Ανάκληση Πληροφορίας

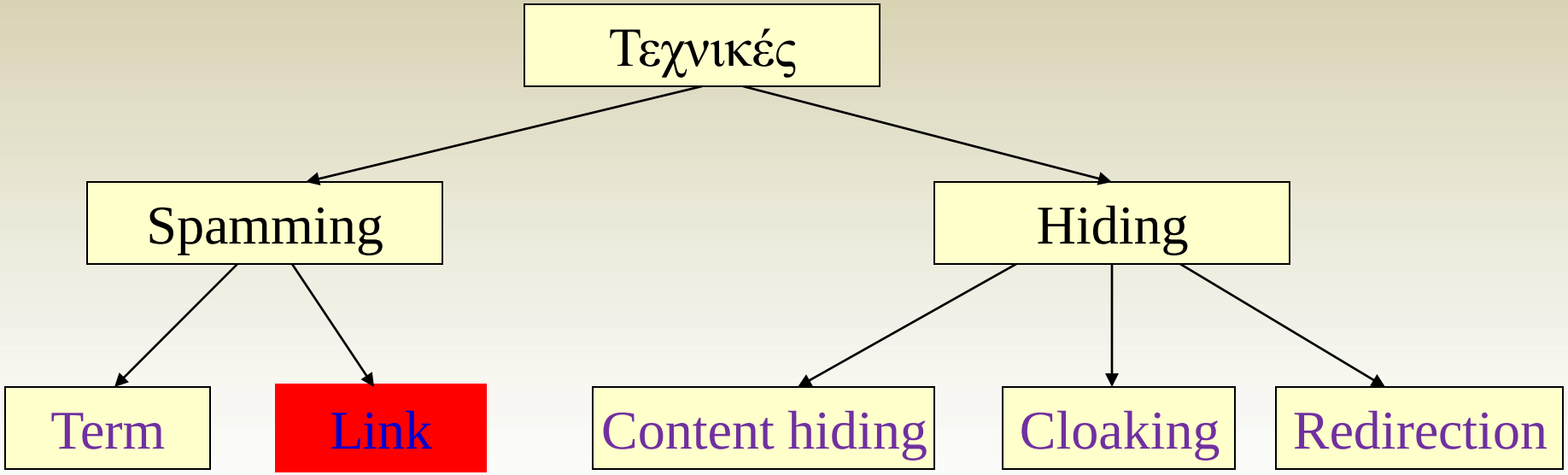
Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 18η

Spamming PageRank

(Link Spam Farms)

- **Spamming**: Παραπλάνηση των μηχανών αναζήτησης για να αποκτηθεί υψηλότερη διάταξη (ranking) για κάποιες σελίδες (ή ιστοτόπους) απ' αυτή που πραγματικά αξίζουν.



Ο αλγόριθμος PageRank

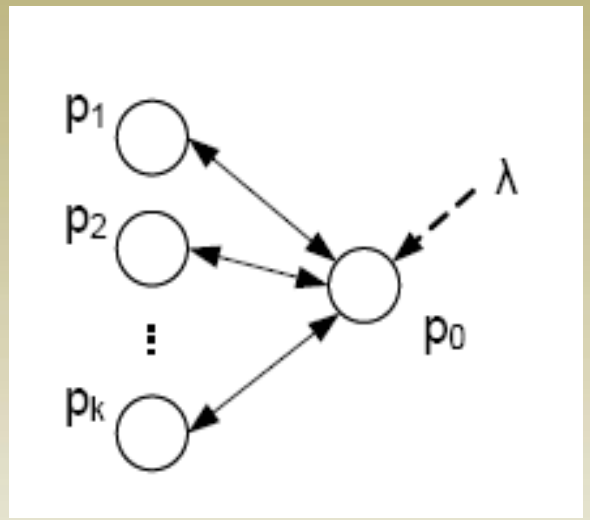
$$\mathbf{p} = c\,\mathbf{T}'\,\mathbf{p} + \frac{1-c}{N}\,\mathbb{1}_N$$

- c: η σταθερά damping
- **T**: ο πίνακας μεταβάσεων
- N: συνολικός αριθμός σελίδων του Web
- $\mathbf{1}_N$: διάνυσμα που όλα τα στοιχεία του είναι ίσα με 1

- Στην ουσία, αυτή η έκφραση οδηγεί σε διατύπωση του προβλήματος PageRank ως γραμμικό σύστημα

Spam Farm για εξύψωση μιας σελίδας (1/4)

- Υποθέσεις:
  - Κάθε σελίδα της φάρμας δείχνει μόνο προς τη μια και μοναδική σελίδα-στόχο, της οποίας ο spammer θέλει ν' αυξήσει το PageRank. Αυτή η σελίδα είναι μέρος της φάρμας
  - Η φάρμα αποτελείται από δεδομένο αριθμό *k* σελίδων, λόγω κόστους συντήρησης, ή πόρων
  - Είναι πιθανό, εκτός των σελίδων της φάρμας, ο spammer να κατορθώσει να αποκτήσει συνδέσμους προς τη σελίδα που θέλει και διαμέσου έγκριτων πηγών, π.χ., από Web directory, ή από unmoderated bulletin boards. Αυτούς τους συνδέσμους θα τους ονομάζουμε *hijacked links* και το PageRank που φτάνει στη φάρμα διαμέσου αυτών θα καλείται *leakage* λ
  - Ενώ ο spammer έχει πλήρη έλεγχο των σελίδων της φάρμας, δεν έχει τον πλήρη έλεγχο των σελίδων που περιέχουν τους hijacked links



Spam Farm για εξύψωση μιας σελίδας (2/4)

- **ΘΕΩΡΗΜΑ**. Η τιμή PageRank p_0 της σελίδας-στόχος του προηγούμενου σχήματος είναι:

$$p_0 = \frac{1}{1-c^2}\left[c\lambda + \frac{(1-c)(ck+1)}{N}\right]$$

- **ΑΠΟΔΕΙΞΗ**. Σύμφωνα με την προηγούμενη διατύπωση του PageRank, η τιμή PageRank των σελίδων της φάρμας είναι:

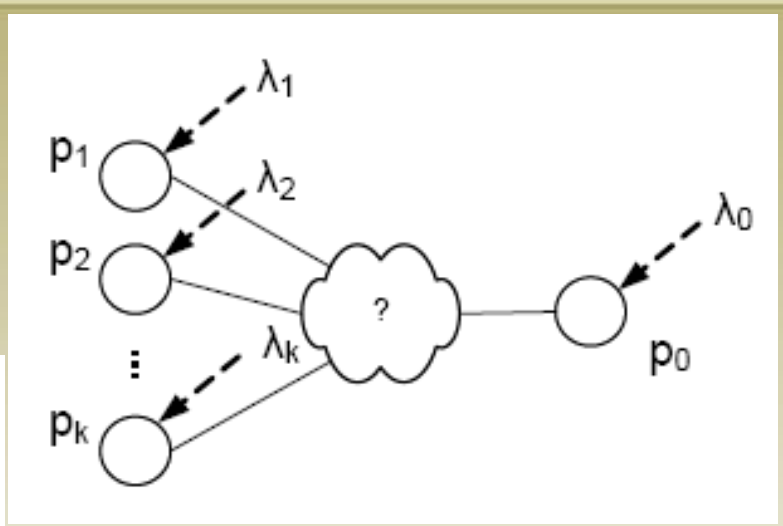
$$\begin{cases} p_0 = c\lambda + c\sum_{i=1}^{k} p_i + (1-c)/N \\ p_i = cp_0/k + (1-c)/N, \text{ for } i = 1, \ldots, k \end{cases}$$

Αντικαθιστώντας την τιμή των p_i, έχουμε:

$$p_0 = c\lambda + ck\left(\frac{cp_0}{k} + \frac{1-c}{N}\right) + \frac{1-c}{N}$$
$$= \frac{1}{1-c^2}\left[c\lambda + \frac{(1-c)(ck+1)}{N}\right].$$

Spam Farm για εξύψωση μιας σελίδας (3/4)

- **Βέλτιστη δομή της φάρμας**. Έστω ότι με **p** και **λ** συμβολίζουμε τα διανύσματα που αντιπροσωπεύουν τις τιμές *PageRank* και του *leakage* των σελίδων της φάρμας:



$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix} \quad \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{pmatrix}$$

Τότε η εξίσωση του PageRank για τις σελίδες της φάρμας είναι:

$$\begin{pmatrix} p_0 \\ \mathbf{p} \end{pmatrix} = c \begin{pmatrix} \lambda_0 \\ \boldsymbol{\lambda} \end{pmatrix} + c \begin{pmatrix} 0 & \mathbf{e}' \\ \mathbf{f} & \mathbf{G} \end{pmatrix} \begin{pmatrix} p_0 \\ \mathbf{p} \end{pmatrix} + \frac{1-c}{N} \mathbb{1}_{k+1}$$

- **ΘΕΩΡΗΜΑ**. Η τιμή PageRank p_0 της σελίδας-στόχος του προηγούμενου σχήματος είναι μέγιστη εάν $\mathbf{e}=\mathbf{1}_k$, $\mathbf{1}'_k\mathbf{f}=1$, $\mathbf{G}=\mathbf{0}_{kxk}$ και $\lambda_0=\lambda$ $(=\lambda_0+\lambda_1+\lambda_2+\ldots)$ και $\lambda_i=0$ $\forall$ $i=1,\ldots,k$

Spam Farm για εξύψωση μιας σελίδας (4/4)

- Μ' άλλα λόγια, η δομή της φάρμας είναι βέλτιστη, εάν:
  - Όλες οι boosting σελίδες δείχνουν και δείχνονται από τη σελίδα-στόχο ($\mathbf{e}=\mathbf{1}_k$)
  - Δεν υπάρχουν σύνδεσμοι μεταξύ των boosting σελίδων ($\mathbf{G}=\mathbf{0}_{kxk}$)
  - Η σελίδα-στόχος δείχνει σε μερικές ή όλες τις boosting σελίδες ($\mathbf{1}'_k\mathbf{f}=1$)
  - Όλοι οι hijacked σύνδεσμοι δείχνουν στη σελίδα-στόχο ($\lambda_0=\lambda$ και $\lambda_i=0\ \forall\ i=1,\ldots,k$)

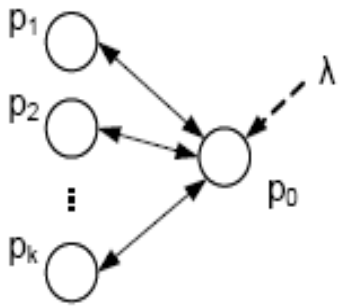


Figure 1: An optimal structure for a single spam farm with one target page.

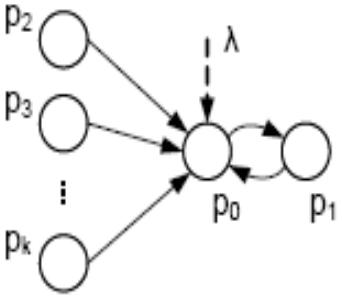
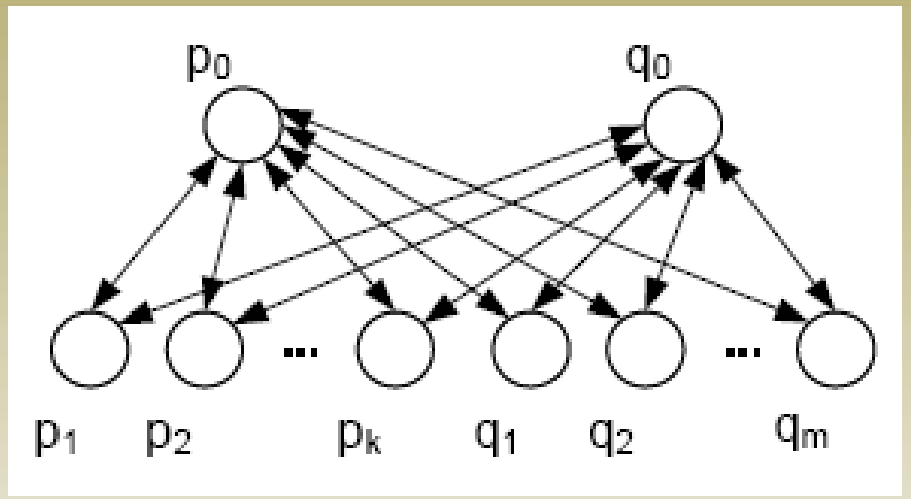


Figure 3: Another optimal structure for a single spam farm with one target page.

Συμμαχίες link spam farms: Δυο φάρμες

- Η μία φάρμα έχει k boosting σελίδες και η άλλη έχει m boosting σελίδες
- Χωρίς να συνδέονται οι φάρμες μεταξύ τους, η μέγιστη τμή της σελίδας-στόχος είναι:



$$\bar{p}_0 = \frac{ck+1}{(1+c)N} \qquad \bar{q}_0 = \frac{cm+1}{(1+c)N}$$

- Εάν κάνουμε την συνδεσμολογία των δυο farms με τον τρόπο που φαίνεται στο πιο πάνω σχήμα, τότε:

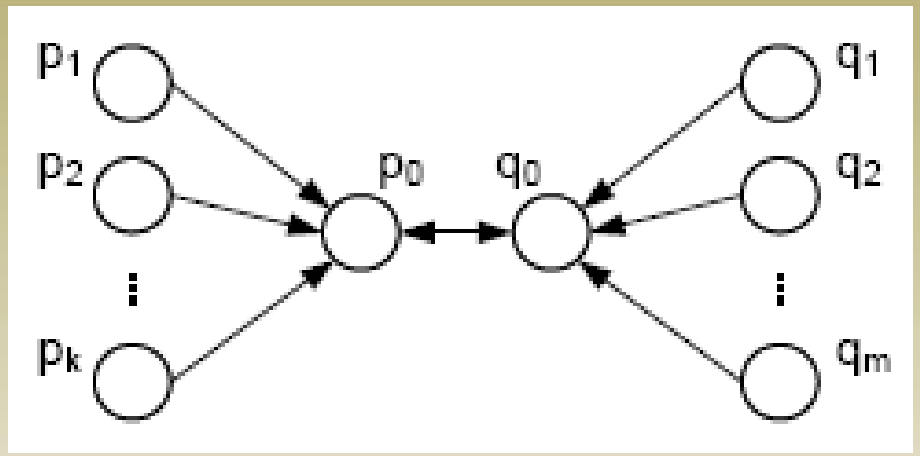
$$p_0 = q_0 = (\bar{p}_0 + \bar{q}_0)/2$$

- Συνεπώς, κερδίζει ο spammer που έχει τις λιγότερες σελίδες στη φάρμα του!

Συμμαχίες link spam farms: Δυο φάρμες

- Εάν εκτελέσουμε τη διπλανή συνδεσμολογία, τότε
  - $p_0 = q_0$
  - και:

$$p_0 = \frac{ck + c^2 m}{(1+c)N} + \frac{1}{N}$$



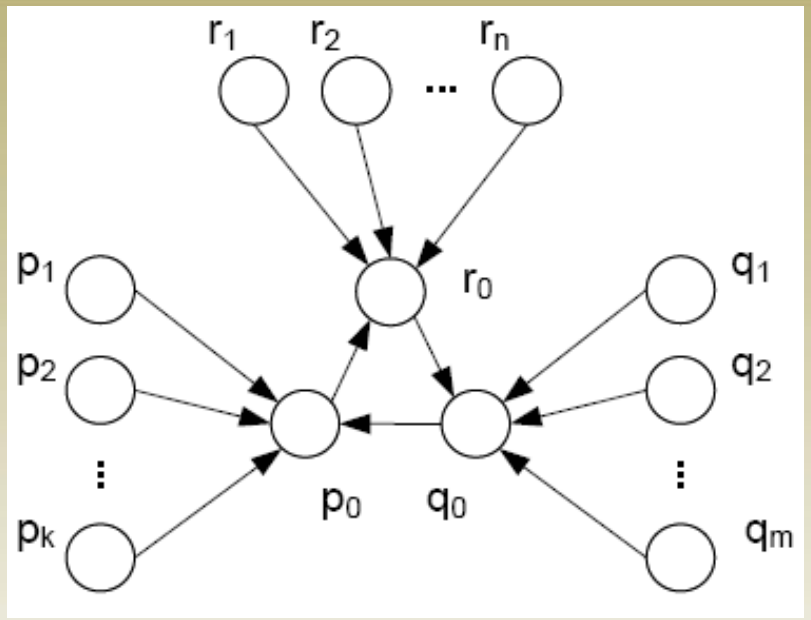
- Άρα ωφελούνται και οι δυο, κατά ποσά ανάλογα του μεγέθους της άλλης φάρμας, που είναι το ζητούμενο για τους spammers:

$$(p_0 - \bar{p}_0) \propto m \qquad (q_0 - \bar{q}_0) \propto k$$

Συμμαχίες link spam farms: Δακτύλιοι

- Εάν έχουμε F φάρμες, και συμβολίσουμε με t_i την τιμή PageRank της σελίδας-στόχου κάθε φάρμας και με b_i τον αριθμό των boosting σελίδων κάθε φάρμας, τότε το PageRank score της πρώτης σελίδας-στόχου θα είναι:

$$t_1 = \frac{\sum_{j=1}^{F} c^j b_j}{N \sum_{j=1}^{F} c^{j-1}} + \frac{1}{N}$$

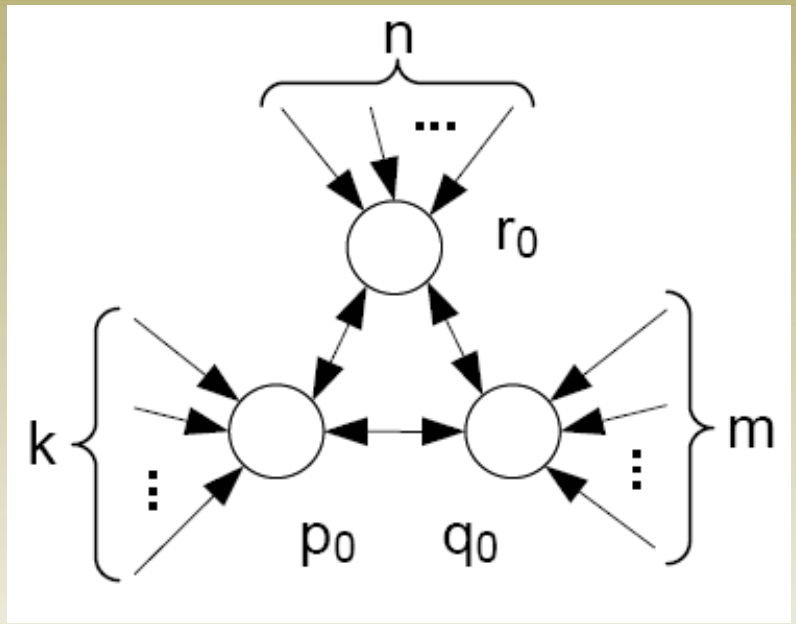


- Γενικά, η τιμή PageRank της i-οστής σελίδας-στόχου θα είναι:

$$t_i = \frac{\sum_{j=i}^{F} c^{j-i+1} b_j + \sum_{j=1}^{i-1} c^{j+F-i+1} b_j}{N \sum_{j=1}^{F} c^{j-1}} + \frac{1}{N}$$

Συμμαχίες link spam farms: Κλίκες

- Εάν συμβολίσουμε με t_i την τιμή PageRank της σελίδας-στόχου κάθε φάρμας και με b_i τον αριθμό των boosting σελίδων κάθε φάρμας, τότε το PageRank score της πρώτης σελίδας-στόχου θα είναι:



$$p_0 = \frac{2ck - c^2k + c^2m + c^2n}{(2+c)N} + \frac{1}{N}$$

- Γενικά, η τιμή PageRank της i-οστής σελίδας-στόχου θα είναι:

$$t_i = \frac{c(1-c)(F-1)b_i + c^2 \sum_{j=1}^{F} b_j}{(F+c-1)N} + \frac{1}{N}$$

Ζητήματα στη δομή των link spam farms

- Πότε έχει νόημα να συμμετάσχει μια νέα φάρμα σε μια ήδη υπάρχουσα συμμαχία;
  - Πόσες σελίδες πρέπει να έχει η νέα φάρμα, ώστε να ωφελήσει και τις υπάρχουσες φάρμες;
- Πότε έχει νόημα να αποχωρήσει μια φάρμα από μια συμμαχία στην οποία συμμετέχει;
  - Υπάρχει κάποιος critical αριθμός κόμβων, πέρα από τον οποίο είναι καλύτερα η φάρμα να υπάρχει μόνη της;
- Αφού οι βέλτιστες δομές των link spam farms είναι εύκολα ανιχνεύσιμες από τις μηχανές αναζήτησης, είναι πιθανό ότι οι spammers θα δημιουργήσουν ακανόνιστες δομές που όμως θα μοιάζουν με τις βέλτιστες; Πώς τις ανιχνεύουμε αυτές;