σαφώς πολύ σημαντικό να επιταχύνουμε τον υπολογισμό του π^*. Όπως θα δούμε και πάλι στο Κεφάλαιο 8, η ταχύτητα σύγκλισης σε αυτή την περίπτωση ρυθμίζεται από τη δεύτερη μεγαλύτερη ιδιοτιμή $\lambda_2(\mathbf{G})$ του **G**, το οποίο μπορεί να αποδειχθεί ότι είναι περίπου θ. Έτσι, αυτή η παράμετρος θ ελέγχει την ανταλλαγή μεταξύ της ταχύτητας σύγκλισης και της σημασίας του γράφου υπερσυνδέσμων στον υπολογισμό των βαθμολογιών σημαντικότητας: το μικρότερο θ (πιο κοντά στο 0) κινεί τη σύγκλιση πιο γρήγορα, αλλά και μειώνει τη σημασία της δομής του γράφου των υπερσυνδέσεων. Αυτό δεν προκαλεί έκπληξη: αν δείτε τις βαθμολογίες της σημαντικότητας των ιστοσελίδων περισσότερο ως τυχαία αντικείμενα, είναι πιο εύκολο να υπολογιστεί η ισορροπία. Συνήθως, το $\theta = 0{,}85$ πιστεύεται ότι είναι μία πολύ καλή επιλογή. Αυτό οδηγεί σε σύγκλιση σε περίπου μερικές δεκάδες επαναλήψεις, ενώ εξακολουθεί να δίνει το μεγαλύτερο μέρος του βάρους στην πραγματική δομή του γράφου υπερσυνδέσεων (αντί του συστατικού τυχαιοποίησης στο **G**).

3.4.2 Ο PageRank ως λύση σε μία γραμμική εξίσωση

Ο PageRank ακούγεται παρόμοιος με τον κατανεμημένο έλεγχο ισχύος του Κεφαλαίου 1. Και οι δύο εφαρμόζουν τη μέθοδο ισχύος για την επίλυση ενός συστήματος γραμμικών εξισώσεων. Οι λύσεις σε αυτές τις εξισώσεις συλλαμβάνουν τον κατάλληλο μηχανικό διαμόρφωσης στο δίκτυο, είτε αυτός είναι η σχετική σημαντικότητα των ιστοσελίδων σε ένα γράφο υπερσυνδέσεων είτε το καλύτερο διάνυσμα της ισχύος εκπομπής σε ένα ασύρματο περιβάλλον με παρεμβολές. Αυτό το εννοιολογικό πλαίσιο μπορεί να μετασχηματιστεί σε έναν ακριβή παράλληλο τύπο παρακάτω.

Πρώτα, μπορούμε να ξαναγράψουμε το χαρακτηρισμό του π^*ως τη λύση στην ακόλουθη γραμμική εξίσωση (και όχι ως το κυρίαρχο αριστερό ιδιοδιάνυσμα του πίνακα **G** (3.4) , η οποία είναι η άποψη που έχουμε ακολουθήσει μέχρι τώρα):

$$(\mathbf{I} - \theta\mathbf{H})^{\mathrm{T}}\pi = \mathbf{v} \tag{3.6}$$

Θα αποδείξουμε σε λίγο ότι η (3.6) είναι αληθής. Συγκρίνοντας την (3.6) με τον χαρακτηρισμό του βέλτιστου διανύσματος εκπομπής ισχύος στον κατανεμημένο αλγόριθμο ελέγχου ισχύος στο Κεφάλαιο 1: