

YAHOO!



Link Spam Detection Based on Mass Estimation

Zoltán Gyöngyi, Pavel Berkhin,
Hector Garcia-Molina, Jan Pedersen

Roadmap

- Web spamming
- Link spamming
- PageRank contribution
- Spam mass
 - Definition
 - Estimation
 - Algorithm
- Experiments

Web Spamming: Example

#3 search result for the query “kaiser pharmacy online”



The screenshot shows the homepage of UNCOMMONLY CLASSICAL.COM. The website has a green and white color scheme. At the top, there is a logo featuring a stylized figure holding a musical instrument, with the text "UNCOMMONLY CLASSICAL.COM" next to it. Below the logo, the website name "UNCOMMONLY CLASSICAL.COM" is displayed in a large, bold, orange font, with the tagline "ONLINE SHOPPING EXTRAVEGANZA!" underneath. The main content area is divided into several sections: a "NAVIGATE" sidebar with a list of categories (Lawyers, Loans, Magazines, Mortgage, Nurse, Pharmacy, Posters And Crafts, Ringtones, Shoes, Stock Market, Sunglasses, Theater, Viagra, Video Games), a "HOT PRODUCTS" section featuring a "ZERO DOWNTIME" advertisement for a website hosting service, and an "ABOUT US" section. The right side of the page contains a green banner with the text "GO SHOPPING NOW!" and a section titled "TAKE A LOOK AT SOME OF OUR TOP SELLERS!" which lists several products and services, including "Compounded Medications", "Pharmacist-CA", "Pharmacy", and "Online Pharmacy". The bottom of the page features an "ARTICLES" section with a paragraph of text about pharmacy.

UNCOMMONLY CLASSICAL.COM
ONLINE SHOPPING EXTRAVEGANZA!

NAVIGATE

- Lawyers
- Loans
- Magazines
- Mortgage
- Nurse
- Pharmacy
- Posters And Crafts
- Ringtones
- Shoes
- Stock Market
- Sunglasses
- Theater
- Viagra
- Video Games

HOT PRODUCTS

ZERO DOWNTIME
Host your website on 5 machines!
\$7.77 PER MONTH

ABOUT US

WHEN PEOPLE THINK ABOUT SHOPPING, THEY THINK OF GOING TO A STORE, BUT HOW ABOUT IF THE STORE CAME TO YOU! THAT'S RIGHT, THROUGH SOMETHING WE LIKE TO CALL THE "INTERNET", WE HAVE REACHED YOU IN YOUR HOME!

GO SHOPPING NOW!

TAKE A LOOK AT SOME OF OUR TOP SELLERS!

Ads by Goooooogle

Compounded Medications
Quality drugs, nationwide delivery Good prices, great customer service
McGuffCompoundingPharmacy.com

Pharmacist-CA
We recruit pharmacists for CA jobs. We represent the top CA employers.
www.pharmacist-ca.com

Pharmacy
Safe, Quality, Prescription Drugs Licensed Canadian Pharmacy
www.CanPharm.com

Online Pharmacy
Save on thousands of drugs. Local Pharmacy quality drugs
www.usamedsonline.com

ARTICLES

If you're looking for PHARMACY, **Click Here!**
Pharmacy is the profession of compounding and dispensing medication. More recently, the term has come to include other services related to patient care including clinical practice, medication review, drug information, etc. Some of these new roles are now mandated by law in various legislatures. Pharmacists, therefore, are the primary health professionals who optimise

Web Spamming: Example

#3 search result for the query “kaiser pharmacy online”

