Link spam detection through spectral clustering on Markov chains

Ing. José Gómer González Hernández
Maestría en Ingeniería de Sistemas y Computación
2007

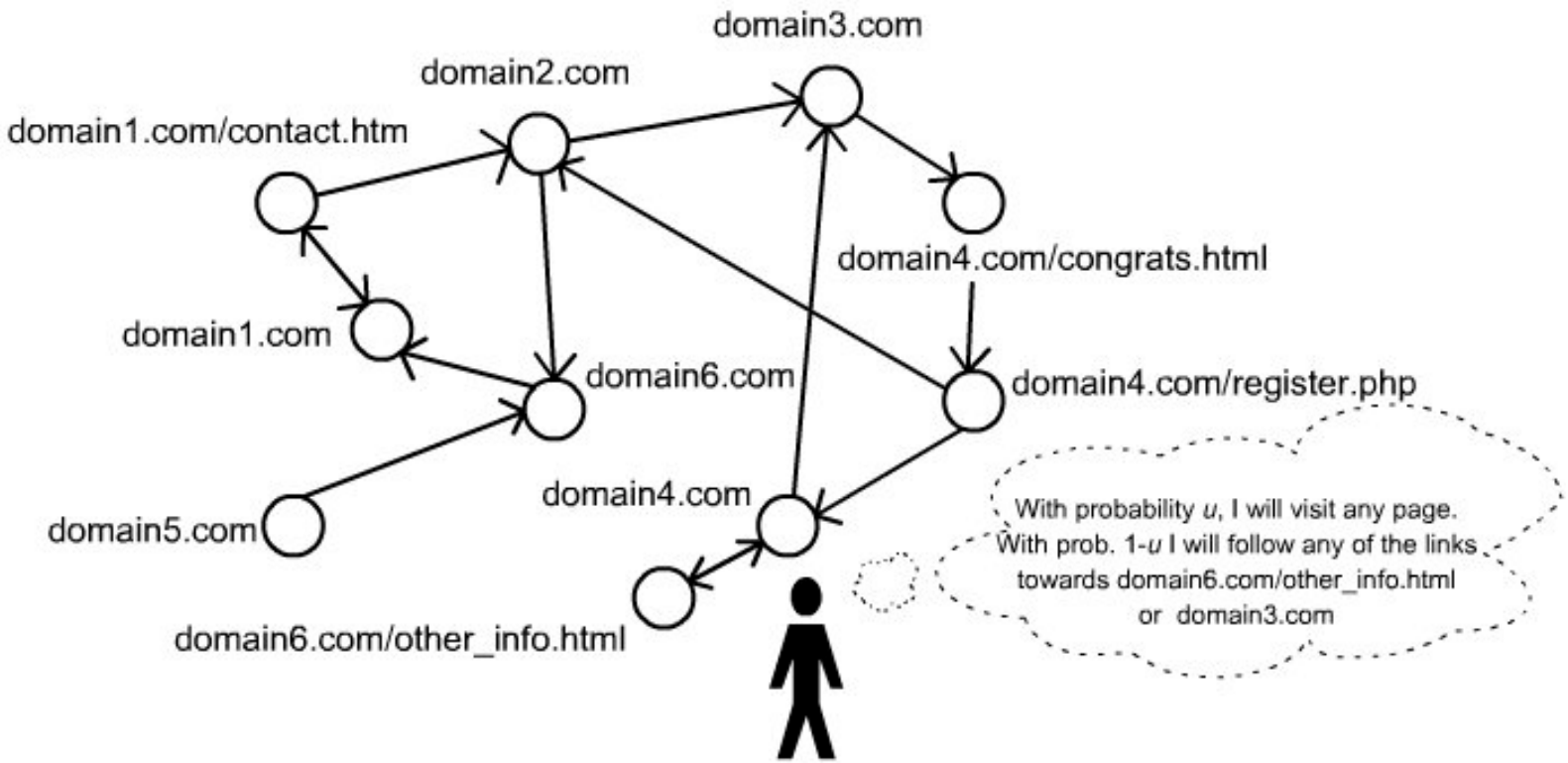
Outline

- Link spam
- PageRank
- Manipulating PageRank
- Previous work on link spam
- A new approach using conductance
- Project’s objectives

Link spam: a recent and compelling problem

- Ranking highly in the web brings commercial advantages for a website owner
- Thus, search engines' algorithms have become target of manipulation: web spam
- Misleading a *ranking* algorithm is known as *link spam*
- Harmful consequences for both users and search engines

