



LiWA
Living Web Archives

Web Spam

Marc Spaniol



mpg

Databases and
Information Systems
Prof. Dr. G. Weikum
MPII-Sp-0709-1/49

Web Dynamics

Web Spam

Marc Spaniol

Saarbrücken, July 23, 2009



Agenda

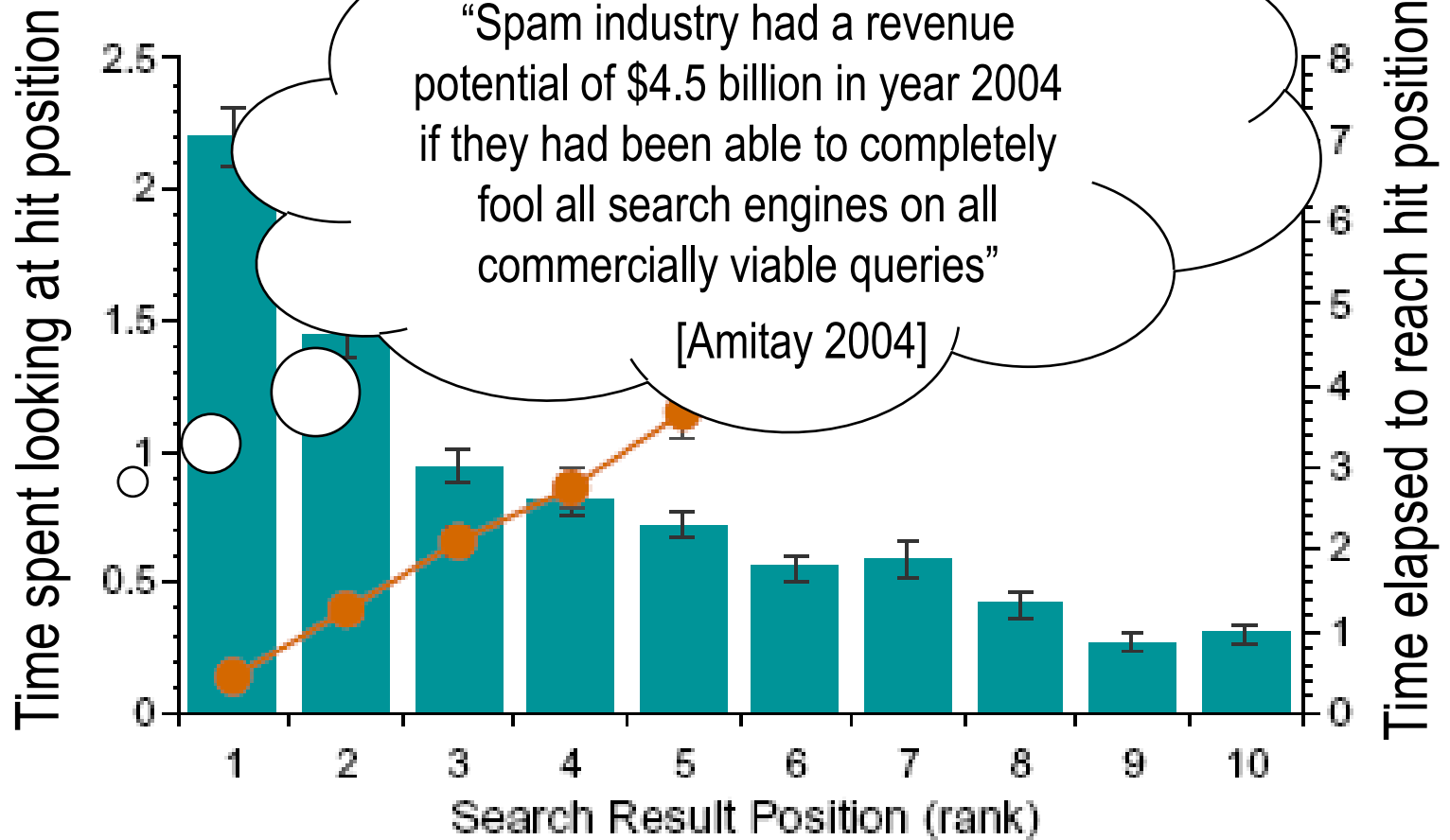
- Web spam
 - Why and what?
 - Spam taxonomy
 - Overview
 - Strategies in detail
 - o Link spam
 - o Link farms
 - Examples
- Countermeasures
 - Spam detection
 - Labeling and assessment
 - Combating spam
 - Web spam challenge
- Conclusion



Web Spam Why?

Web Spam

Marc Spaniol

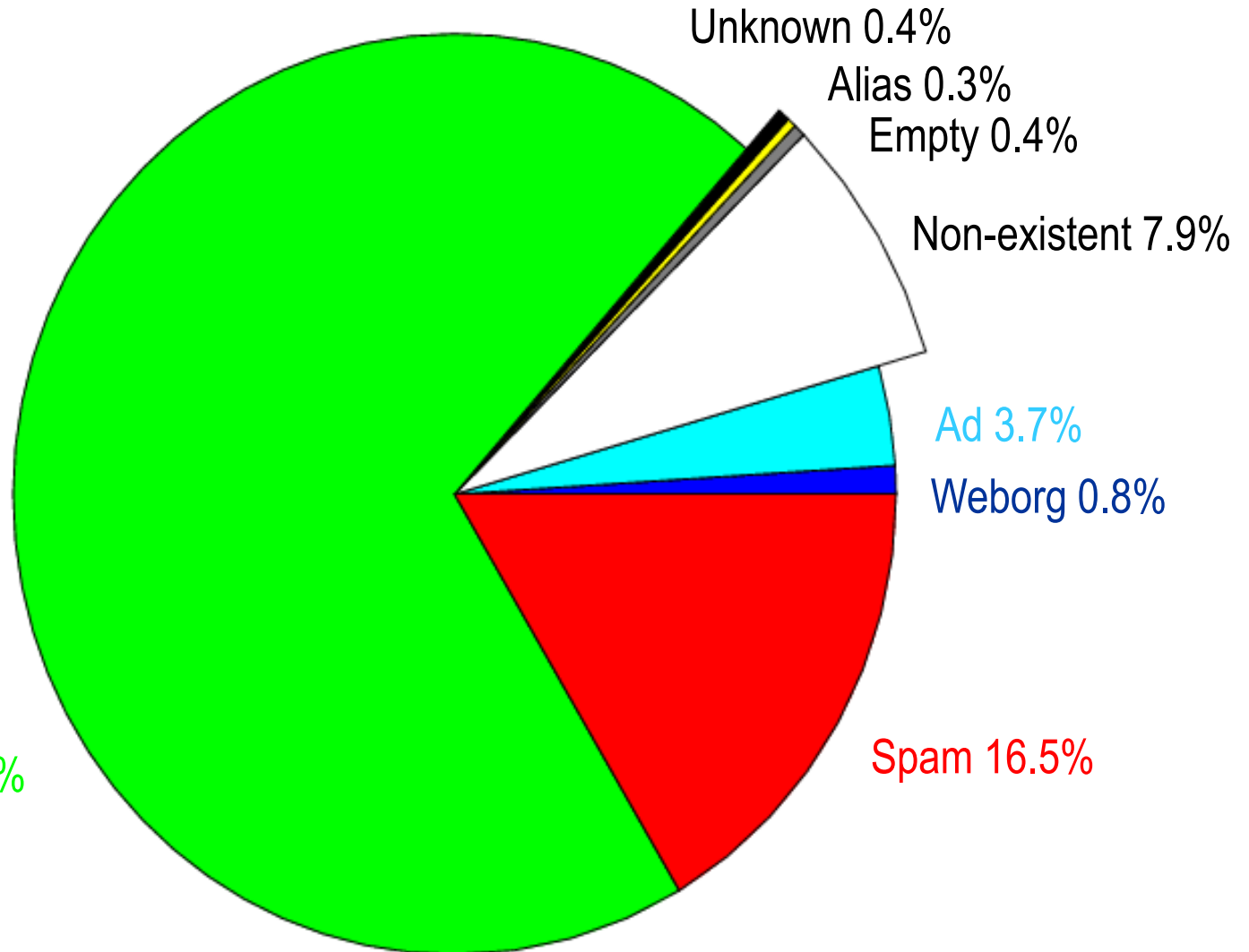




Web Spam

What's the Problem?

2004 .de crawl
Courtesy: T. Suel





Web Spam

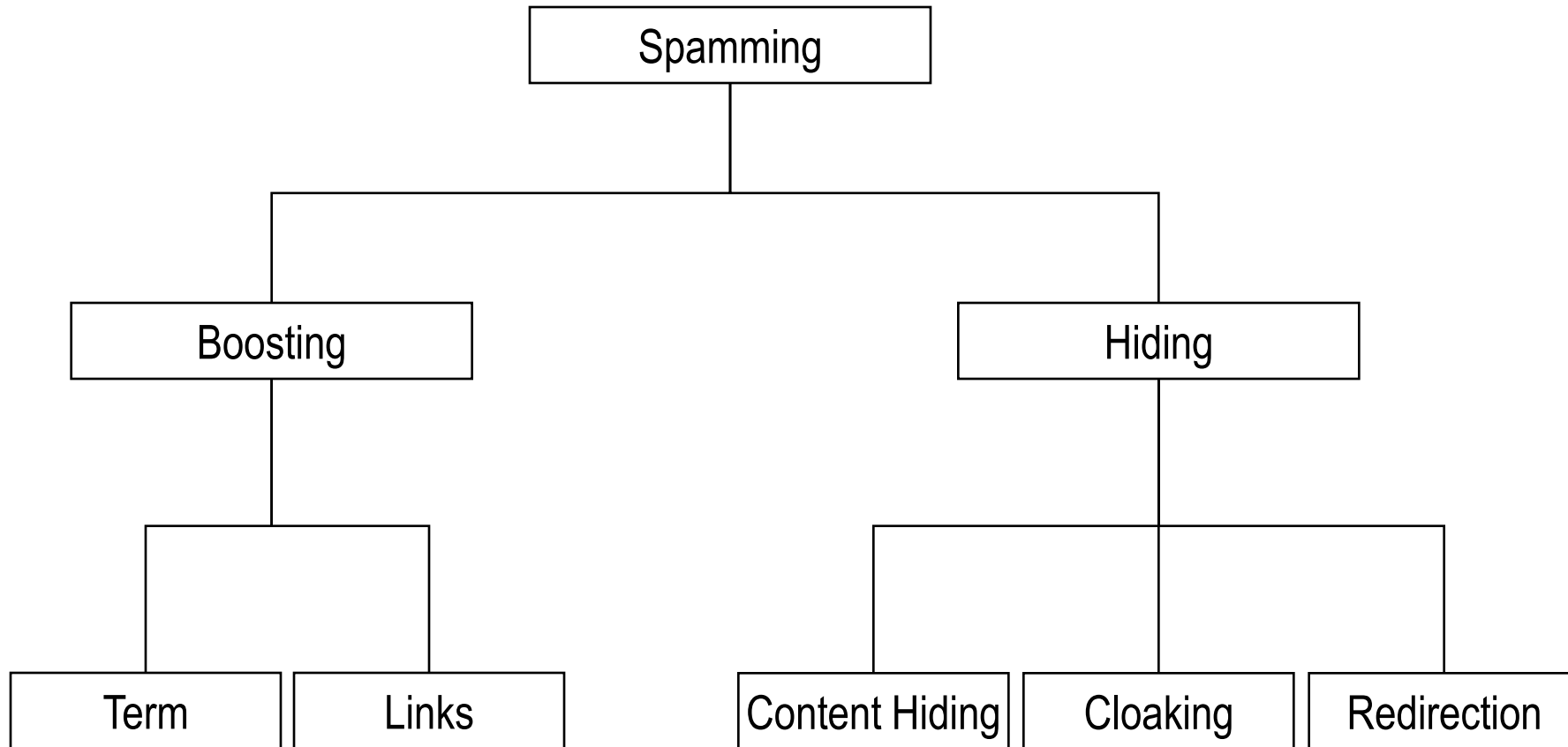
- Target of spammers
 - Not end users (directly)
 - High revenue from customers for search engine “optimization” (especially Google)
 - Indirect revenue
 - Affiliate programs, Google AdSense
 - Ad display, traffic funneling
- Spam taxonomy
 - Content spam
 - Keywords
 - Popular expressions
 - Mis-spellings
 - Link spam “farms”
 - Densely connected sites
 - Redirects
 - Cloaking and hiding
 - Spam in social media



Overview

Web Spam

Marc Spaniol





Spammed Ranking Elements

- Term frequency (tf in the tf.idf, Okapi BM25 etc. ranking schemes)
- Term frequency weighted by HTML elements
 - Title
 - Headers
 - Font size
 - Face
- Heaviest weight in ranking
 - URL, domain name part
 - Anchor text: `Best Saarbruecken nightlife`
- Structural information
 - URL length
 - Depth from server root
 - Indegree
 - PageRank
 - Link based centrality

⇒ All Web information retrieval ranking elements spammed



Content Spam

- Domain name

adjustableloanmortgagemastersonline.compay.dahannusaprima.co.uk

buy-canon-rebel-20d-lens-case.camerasx.com

- Anchor text (title, H1, etc)

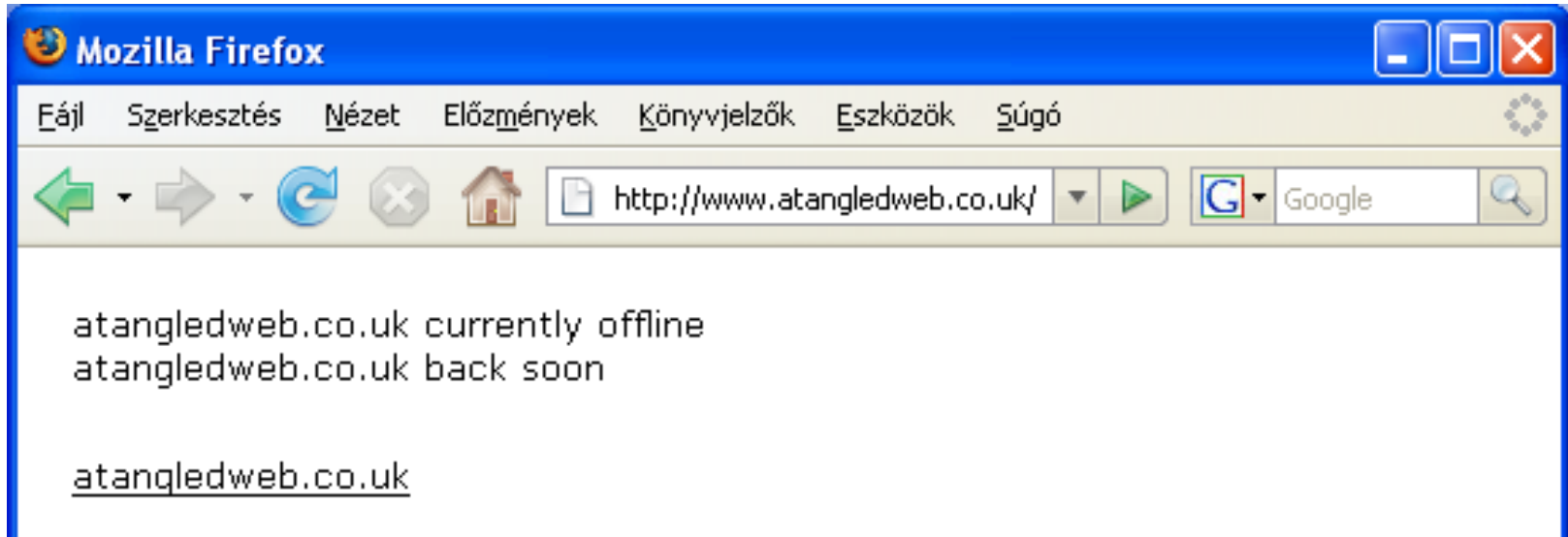
`free, great deals, cheap, inexpensive, cheap, free`

- Meta keywords

`<meta name="keywords" content="UK Swingers, UK, swingers, swinging, genuine, adult contacts, connect4fun, sex, ...">`



Parking Domain



```
<div style="position:absolute; top:20px; width:600px; height:90px; overflow:hidden;"><font size=-1>atangledweb.co.uk  
currently offline<br>atangledweb.co.uk back soon<br></font><br><br><a href="http://www.atangledweb.co.uk"><font  
size=-1>atangledweb.co.uk</font></a><br><br><br>
```

Soundbridge HomeMusic WiFi Media Play-...
SanDisk Sansa e250 - 2GB MP3 Player --...
AIGO F820+ 1GB Beach inspired MP3 Pla-...
Targus I-Pod Mini Sound Enhancer-...
Sony NWA806FP.CE7 4GB video WALKMAN -...
Ministry of Sound 512MB MP3 player-...
Nokia 6125 - Fold Design - 1.3 Megapi-...





Keyword Stuffing & Generated Copies

wrjk.frinzezz.net

Web Spam

belmajdoub

– From "Seductions of Rice" by Jeffrey Alford and Naomi Duguid (Artisan, \$24. Als erste 32 GB Karte wird sie dabei der Class 6 Geschwindigkeitsspezifikation genügen, die eine minimale Datenübertragungsrate von sechs MB/s bei einer leeren Karte vorsieht. It's pronounced incorrectly sometimes, but they know me. The Cospicua school has decided to use the Belgian and Scottish schools' approaches, which are entitled 'The Achievement Wall' and 'The Box of Feelings'. "It's more of the smaller stuff. I think it would be wise to not get in knee deep with ideas and plans once I have everything, in every room, cleaned and organized. In the turbulent days preceding the Spanish civil war, Lorca, who was living in Madrid, was uncertain whether or not to return home to Granada as he did each summer, unclear where he would be safest in the event of a Nationalist coup. "If it's a significant customer we can go quite upmarket - when you go down the bespoke route, it can be almost anything. 4 ranked Lady Mustangs (12 3, 2 1) beat Northside in three of the four meetings between the two last season. No wonder the Sena has asked BPOs across the city for details of security measures taken for female staff during night "Will

Marc Spaniol

article

[bon jovi crush tour dvd](#)
[megaupload](#)
[biphosphonates dialysis](#)
[descargar solucionario](#)
[tanenbaum](#)
[carla giraldo con sus posturas sexuales](#)
[epileren touw](#)
[construccion del teleton](#)
[tlalnepantla](#)
[feuerwehr gisingen](#)
[termine](#)
[concepto de pterigium](#)
[configuracion pagina con](#)





Google ads

admin-to-go.co.uk

Office and secretarial services

Welcome back!

Friday 25 April 2008



Looking for office and secretarial services?
Compare companies and solutions here

The following companies may be of interest to you . . .

1. Next Home Collection

Collection of Homeware at Next. Next day delivery and free returns.
[next.co.uk](http://www.next.co.uk)

2. Shopping

Looking for discount vouchers codes? Discount Code has 100's of free to use promo codes, discount codes and voucher code for many UK online shops. Get you voucher codes now.
www.discountcodes.co.uk

3. Home Shopping

Huge Range of Items From Top Brands Order Online & Get Free Delivery.
www.empirestores.co.uk

4. Additions Direct

All the latest fashion delivered to your door the next day for £3.134.
www.additionsdirect.co.uk

5. Cheap Products - UK

Buy any products at web prices with Kelkoo. Find Great deals.
www.kelkoo.co.uk

6. Spring 2008 Collection

admin-to-go.co.uk



Other suggested searches . . .

> [Car Hire Company](#)

> [Four W](#)

> [Buy New Car](#)

> [Car Par](#)

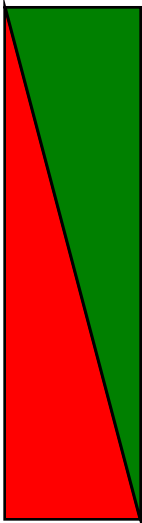


Link Spam

“Hyperlink structure contains an enormous amount of latent human annotation that can be extremely valuable for automatically inferring notions of authority.”

(Chakrabarti et. al. '99)

- Hyperlinks: Good, Bad, Ugly



Honest link, human annotation

No value of recommendation, e.g. “affiliate programs”, navigation, ads ...

Deliberate manipulation, link spam





PageRank

PageRank of page p_0 :

$$p_0 = c \sum_i p_i / |F(i)| + (1-c)$$

Outgoing links from p_i

damping factor

PageRank of p_i pointing to p_0

random jump

Generalized (vector):

$$p = c T^T p + \frac{(1-c)}{N} \mathbf{1}_N$$

Transition matrix

Score vector

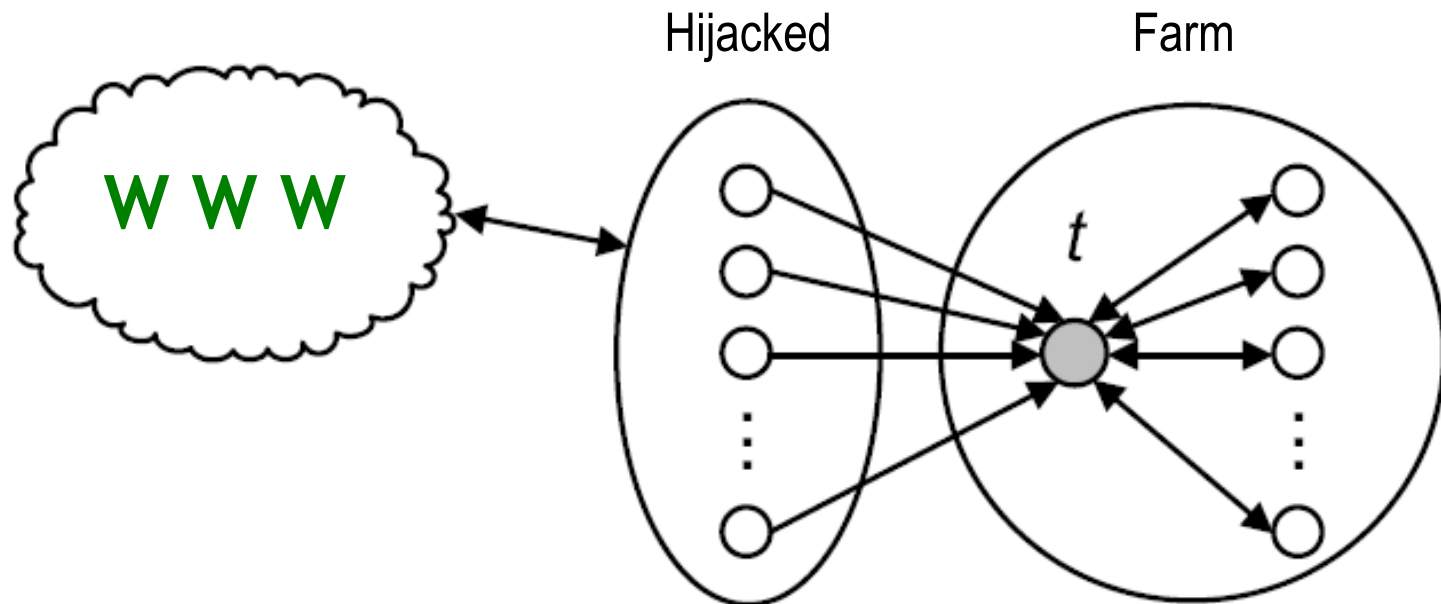
"1" vector

- One page is important if it is pointed to by many other pages
 - Based on the link structure
- ⇒ The algorithm of PageRank is vulnerable to link spamming



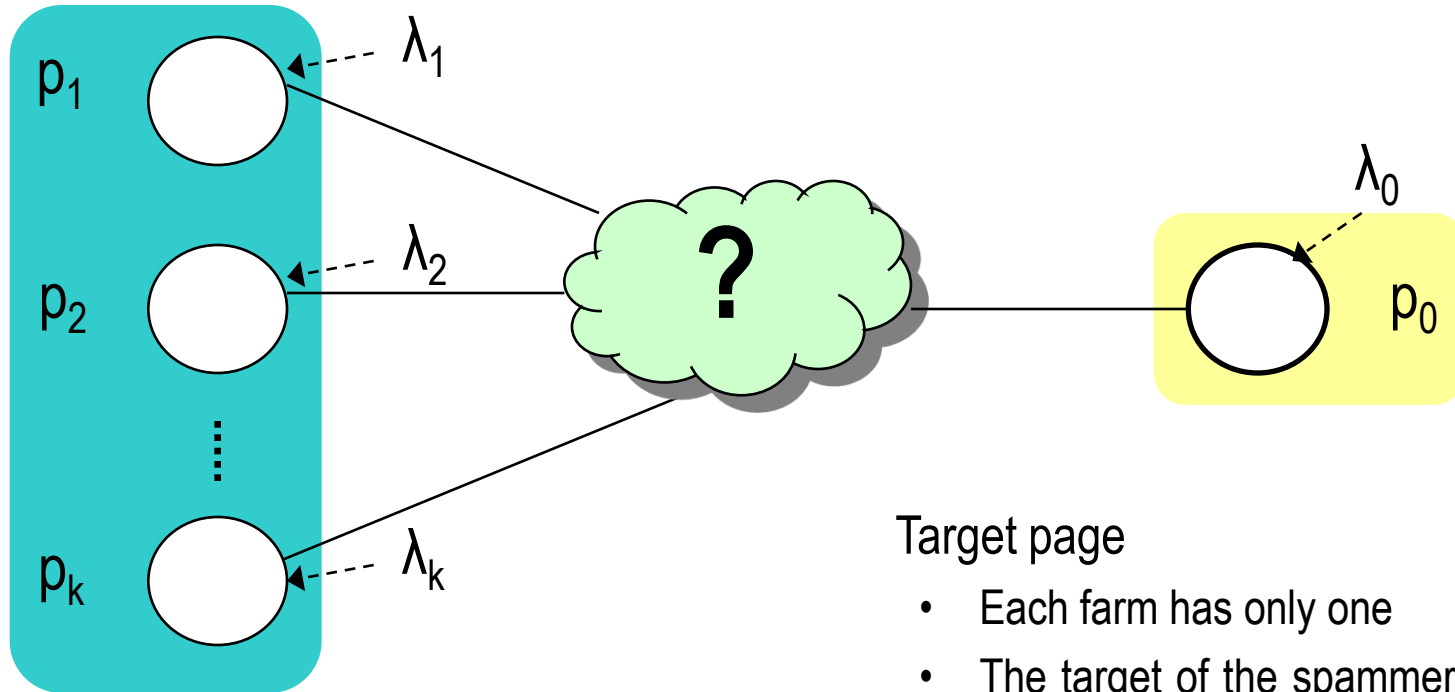
Link Farms

- Entry point from the honest web
 - Honey pots: Copies of quality content
 - Dead links to parking domain
 - Blog or guestbook comment spam





Spam Farm: Pages



Target page

- Each farm has only one
- The target of the spammer is to increase this page's ranking

Boosting pages

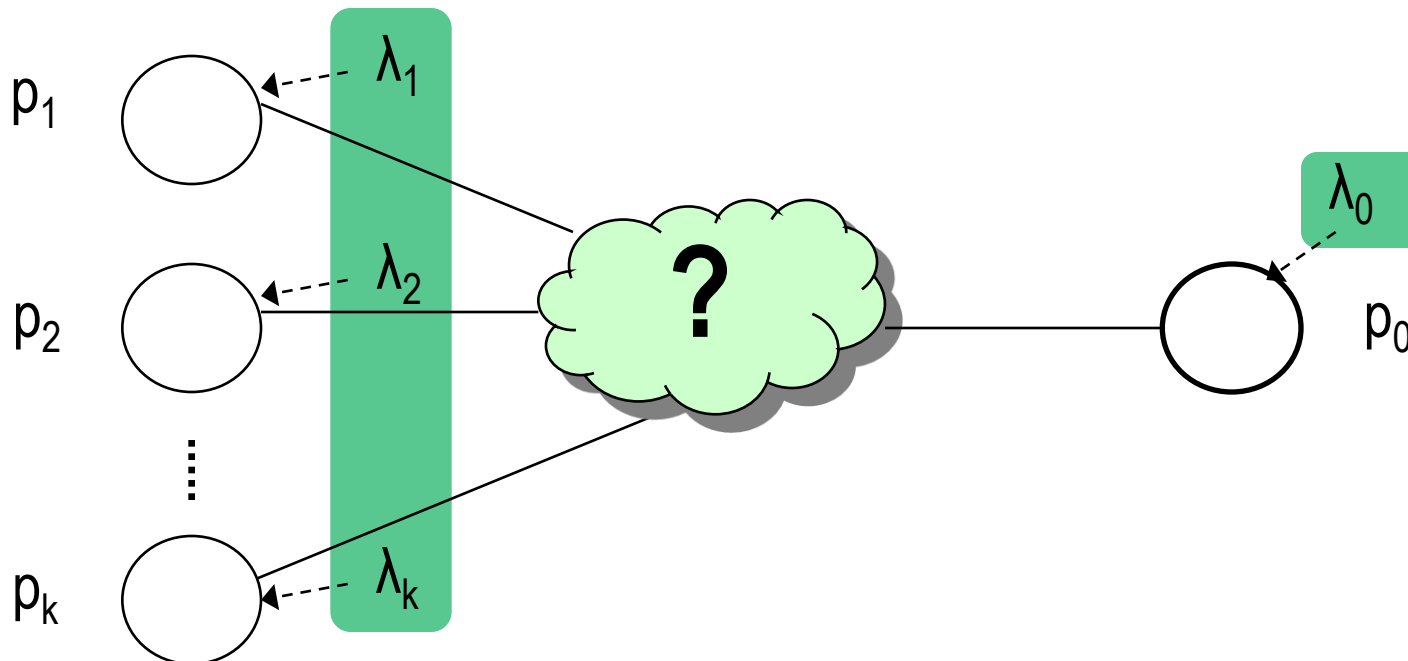
- Controlled by the spammer
- Pointing to the target page in order to increase its PageRank

Topics of interest on cheap

cheap airline discount tickets
cheap airline tickets online
cheap airline tickets
cheap airline tickets to london
cheap airline tickets canada
cheap airline tickets belgium
cheap and airline and tickets



Spam Farm: External Links



Web Spam

Marc Spaniol

Leakage

- Fractions of PageRank
- Link to the pages are added from pages outside the Farm (forum, blog, ...)
- The spammer has no or limited control on them
- $\lambda = \lambda_0 + \dots + \lambda_k$

#21 SergBin (SergBin[at]mymail.com)

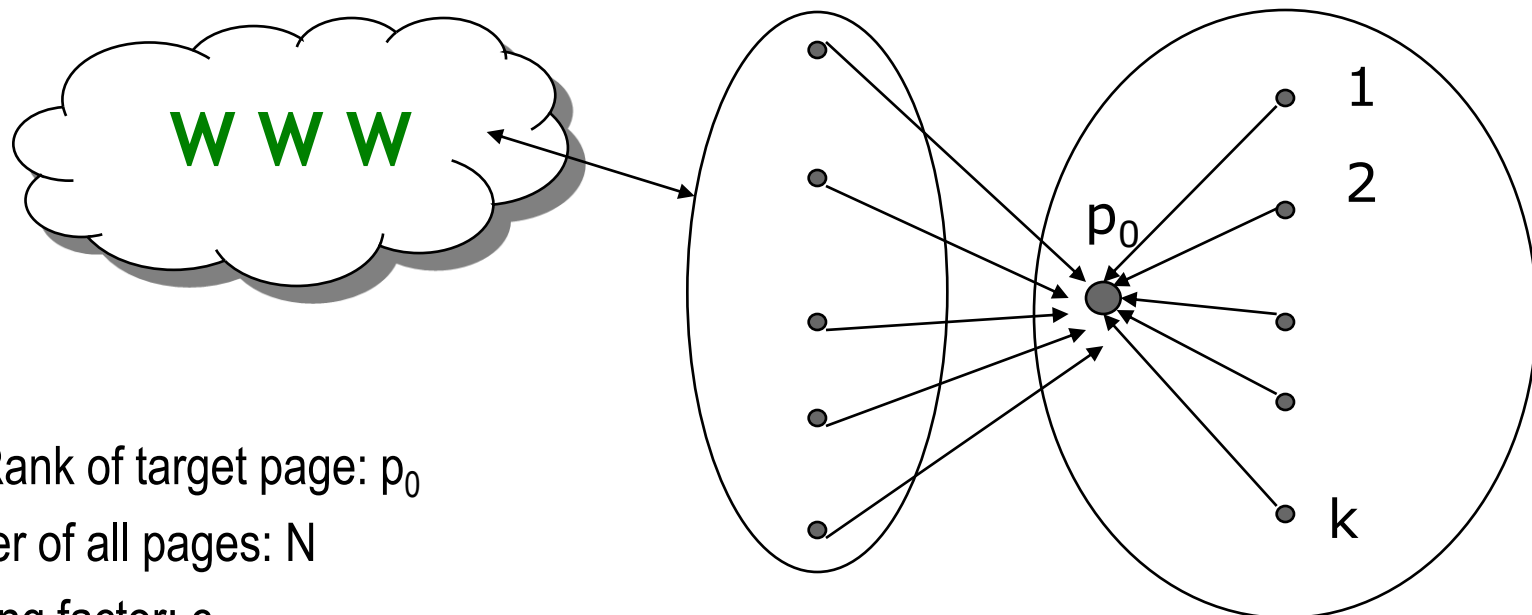
SergBin

Good stuff dude, thanks! <http://rik.tag-host.com>
<http://phenterminerx.cl.nu/> phentermine cod ph
<http://k.domaindx.com/rxsyst/> buy viagra purch
<http://messageboard694583.aimoo.com/> buy ph





Simple Farm Model



- PageRank of target page: p_0
- Number of all pages: N
- Damping factor: c
- Leakage contributed by accessible pages: λ
- PageRank of each farm page: $(1-c)/N$

$$p_0 = \lambda + k \cdot c \cdot [(1-c)/N] + (1-c)/N$$
$$= \lambda + [(1-c)(ck+1)]/N$$

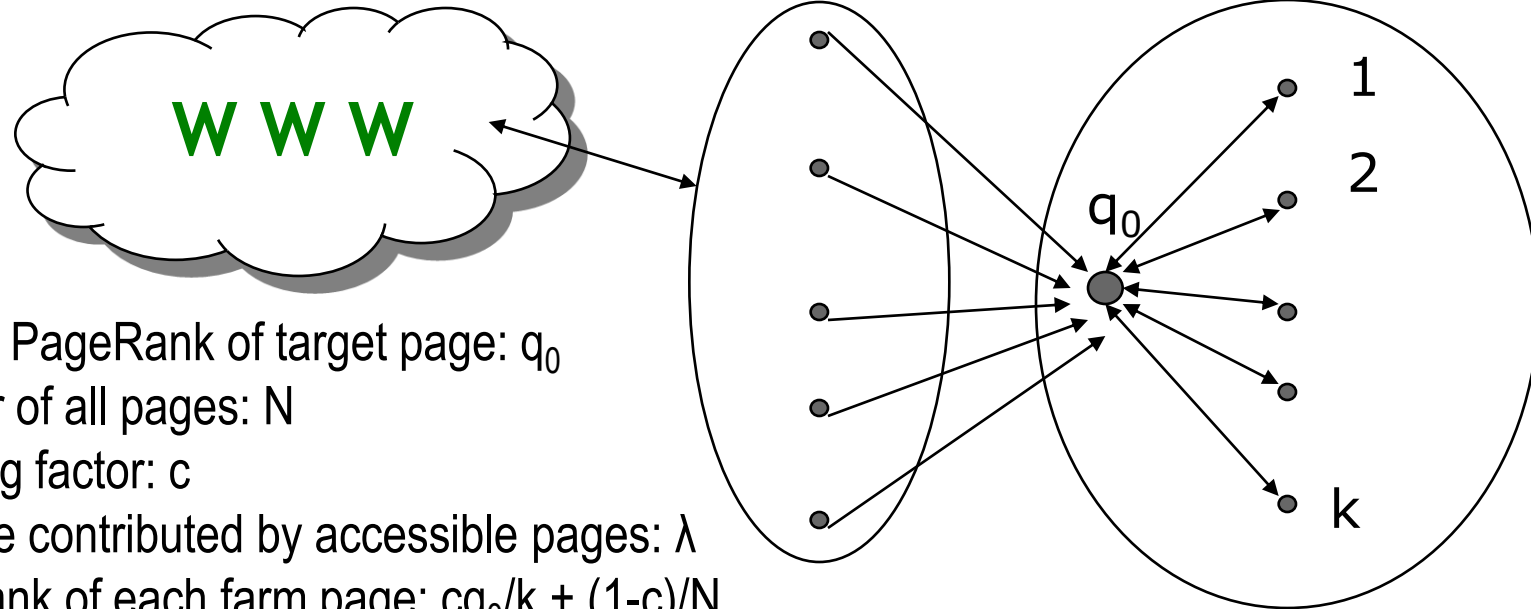
⇒ By making k large, we can make p_0 as large as we want

⇒ No multiplier effect for “acquired” page rank





Optimal Farm Model



- Optimal PageRank of target page: q_0
- Number of all pages: N
- Damping factor: c
- Leakage contributed by accessible pages: λ
- PageRank of each farm page: $cq_0/k + (1-c)/N$

$$q_0 = \lambda + ck[cq_0/k + (1-c)/N] + (1-c)/N$$

$$= \lambda + c^2q_0 + c(1-c)k/N + (1-c)/N$$

...

$$= \lambda/(1-c^2) + [(1-c)(ck+1)]/N(1-c^2)$$

$$= p_0/(1-c^2)$$

⇒ By making k large, we can make q_0 as large as we want

⇒ For $c = 0.85$ “performance” gain: $1/(1-c^2) = 3.6$

⇒ Multiplier effect for “acquired” page rank



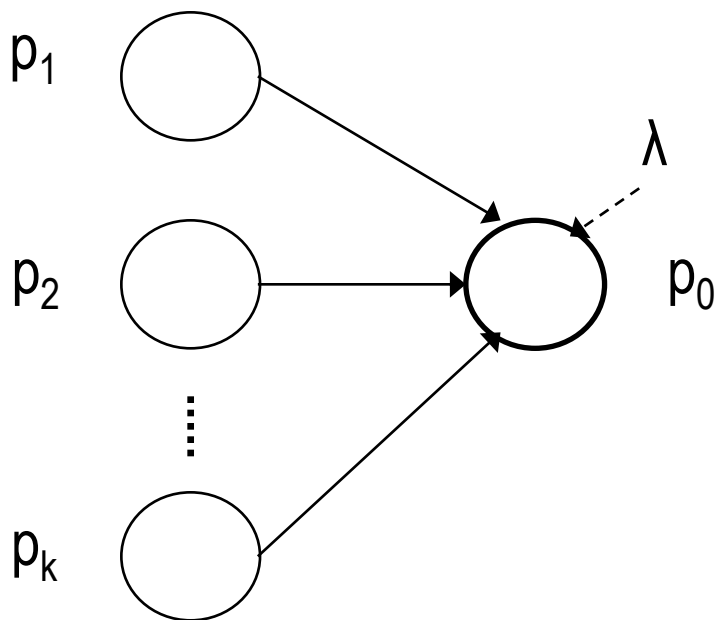


Simple vs. Optimal Farm

Simple:

Each boosting page only points to the target page

$$p_0 = c\lambda + \frac{(1-c)(ck+1)}{N}$$

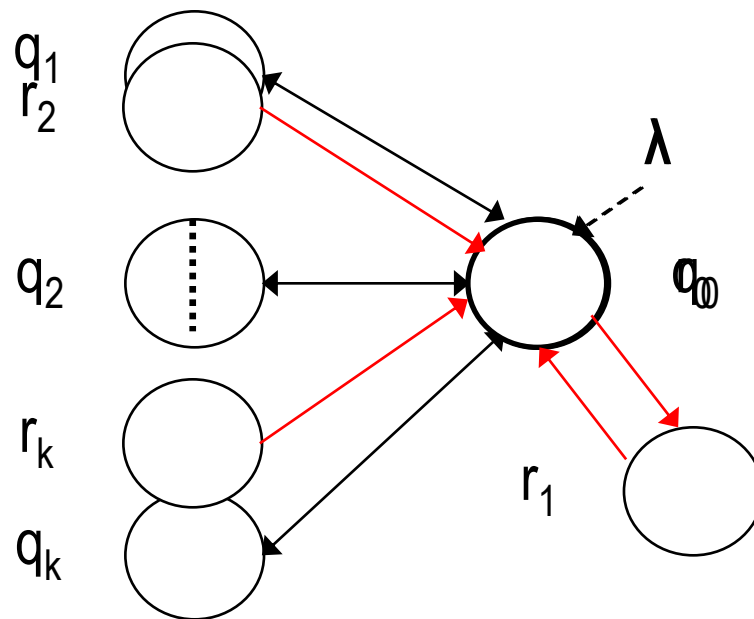


Optimal:

Shorter ref. paths to all boosting pages

There are also links among boosting pages

$$r_0 = p_0 / (1 - c^2)$$





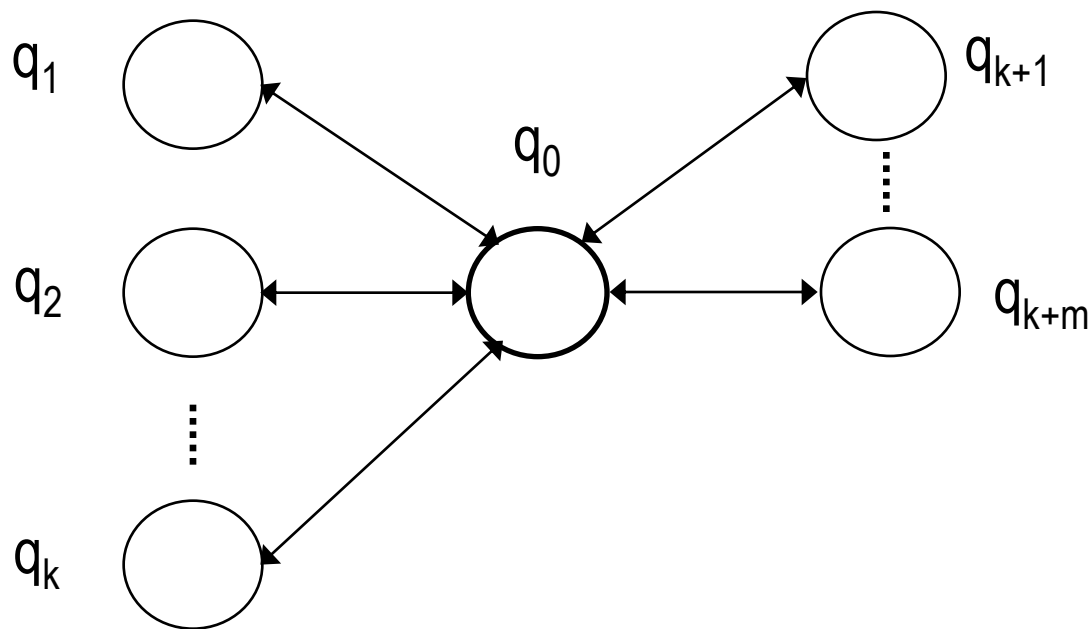
Optimality without Leakage

For mathematical simplification only

Idea: Interpret leakage as additional boosting pages

Web Spam

Marc Spaniol



$$\frac{c\lambda}{(1 - c^2)} + \frac{(1 - c)(ck + 1)}{(1 - c^2)N} \stackrel{!}{=} \frac{(1 - c)[c(k + m) + 1]}{(1 - c^2)N}$$

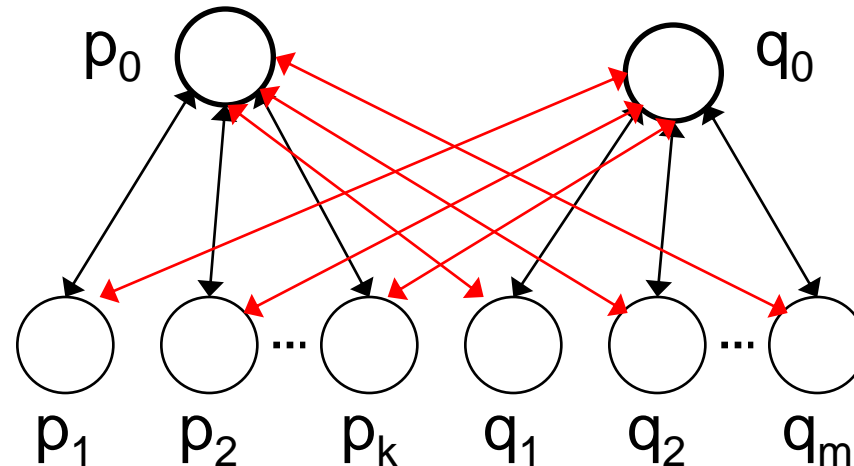




Alliance of two Farms

Intuitive:

Each boosting page points to both targets



$2(k + m)$ new links

$$p_0 = c \sum_{i=1, \dots, k} p_i / 2 + c \sum_{j=1, \dots, m} q_j / 2 + (1-c)/N$$

$$q_0 = c \sum_{i=1, \dots, k} p_i / 2 + c \sum_{j=1, \dots, m} q_j / 2 + (1-c)/N$$

$$p_i = c(p_0 + q_0) / (k + m) + (1-c)/N, \quad i = 1, \dots, k$$

$$q_j = c(p_0 + q_0) / (k + m) + (1-c)/N, \quad j = 1, \dots, m$$

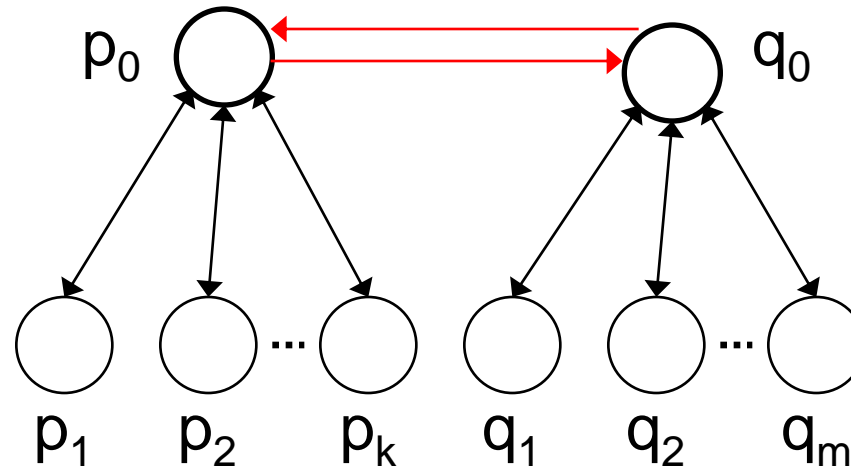




Alliance of two Farms

Better:

Only the target pages are interconnected with each other



only 2 new links

Redistribution of PageRank

$$p_0 = q_0 = \frac{c(k + m)/2 + 1}{(1 + c)N}$$

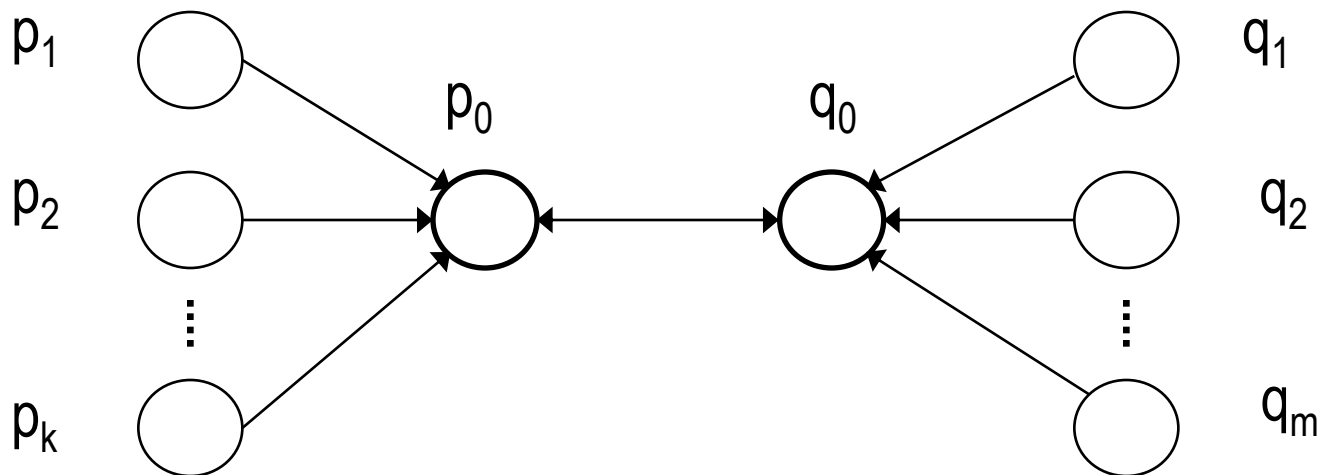
Convenient for the smaller Farm



Alliance between two Farms

Optimal:

Each target points to the other target
The targets have no links to the boosting pages



$$p_0 = c(\sum_{i=1, \dots, k} p_i + q_0) + (1-c)/N$$

$$q_0 = c(\sum_{j=1, \dots, m} q_j + p_0) + (1-c)/N$$

$$p_i = (1-c)/N, \quad i = 1, \dots, k$$

$$q_j = (1-c)/N, \quad j = 1, \dots, m$$



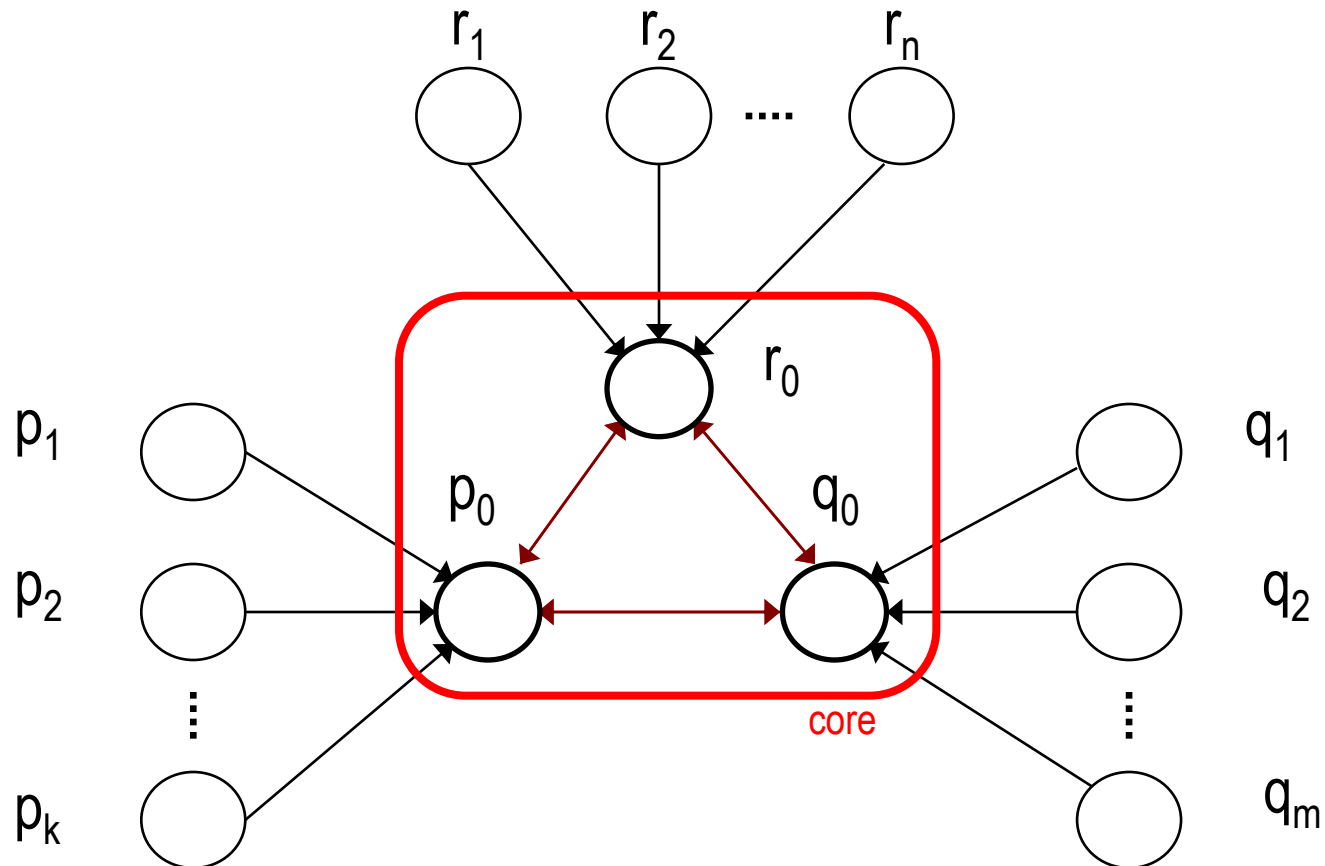


Multi-Farm Alliance

Two fundamental structures:

Web ring

Complete core





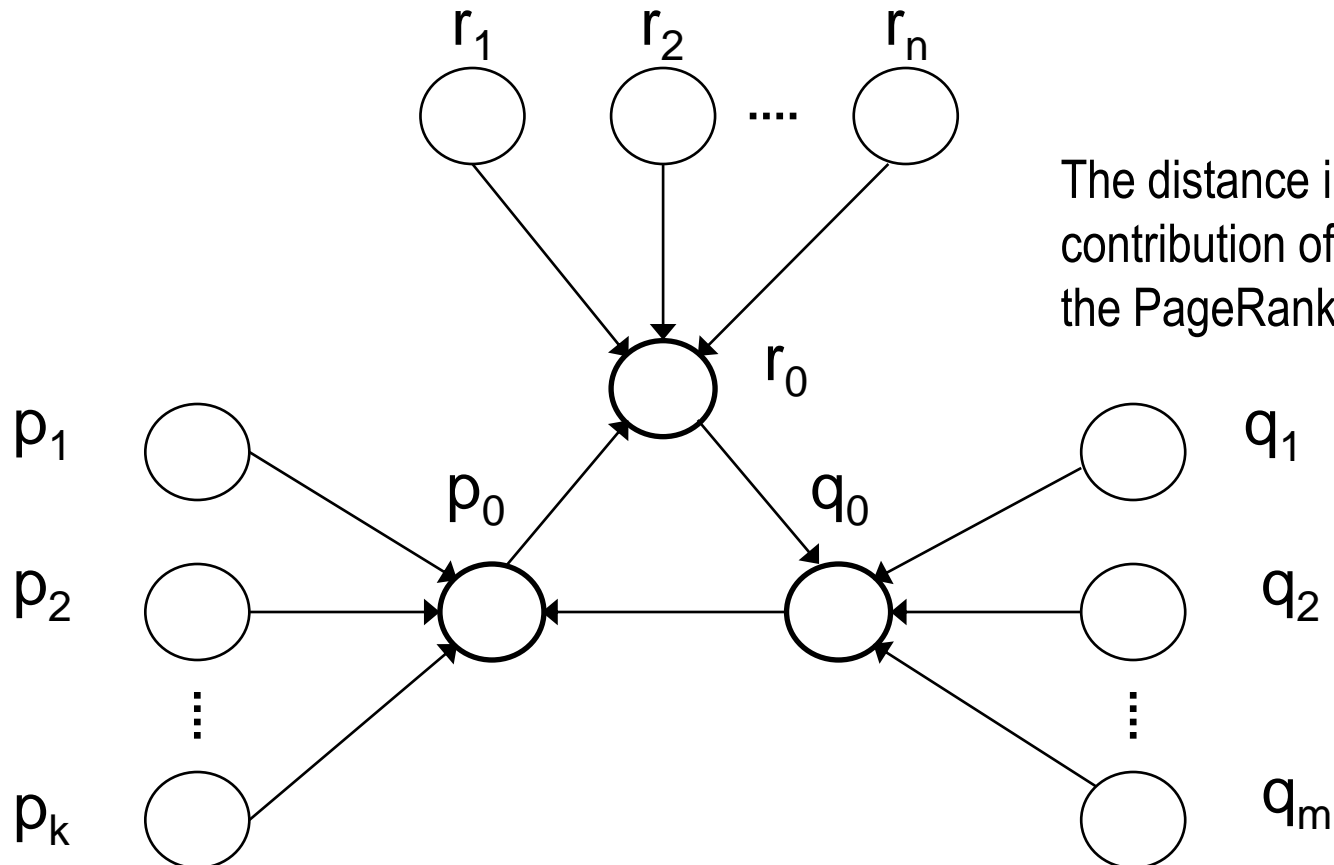
Web Ring

Simple and intuitive connection model

Web Spam

Marc Spaniol

$$p_0 = \frac{ck + c^2m + c^3n}{(1 + c + c^2)N} + \frac{1}{N}$$



The distance influences the contribution of each farm to the PageRank of the others