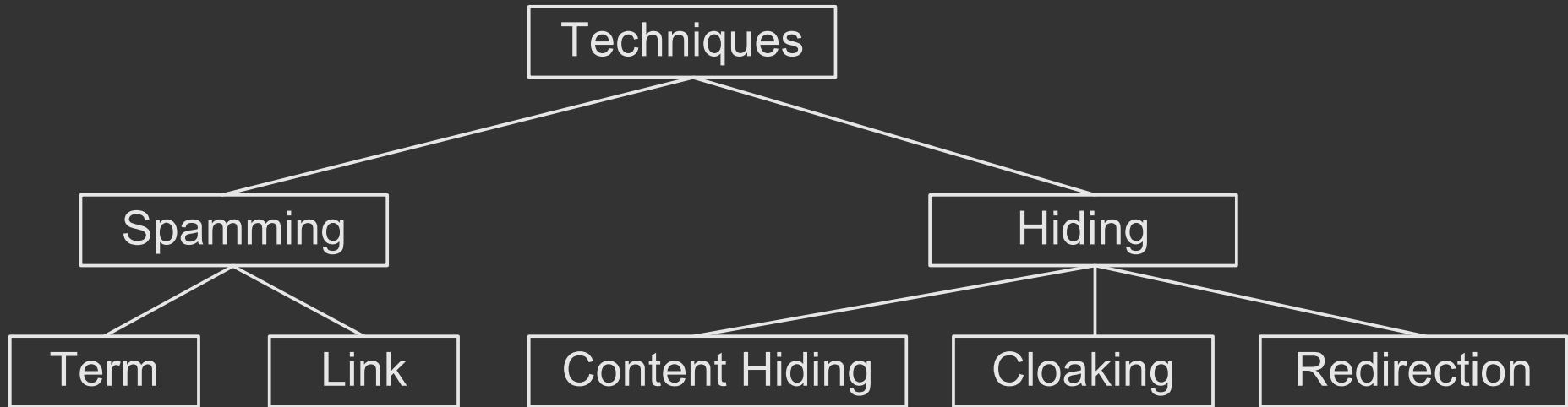
The screenshot shows the homepage of 'UNCOMMONLY CLASSIC ONLINE SHOPPING EXTRAVEGANZA!'. The site features a navigation menu with links to various categories: Lawyers, Loans, Magazines, Mortgage, Nurse, Pharmacy, Posters And Crafts, Ringtones, Shoes, Stock Market, Sunglasses, Theater, Viagra, and Video Games. A 'HOT PRODUCTS' section displays an advertisement for 'ZERO DOWNTIME' hosting services at '\$7.77 PER MONTH'. The 'ARTICLES' section contains a text-based advertisement for 'Pharmacy' that reads: 'Pharmacy is the profession of compounding and dispensing medication. More recently, the term has come to include other services...'. Two yellow callout boxes are overlaid on the image. The first box, on the left, lists the categories: 'Lawyers', 'Loans', 'Mortgage', 'Ringtones', and 'Viagra'. The second box, on the right, contains the text: 'Pharmacy is the profession of compounding and dispensing medication. More recently, the term has come to include other services...'. Both boxes have lines pointing to their respective content on the website.