Google's PageRank: a random surfer



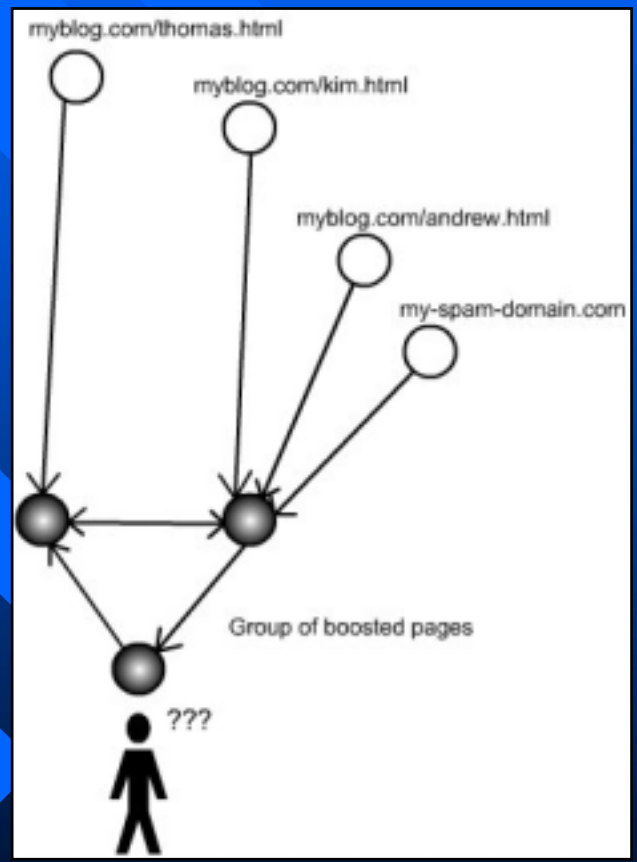
A creature that crawls the web, visiting one page at a time, deciding which one to visit next from the outlinks in the current page. At each visit, he gets bored with probability *u*. In that case he jumps to any of all the pages.

The PageRank of a page

- In the long run, the random surfer visits a page *i* with probability π_i
- π_i is the PageRank of *i*: the (global) measure of importance/popularity of page *i* in the whole web
- The walk followed by the creature can be regarded as a Markov chain whose steady-state probability distribution is π
- This chain is ergodic because of the random jump, ensuring existence and uniqueness of π

Manipulating the algorithm

- Link nepotism as a form link spamming
- Point to one's pages as much as possible (through forums, blogs, wikis, etc.) to boost the probability of being visited in the random walk.
- Once visited, manage to trap the surfer (in probabilistic terms) within the group of pages.

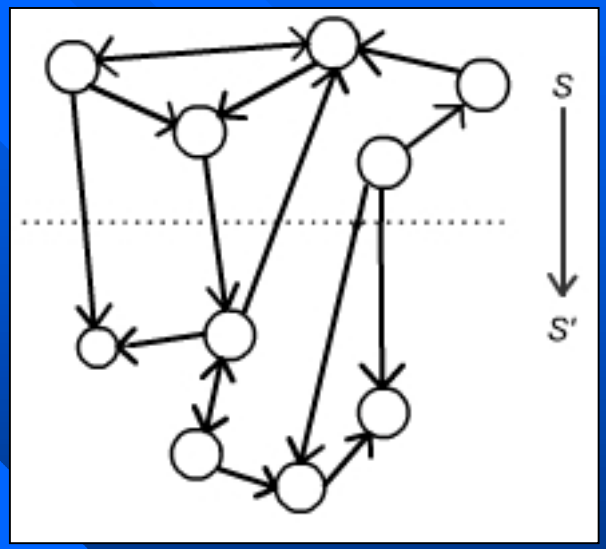


Actions to prevent manipulation

- Naïve approach: use a high jumping factor ($u \to 1$)
- Jump to trusted sites only
- Maintain white/black lists to propagate notions of trust/distrust
- Build classifiers from features: title, keywords, content (HTML code), words in the URL, IP address, out-degree, in-degree, etc.

A new direction: conductance

- In a Markov chain, conductance ϕ measures the chance of leaving a subset in one step
- A low conductance $\phi(S)$ implies the random surfer can be easily trapped inside S
- Low conductance is a necessary condition in a colluding group of pages



$$\phi(S) = \frac{\sum_{i \in S, j \in S'} \pi_i p_{ij}}{\sum_{i \in S} \pi_i}$$

The problem

- Find subsets of pages where conductance is below a certain threshold
- A similar problem in its formulation is that of spectral clustering
  - On an undirected graph $G=(V,E)$ find disjoint subsets $\{C_1,C_2,...,C_l\}$ such that $C_i \subset V$ and $\Phi(C_i) \leq \alpha$
  - Φ is called conductance
- Markov chains and graphs are not the same thing, so ϕ and Φ does not reflect the same. How to relate them?

The connection

- If a new Markov chain is built such that transition probabilities are $p'_{ij} = \frac{1}{2}(p_{ij} + \pi_j p_{ji} / \pi_i)$, we have:
  - Stationary distribution is still being π
  - Conductance remains $\phi(S)=\phi'(S)$ for all S
  - The new chain is reversible $\pi_i p'_{ij} = \pi_j p'_{ji}$
- If a matrix is built so that $w_{ij} = \pi_i p'_{ij}$ we have that w represents an undirected graph where $\phi(S)=\Phi(S)$
- Conclusion: apply spectral clustering on such a graph will lead to pages under presumable collusion

Aims of the project

- By applying spectral clustering on graphs obtained from modest-size portions of the web, determine whether conductance is actually a good criterion in the practice for link spam detection
- Contrast this new approach against already seen strategies in terms of quality and feasibility
- Analyse the computational complexity of the resultant algorithm to derive conclusions about scalability (application to real-size web graphs)

References

[1] S. Brin and L. Page. Anatomy of a large-scale hypertextual web search engine. In World Wide Web Conference, 1998.

[2] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. Technical report, Computer Science Department, Stanford University, 2005.

[3] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. Journal of the ACM, 51(3): 497-515, 2004.

[4] R. Montenegro and P. Tetali. Mathematical aspects of mixing times in markov chains. Foundations and Trends in Theoretical Computer Science, 1(3): 237-354, 2006.