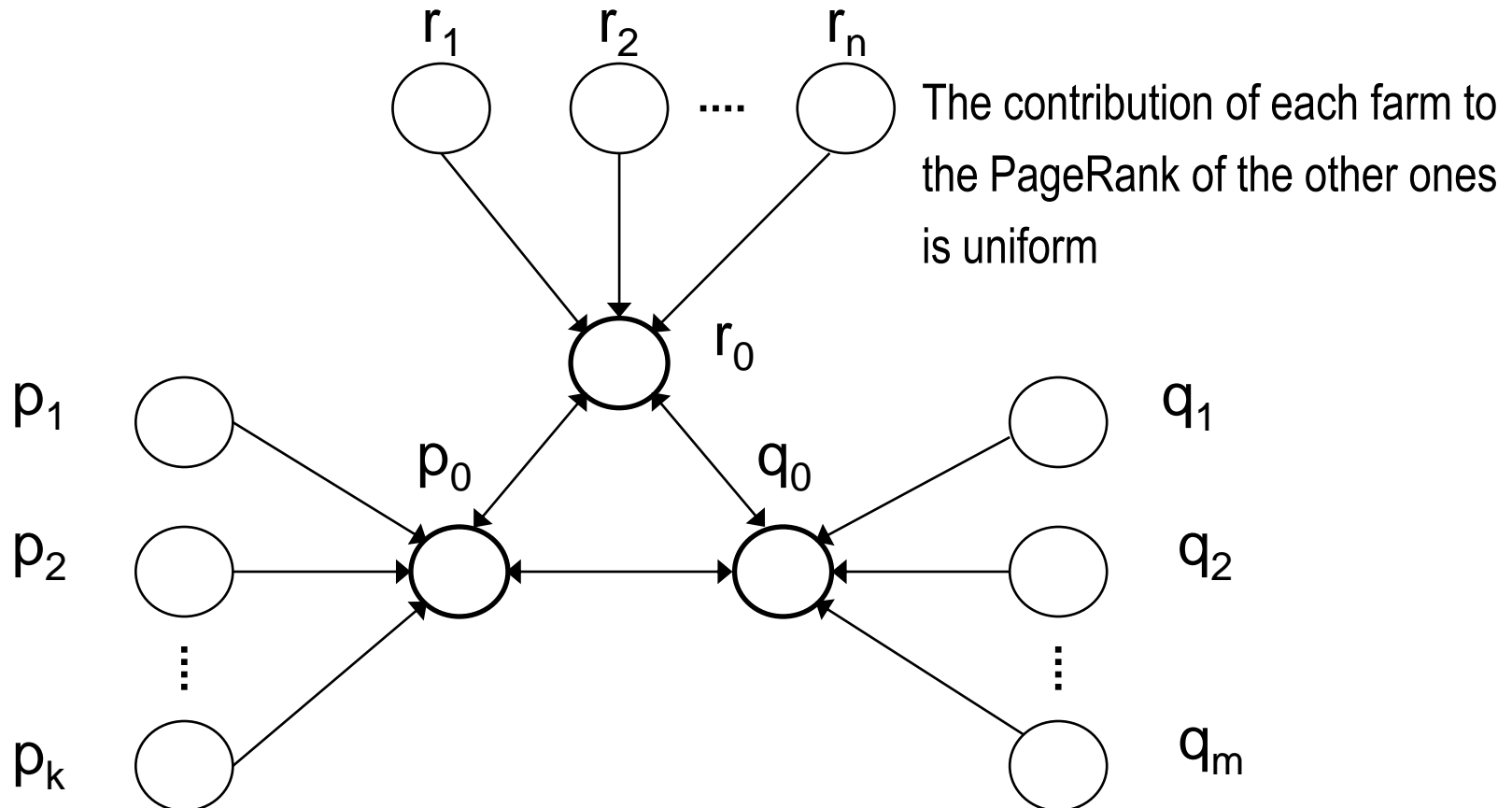
Complete Core

The core is a completely connected sub-graph

Web Spam

Marc Spaniol

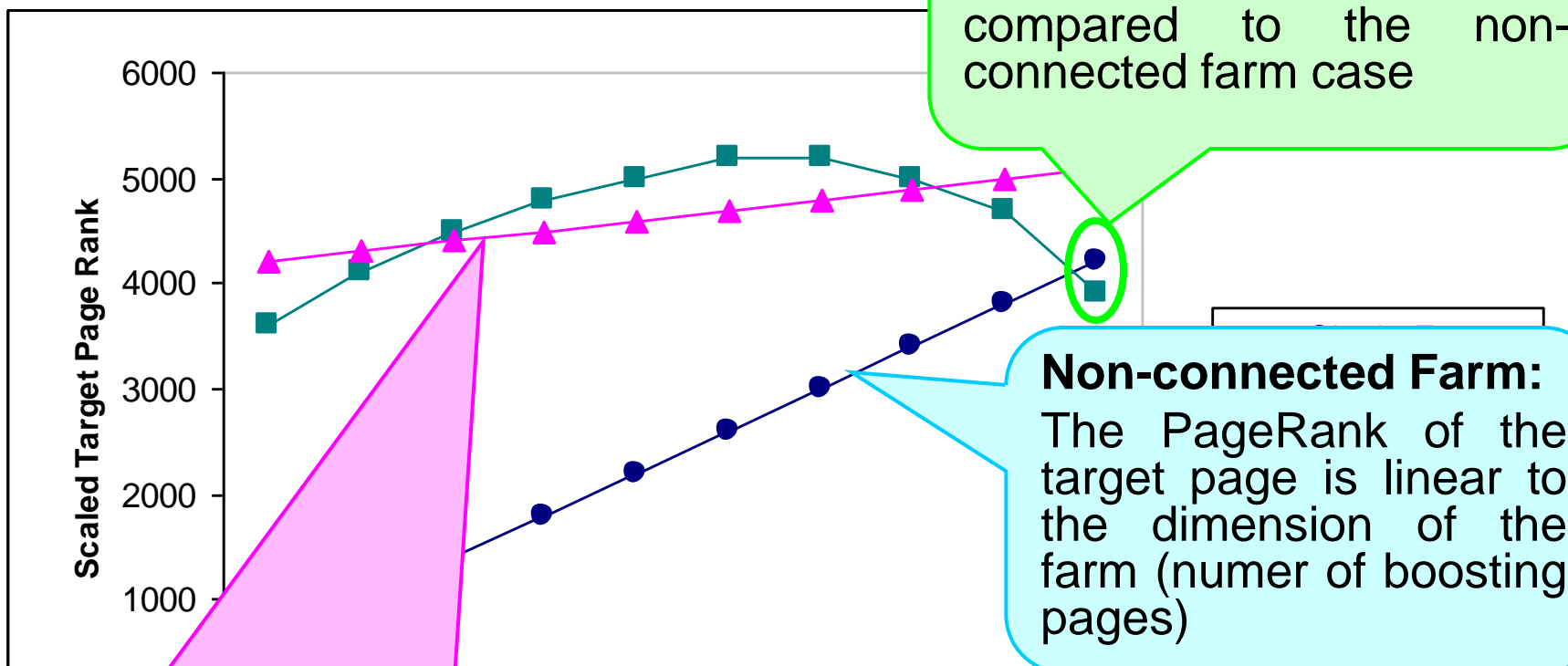
$$p_0 = \frac{2ck - c^2k + c^2m + c^2n}{(2 + c)N} + \frac{1}{N}$$





Evaluation

Target scores for ring/complete cores:
10 farms of sizes 1.000, 2.000, ... , 10.000



Web ring:

The PageRank of the target of farm 10 decreases compared to the non-connected farm case

Non-connected Farm:

The PageRank of the target page is linear to the dimension of the farm (number of boosting pages)

Complete core:

All PageRanks increase, especially those of the target of farms with low dimensions