Web Spamming: Techniques

Spamming = misleading search engines to obtain higher-than-deserved ranking

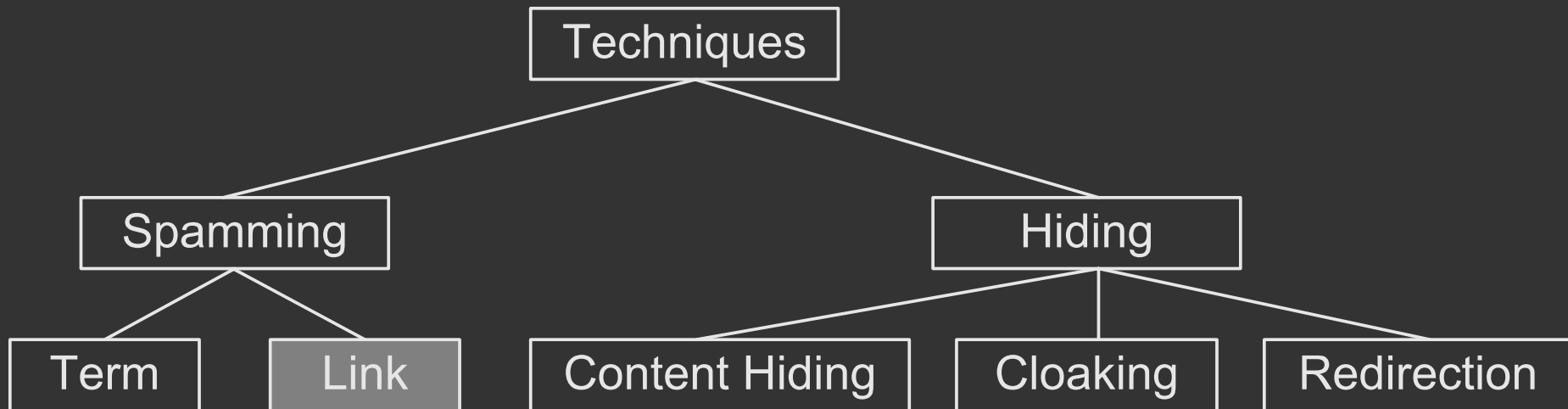
Web Spamming: Techniques

Spamming = misleading search engines to obtain higher-than-deserved ranking



Web Spamming: Techniques

Spamming = misleading search engines to obtain higher-than-deserved ranking



Link spamming = building link structures that boost PageRank score

Link Spamming: Example

#1 search result for the query “switzerland ski”

Switzerland, Switzerland Travel Advice, Ski Switzerland, Skiing in Switzerland, Swiss Ski Resor - Microsoft Internet Explorer

Address: <http://www.skiswitzerland.com/>

The Activelifestyle Travel Network. Focused travel targeting at its best = perfect results for buyer and seller.

Austria ski/resorts	Swiss/ski/resorts	Italy/ski/resorts	France/ski/holidays	Last Minute
skiaustria.com	skiswitzerland.com	skiitaly.com	skifrance.com	dive-lastminute.com
stantonaustria.com	zermatt.com	aostaitaly.com	holidayfrancais.com	golf-lastminute.com
austrianarberg.com	jungfrauregion.com	courmayeur.com		holidays-lastminute.com
lechaustria.com	verbierswitzerland.com	dolomitesitaly.com		ski-lastminute.com
stubaiaustria.com	zermattswitzerland.com	lignoitally.com		
tirolaustria.com	holidaysswitzerland.com	holidaysswitzerland.com		
holidaysaustria.com				

Asia/activities/dest	Asia/activities/dest	Holidays Europe	Luxury	Luxury
travelthailand.com	asiandiveholidays.com	holidayseurope.com	luxuryalpinehotels.com	luxuryhotelsamerica.com
bangkokthailand.com	asianmp3.com	holidaysineurope.com	luxuryasianhotels.com	luxuryhotelscanada.com
pattayathailand.com	mp3thailand.com	europeanreservations.com	luxuryasianresorts.com	luxuryislandresorts.com
phuketthailand.com	thailandhealthcaretimes.com	croatiancoastholidays.com	luxurygolfdestinations.com	luxuryhotelsbangkok.com
thailandgolfmaps.com	thailandpropertytimes.com	sloveniancoast.com	luxuryyachtholidays.com	luxuryski.com

Best Price	Best Price	Best Price	Alpine Sun	Special travel
bestpriceeurope.com	bestpricethailand.com	bestpricetouring.com	alpineholidays.com	activelifestylewoman.com
bestpriceaustria.com	bestpricezermatt.com	bestpriceverbier.com	alpinesecrets.com	euroski-on-line.com
bestpriceitaly.com	bestpricecourmayeur.com	bestpriceairlineickets.com	alpinsummer.com	businesstraveltoday.com
bestpriceswitzerland.com	bestpriceskiing.com	bestpriceairtickets.com	lakesmountainseurope.com	bookhotelsdirect.com
bestpricefrance.com	bestpricegolfing.com	bestpricetravelnetwork.com	hotelsinthealps.com	activelifestyle.com
			alpinegolf.com	activelifestylemall.com
			alpineskimaps.com	gullibletraveler.com

Available Accommodation	Available Accommodation	global apartments	global apartments
availableroomsthailand.com	availableroomsswitzerland.com	alpineapartmentregister.com	lakesmountainsapartments.com
availableroomszermatt.com	availableaccommodationitaly.com	apartmentaustria.com	lignoapartments.com
availableroomsitaly.com	zermattaccommodation.com	apartmentsinthealps.com	matterhornapartments.com
availableroomsfrance.com	zermattapartmentregister.com	apartmentsligno.com	privatealpinehomes.com
availableroomsaustria.com		apartmentswitzerland.com	verbierapartments.com
		apartmentsverbier.com	alpineholidayhomes.com

Link Spamming: Example

#1 search result for the query “switzerland ski”

Switzerland, Switzerland Travel Advice, Ski Switzerland, Skiing in Switzerland, Swiss Ski Resor - Microsoft Internet Explorer

Address <http://www.skiswitzerland.com/>

The Activelifestyle Travel Network. Focused travel targeting at its best = perfect results for buyer and seller.

