MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Martin Olsen
University of Aarhus

Link Building and Communities in Large Networks

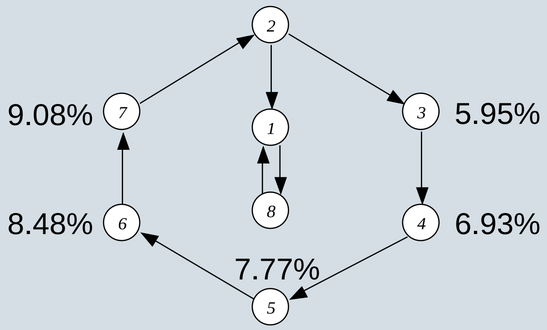
Link Building

In the real world people try to identify optimal places for signs. The corresponding problem in cyberspace is to identify optimal places (web pages) for links to web pages.

The PageRank Algorithm

The PageRank algorithm used by Google assumes that a web surfer visiting a page p will choose the next page to visit following these rules: With probability 0.85 the surfer follows a link from p chosen uniformly at random. With probability 0.15 the surfer chooses to visit another page chosen uniformly at random. The PageRank value of a page is the probability that a web surfer will visit the page after s steps for large s.

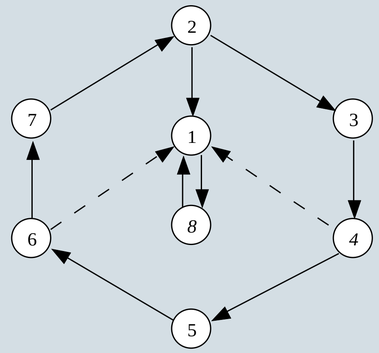
A Simple Link Building Problem



Problem: Find the page u in {3, 4, 5, 6, 7} for which we would achieve the maximum increase in the PageRank value of page 1 by adding the link (u, 1). The percentages are the current PageRank values of the nodes. The solution appears somewhere on the poster.

This problem can be solved in time corresponding to a *constant* and *small* number of PageRank computations [1].

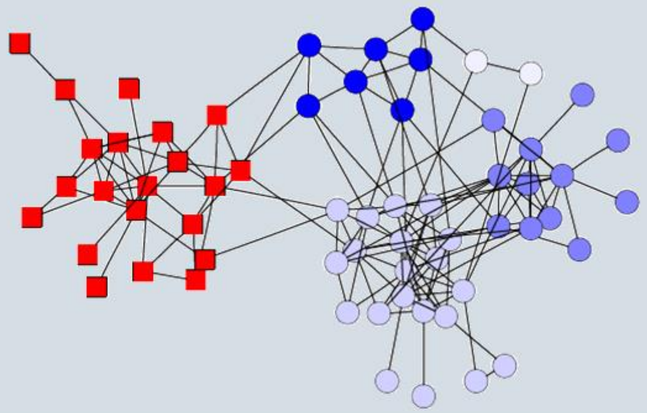
Link Building is NP-Hard



The dashed links show the set of two new links maximizing the PageRank value of page 1. Computing a set of k new links maximizing the minimum PageRank value for a given set of web pages is NP-hard [1].

Communities

You might want to restrict your search for optimal new links to the community of a page (or a set of pages). But how can you define a community of nodes in a graph?

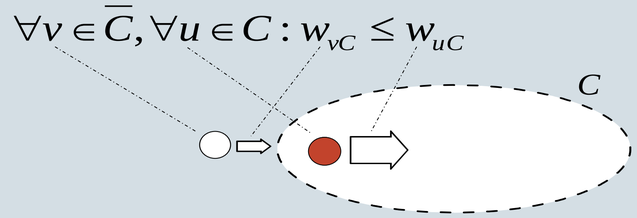


62 dolphins in Doubtful Sound (New Zealand) were observed. The graph shows a real community structure of the dolphins. A dolphin has at least as many friends in its own community compared to any other community. Such structures are NP-hard to compute [2].

Identification of Community Members

In [3] a community is defined as a set C for which every member of C has relatively more links to nodes in C compared to any non-member:

$$\forall v \in \overline{C}, \forall u \in C : w_{vC} \leq w_{uC}$$



$$w_{vC} = \frac{\#\text{links from } v \text{ to } C}{\#\text{links from } v}$$

Such communities are also NP-hard to compute. A simple and efficient greedy approach was used to find a community of 556 Danish computer science sites. The sites were ranked with a local version of the PageRank algorithm:

1) www.daimi.au.dk (CS U Aarhus)
2) www.diku.dk (CS U Copenhagen)
3) **www.itu.dk (ITU Copenhagen)**
4) **www.cs.auc.dk (CS U Aalborg)**
5) www.brics.dk (CS PhD School)
6) **www.imm.dtu.dk** (Informatics/Mathematical modeling DTU Copenhagen)
17) **www.imada.sdu.dk** (CS/Mathematics U Southern Denmark)

(The sites marked with bold font were given to the greedy approach as ”known” members of the community)

References

[1] *The Computational Complexity of Link Building*, Olsen, submitted
[2] *Nash Stability in Additively Separable Hedonic Games is NP-hard*, Olsen, Computability in Europe 2007
[3] *Communities in Large Networks: Identification and Ranking*, Olsen, Fourth Workshop on Algorithms and Models for the Web-Graph, 2006

Solution to the problem: $u = 5$