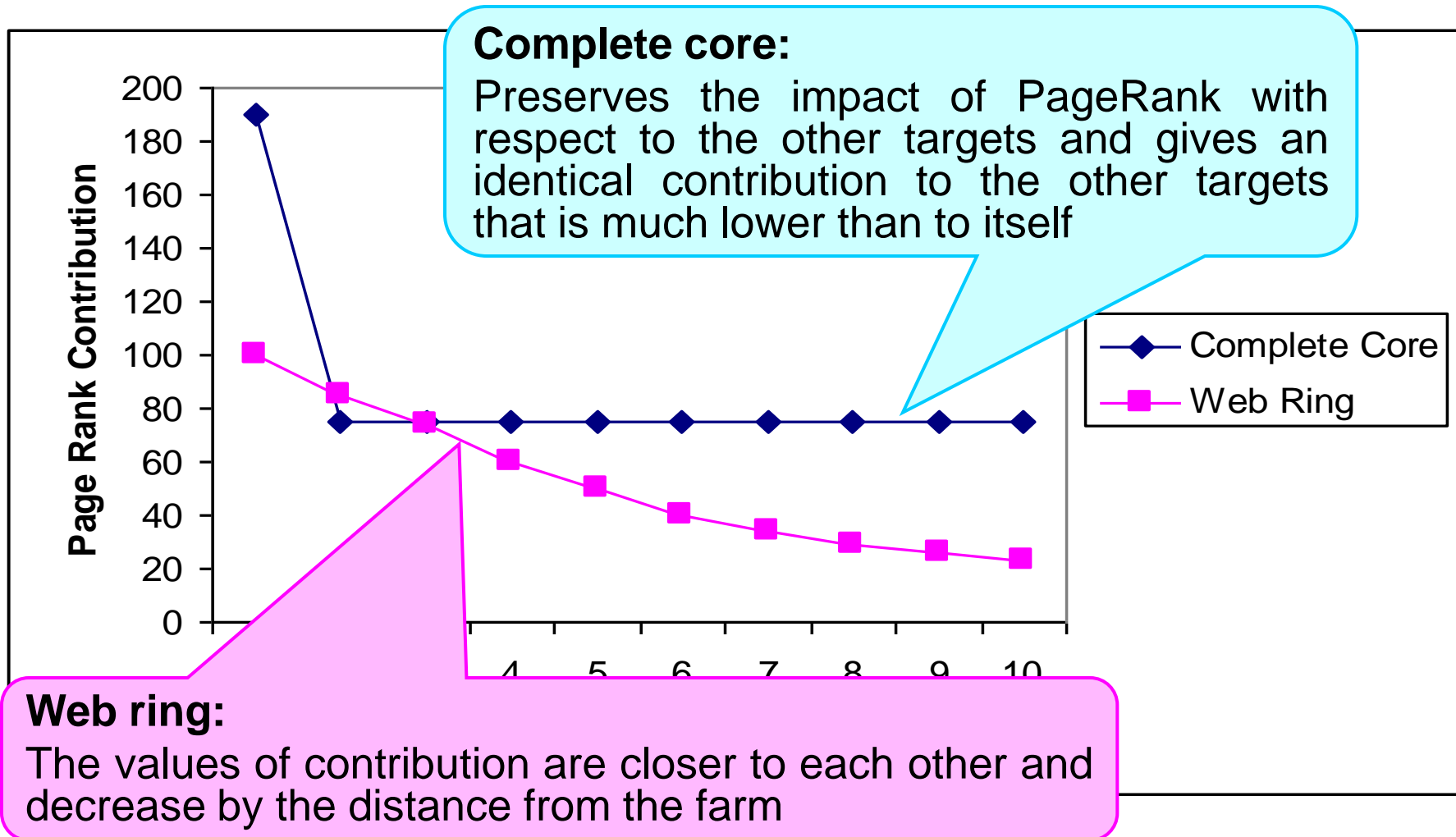


Evaluation

Contribution of farm 1 to the other targets

Web Spam

Marc Spaniol





Lessons Learned

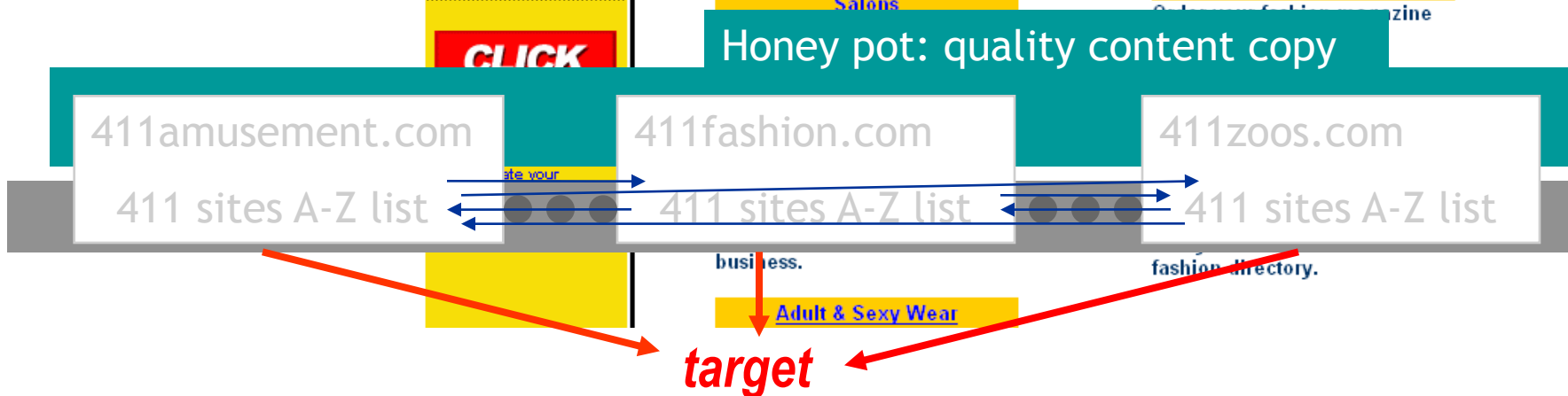
- Single farm
 - Short loop(s) increase target PageRank
 - Increase of PageRank is linear with the amount of boosting pages
 - Leakage should only point to the target page
- Leakage can be interpreted as an additional number of boosting pages
- Two farms:
 - Target pages should only link to other targets
 - In an alliance of two, both participants win
- Larger alliances
 - Need to be stable to keep all participants happy
 - Complete core topology:
Contribution to the PageRank of others at a relatively “low level”
 - Web ring topology
Contribution to the PageRank of others “slowly” decreasing by distance





- Multi-domain
- Multi-IP

Link Farms – Example





Cloaking and Hiding

- Formatting tricks: Trapping crawlers with simple HTML processing only

- One-pixel image

```
<a href= "target.html" ><img src= "tinyimg.gif" ></a>
```

- White over white

```
<body background= "white" >  
  <font color= "white" >hidden text</font>
```

...

```
</body>
```

- Color, position from stylesheet

- Redirection

- Script
- Meta-tag with refresh time 0

- ...



Obfuscated JavaScript

```
<SCRIPT language=javascript>
```

```
var1=100;
```

```
var3=200;
```

```
var2=var1 + var3;
```

```
var4=var1;
```

```
var5=var4 + var3;
```

```
if(var2==var5)
```

```
document.location="http://umlander.info/mega/free_software_downloads.html";
```

```
</SCRIPT>
```

- More sophisticated tricks

- Redirection through window.location
- Spam content (text, link) from random looking static data via eval calls
- Content generation by document.write



HTTP Level Cloaking

- User agent, client host filtering

```
GET /db_pages/members.html HTTP/1.0
```

```
Host: www-db.stanford.edu
```

```
User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
```

- Different for users and for GoogleBot
- “Collaboration service” of spammers for
 - Crawler IPs
 - Agents
 - Behavior



Spam in Social Media

Guest books

Гостевая Книга Guestbook

Спасибо, что посетили мою страницу. Вы можете оставить запись в моей [Гостевой Книге](#).
Thank you for visiting our pages. We would love it if you would [Add](#).

Enjoyed your website and found it informative. [url=http://nazar.onlyhot.info/russell-grant-horoscope/]russell grant horoscope[/url]
[John en Lia Maan](#) <buka_sm@yahoo.com>Miami , USA - Monday, April 3, 2006 at 21:34:58

phentermine
hydrocodone
ханax

[ханax](#) <@size>Москва, Россия - Monday, April 3, 2006 at 21:17:19

Enjoyed your website and found it informative. [url=http://meds.onlyhot.info/russell-grant-horoscope/]russell grant horoscope[/url]
[Rosina May](#) <sigmroni@hotmail.com>Denver, USA - Monday, April 3, 2006 at 20:37:47

I like it because is very useful. [url=http://top.onlyhot.info/russell-grant-horoscope/]russell grant horoscope[/url]
[Jurg Bollinger](#) <annelies.hesp@wanadoo.nl>Memphis, USA - Monday, April 3, 2006 at 19:56:12

Thank you for your site. I have found here much useful information...

[hoodia patch](#) Boston, USA - Monday, April 3, 2006 at 19:30:34

uggs
phentermine
cialis
carisoprodol
fioricet
ambien
-



Fake Blogs

Political Concepts

A Working Paper Series of the Committee on Concepts and Method

Working Paper

Svend-Erik Skaaning, "Measuring Civil Liberty"

April 2008

Comments

[viagra doses prices com net org](#)

21 April 2008

Nice site. Thank you!! [viagra doses prices com net org](#)

[Lane](#)

21 April 2008

Well done! [roulette games online](#) | [fun play slots](#) | [no download online free slots](#) | [free play online no deposit bonus](#) | [cleopatra slot](#) | [online slot game](#) | [free slot machines to play online slot machine](#)





Spam Detection

- Crawl-time vs. post-processing
- Simple filters in crawler
 - Cannot handle unseen sites
 - Require large bootstrap crawl
 - Need to run rendering and script execution
- Crawl time feature generation and classification
 - Needs interface in crawler to access content
 - Needs model from bootstrap or external crawl (may be smaller)
 - Sounds expensive but needs to be done only once per site
- The hard work is done post-processing both cases



Assessment Interface and Collaboration Infrastructure

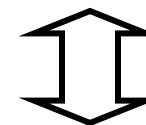
Web Spam

Local storages

May share features, extracts across institutions

“Interaction”
Active learning

Marc Spaniol



Docs
(WARC)

access

Feature generation
(crawl-time)

feature feed
text files

Classifier

- Build model
- Apply model (crawl-time)





Spam Labeling

- Manual labels (black AND white lists) primarily determine quality
- Can blacklist only a tiny fraction
 - Recall 10% of sites are spam
 - Needs machine learning
- Central to the service
 - Aid manual assessment
 - Aid information and label sharing
 - Catch spam farms that span different top-level domains

⇒ No free lunch: No fully automatic filtering



Web Spam Interface

Web Spam



Assessment interface

User: Gideon > [Review](#)

www.liwa-project.eu

- normal
- borderline
- spam
- don't know

Comment:

[Next](#) [Help](#) [Back](#)



<http://tagesschau.de>

The screenshot shows the homepage of tagesschau.de, a German news website. The header includes the site name and the slogan "Die Nachrichten der ARD". Navigation tabs for "ARD Home", "Nachrichten", "Sport", "Börse", "Ratgeber", "Wissen", "Kultur", "Kinder", "Fernsehen", "Radio", "ARD Mediathek", and "ARD Intern" are visible. The main content area features a large image of a hand holding a stethoscope, with the headline "Streit um Hausarztmodelle" and sub-headline "In der Praxis nur Ärger". Below this is a news item about a massive order backlog in the machine building industry, titled "Die Talsohle ist noch nicht erreicht". The right sidebar contains a search box, a video player for a report on the EU Council Presidency, and a "TV-Tipp" section for "ARD EXCLUSIV". The left sidebar lists various categories like "Startseite", "Inland", "Ausland", "Wirtschaft", "Regional", "Wetter", "Multimedia", "Weltatlas", "Info-Services", "Forum", "Blog", "News in English", and "Haberler".



Combating Spam

- Refresh detection
 - Conceal crawling by
 - Headers: Browser vs. crawler
 - Access: "Random" vs. BFS
 - TrustRank method
 - Supervised learning features
 - Number of words in the pages
 - Average word length
 - Number of words in the page title
 - Fraction of visible content
 - Amount of anchor text
 - Compressibility
 - Partition the Web into different blocks
- ⇒ Never stop! On-going process





Google AdWords Competition

10k
10th wedding anniversary
128mb, 1950s, ...
abc, abercrombie, ...
b2b, baby, bad credit, ...
digital camera
earn big money, easy, ...
f1, family, flower, fantasy
gameboy, gates, girl, ...
hair, harry potter, ...
ibiza, import car, ...
james bond, janet jackson
karate, konica, kostenlose
ladies, lesbian, lingerie, ...
...



Query Marketability

Navigation icons: back, forward, refresh, close, home, search.

Address bar: <https://adwords.google.com/select/K>

Keywords related to **conference** - sorted by relevance [?]

<u>Keywords</u>	<u>April Search Volume</u> [?]	<u>Advertiser Competition</u> [?]	Match Type: [?] Broad <input type="button" value="v"/>
conference meeting	<input type="checkbox"/>	<input type="checkbox"/>	Add »
conference proceedings	<input type="checkbox"/>	<input type="checkbox"/>	Add »
conference exhibit	<input type="checkbox"/>	<input type="checkbox"/>	Add »
conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
europa conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
conference speakers	<input type="checkbox"/>	<input type="checkbox"/>	Add »
annual conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
conference recording	<input type="checkbox"/>	<input type="checkbox"/>	Add »
record conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
investment conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
conferences	<input type="checkbox"/>	<input type="checkbox"/>	Add »
banff conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
investor conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »
privacy conference	<input type="checkbox"/>	<input type="checkbox"/>	Add »



Generative Content Models

Web Spam

<i>Spam topic 7</i>
loan (0.080)
unsecured (0.026)
credit (0.024)
home (0.022)

Marc Spaniol

<i>Honest topic 4</i>	<i>Honest topic 10</i>
club (0.035)	music (0.022)
team (0.012)	band (0.012)
league (0.009)	film (0.011)
win (0.009)	festival (0.009)

Excerpt: 20 spam and 50 honest topic models

[Bíró, Szabó, Benczúr 2008]





TrustRank

- Basic idea: Approximate isolation
 - Honest pages rarely point to spam
 - Spam cites many, many spam
 - Sample a set of “seed pages” from the web
 - “Oracle” (human) identifies good and spam pages in the seed set
 - The subset of seed pages that are identified as “good” are called “trusted pages”
- ⇒ Expensive! Make seed set as small as possible



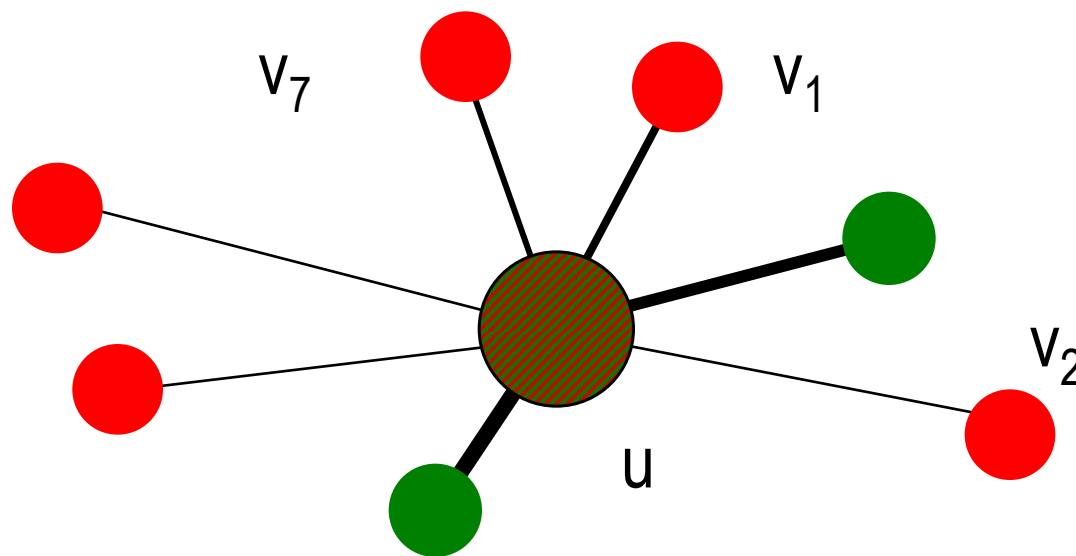
Trust Propagation

- Set trust of each trusted page to 1
- Propagate trust through links
 - Each page gets a trust value between 0 and 1
 - Use a threshold value and mark all pages below the trust threshold as spam
- Trust attenuation
 - The degree of trust conferred by a trusted page decreases with distance
- Trust splitting
 - The larger the number of outlinks from a page, the less scrutiny the page author gives each outlink
 - Trust is “split” across outlinks