Austria ski/resorts	Swiss/ski/resorts	Italy
skiaustria.com	skiswitzerland.com	skiitaly.com
stantonaustria.com	zermatt.com	aostaita.com
austrianarberg.com	jungfrauregion.com	courmayeur.com
lechaustria.com	verbierswitzerland.com	dolomiten.com
stubaiaustria.com	zermattswitzerland.com	livigno.com
tirolaustria.com	holidaysswitzerland.com	holidays.com
holidaysaustria.com		

Asia/activities/dest	Asia/activities/dest	Europe
travelthailand.com	asiandiveholidays.com	holidays.com
bangkokthailand.com	asianmp3.com	holidays.com
pattayathailand.com	mp3thailand.com	europe.com
phuketthailand.com	thailandhealthcaretimes.com	croatian.com
thailandgolfmaps.com	thailandpropertytimes.com	slovenia.com

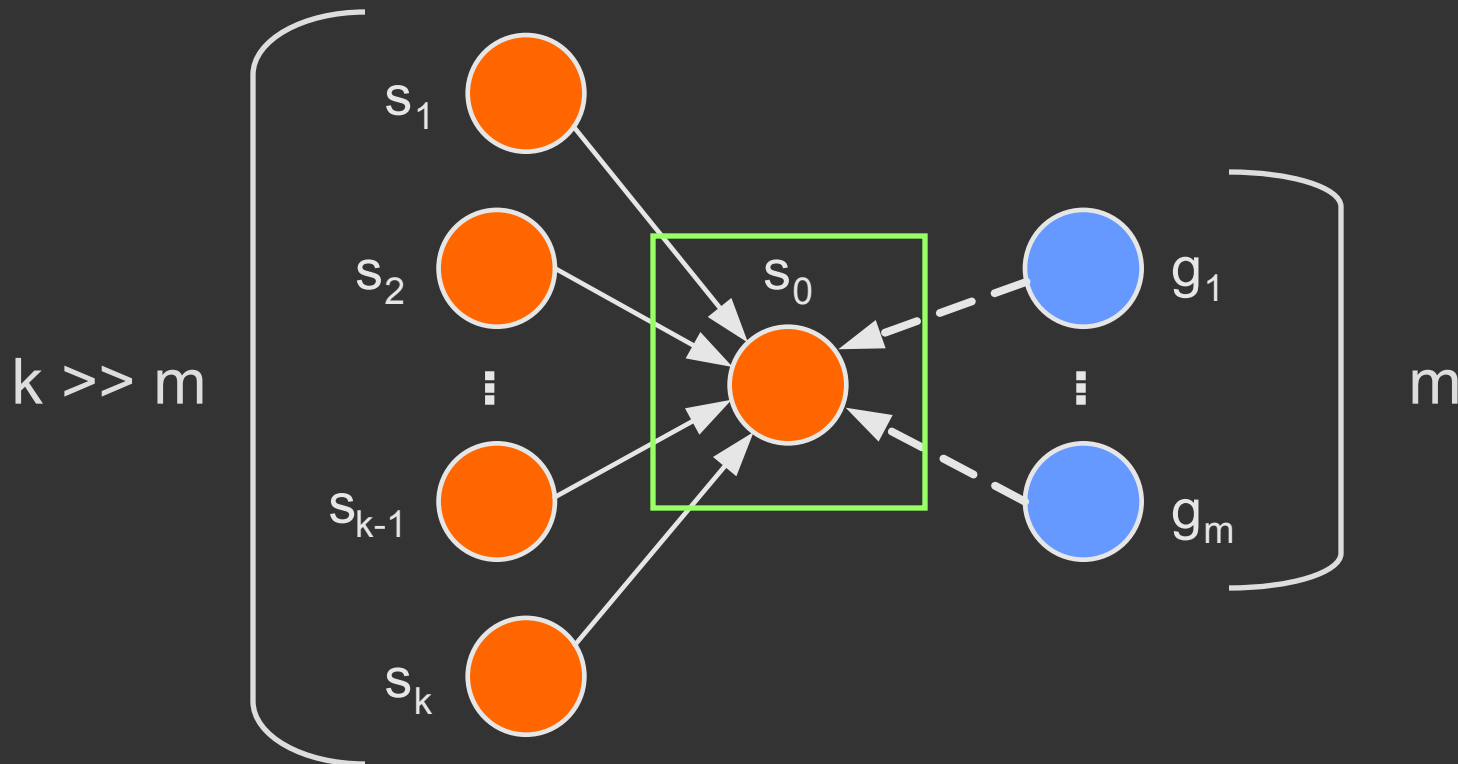
Best Price	Best Price	Best Price	Alpine Sun	Special travel
bestpriceeurope.com	bestpricethailand.com	bestpricetouring.com	alpineholidays.com	activelifestylewoman.com
bestpriceaustria.com	bestpricezermatt.com	bestpriceverbier.com	alpinesecrets.com	euroski-on-line.com
bestpriceitaly.com	bestpricecourmayeur.com	bestpriceairlineickets.com	alpinesummer.com	businesstraveltoday.com
bestpriceswitzerland.com	bestpriceskiing.com	bestpriceairtickets.com	lakesmountainseurope.com	bookhotelsdirect.com
bestpricefrance.com	bestpricegolfing.com	bestpricetravelnetwork.com	hotelsinthealps.com	activelifestyle.com
			alpinegolf.com	activelifestylemall.com
			alpineskimaps.com	gullibletraveler.com