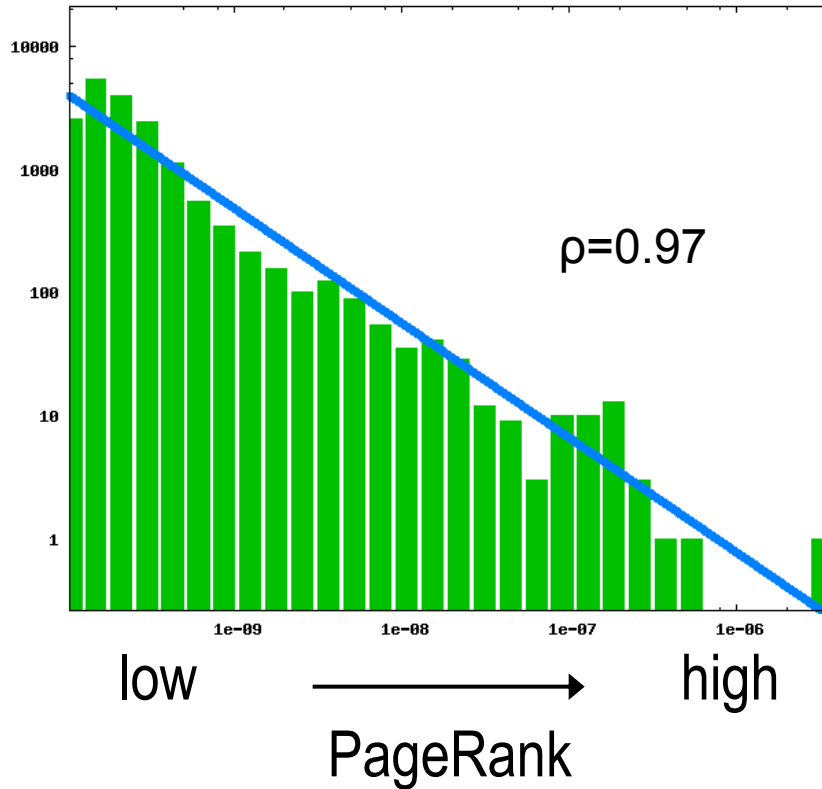
Trust Computation

1. Predicted spamicity
 $p(v)$ for all pages
2. Target page u ,
new feature $f(u)$
by neighbor $p(v)$
aggregation
3. Reclassification by
adding the new feature

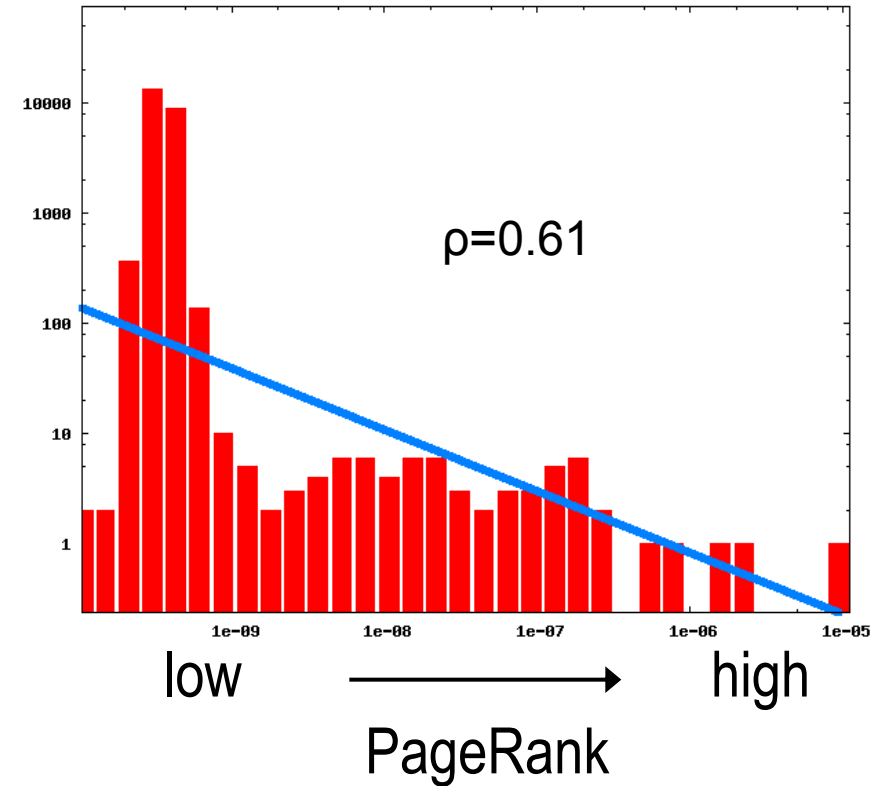




PageRank Supporter Distribution



Honest:
fhh.hamburg.de



Spam:
radiopr.bildflirt.de
(part of www.popdata.de farm)

[Benczúr, Csalogány, Sarlós, Uher 2005]





Web Spam Challenge

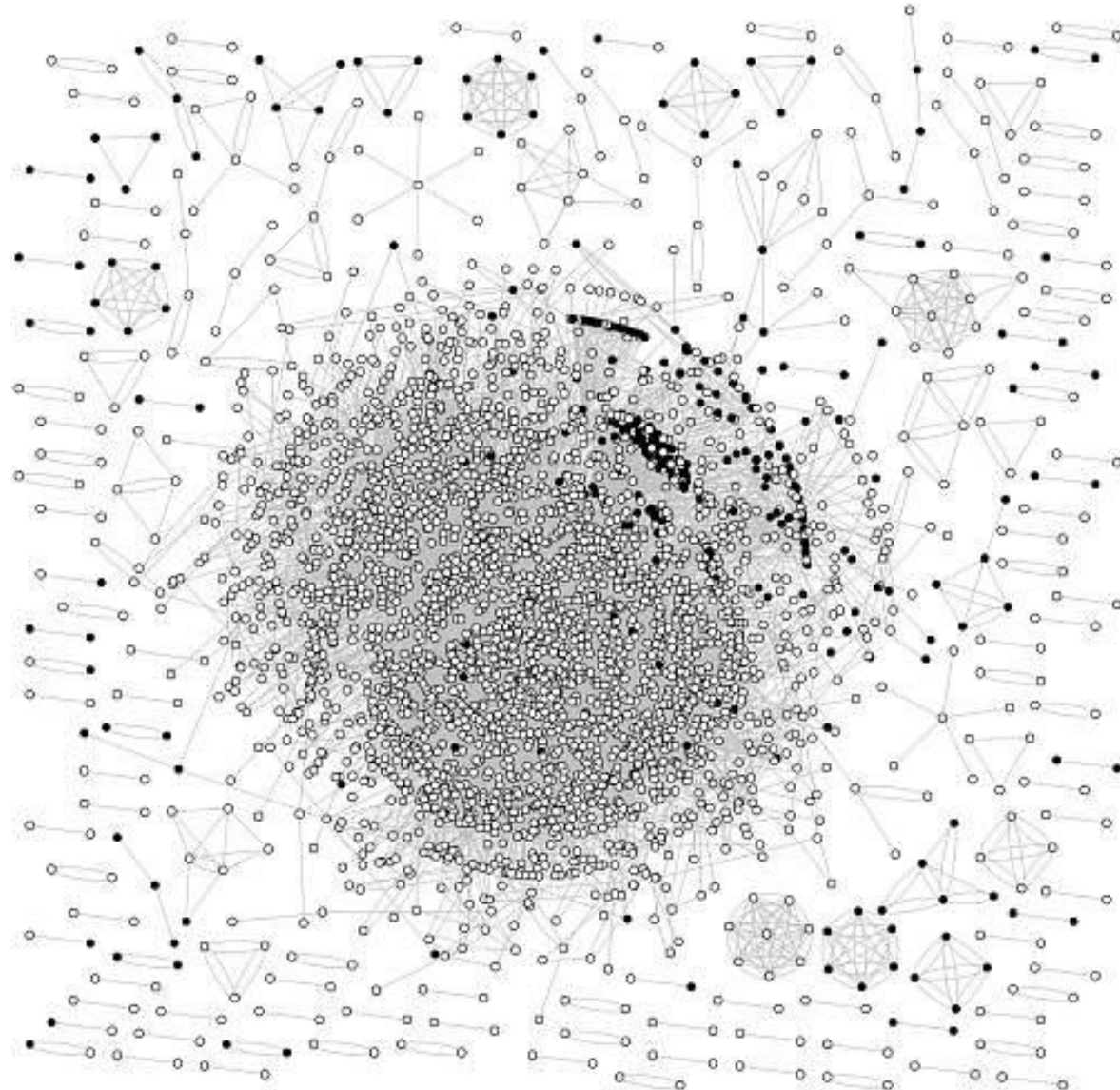
- **WEBSPAM-UK2006**

- 77M pages
- 11,402 hosts
- 7,373 labeled
- 26% spam

- **WEBSPAM-UK2007**

- 100M pages
- 114,529 hosts
- 6,479 labeled
- 6% spam

[WebSpam08]





Summary

- Web Spam
 - Aims at search engine optimization
 - “Attacks” indexing crawlers
 - Slows down archiving crawlers
 - “Pollutes” the archive
 - Social Web is a “threat”
 - Spamming techniques
 - Boosting
 - Hiding
 - Combinations of various techniques
 - Countermeasures
 - Manual assessment
 - Machine learning techniques
 - Hybrid approaches
- ⇒ All Web information retrieval ranking elements spammed



References

- [Benc08] A. Benczur: “Web spam survey for the Archivist”. 8th International Web Archiving Workshop (IWAW 2008), Århus, Denmark, Sept. 18, 2008.
<http://liwa-project.eu/index.php/video/33/>
[last access: July 22, 2009]
- [BSS*08] A. Benczur, D. Siklosi, J. Szabo et al.: “Web spam survey for the Archivist”. Proceedings of the 8th International Web Archiving Workshop (IWAW 2008), Århus, Denmark, Sept. 18, 2008.
<http://iwaw.net/08/IWAW2008-Benczur.pdf>
[last access: July 22, 2009]
- [GyGa05] Z. Gyöngyi and H. Garcia-Molina: “Link Spam Alliances”. Technical Report, Stanford University, March 2, 2005.
<http://www-db.stanford.edu/~zoltan/publications/gyongyi2005link.pdf>
[last access: July 22, 2009]
- [WebSpam08] Web Spam Challenge: “Home page”. 2008.
<http://webspam.lip6.fr/wiki/pmwiki.php>
[last access: July 22, 2009]