Available Accommodation	Available Accommodation	global apartments	global apartments
availableroomsthailand.com	availableroomsswitzerland.com	alpineapartmentregister.com	lakesmountainsapartments.com
availableroomszermatt.com	availableaccommodationitaly.com	apartmentaustria.com	livignoapartments.com
availableroomsitaly.com	zermattaccommodation.com	apartmentsinthealps.com	matterhornapartments.com
availableroomsfrance.com	zermattapartmentregister.com	apartmentslivigno.com	privatealpinehomes.com
availableroomsaustria.com		apartmentswitzerland.com	verbierapartments.com
		apartmentsverbier.com	alpineholidayhomes.com

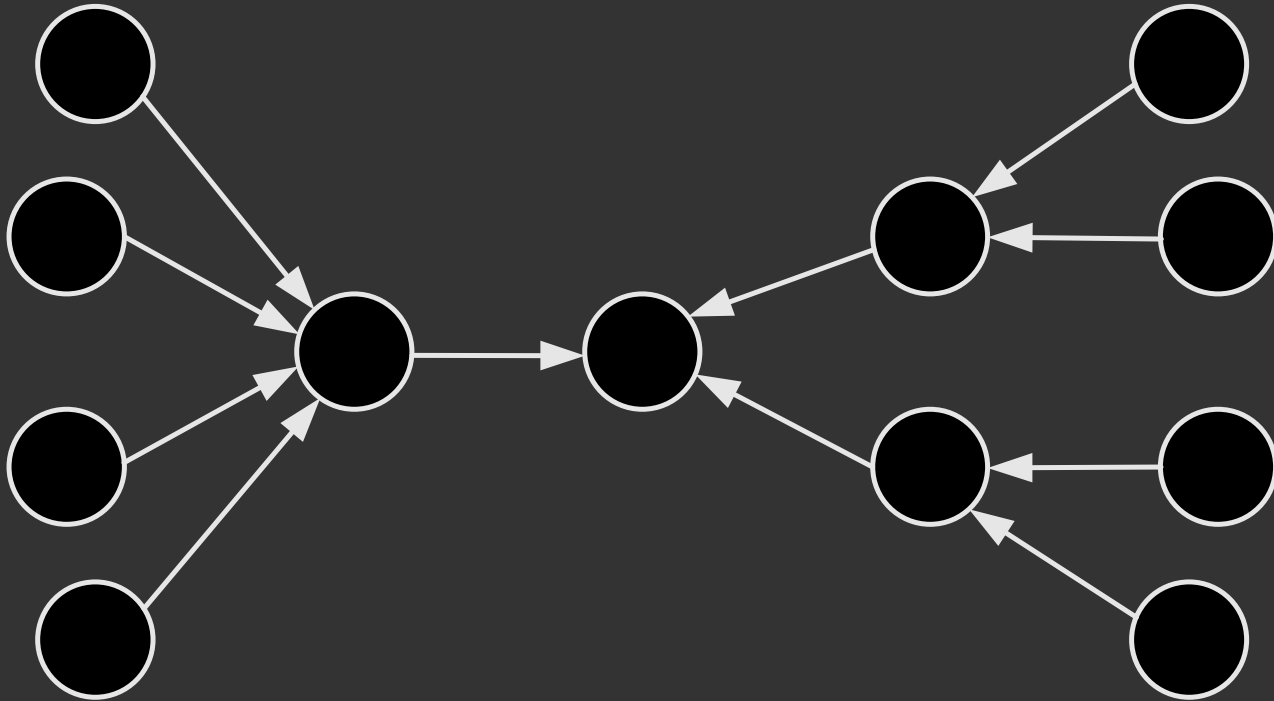
Internet

Link Spamming: Our Goal

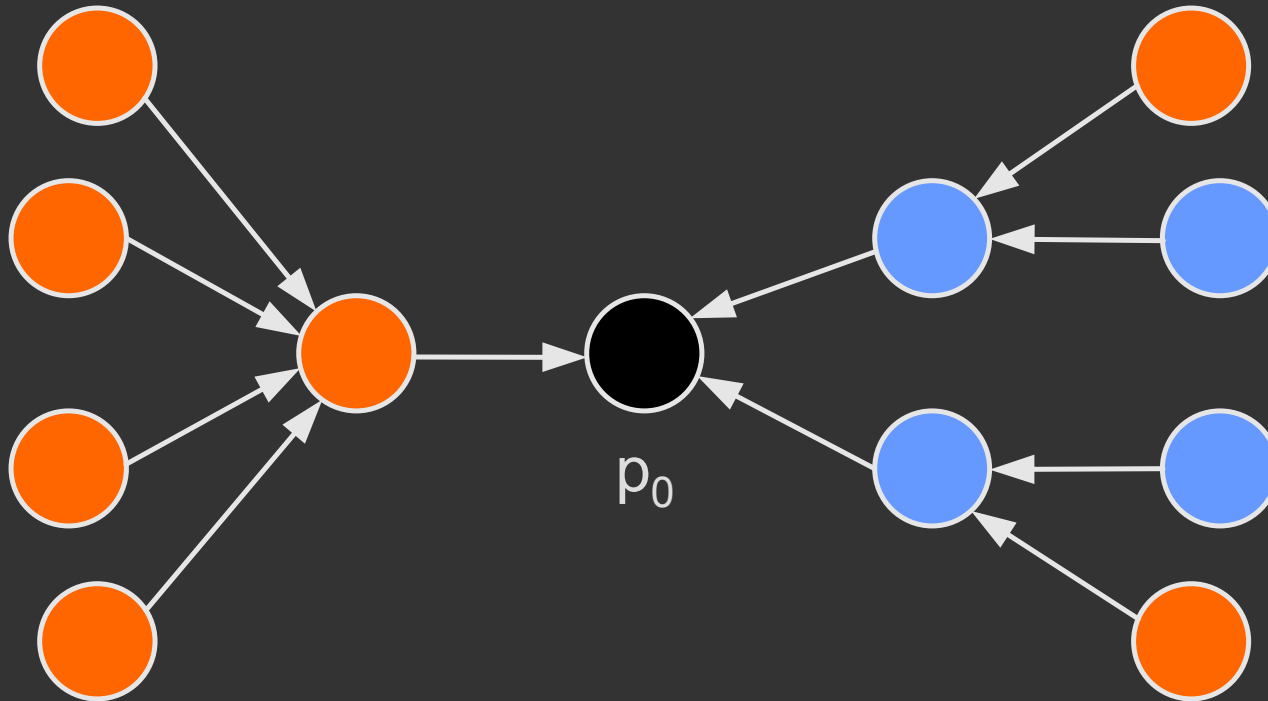
Detect pages that achieve high PageRank through link spamming



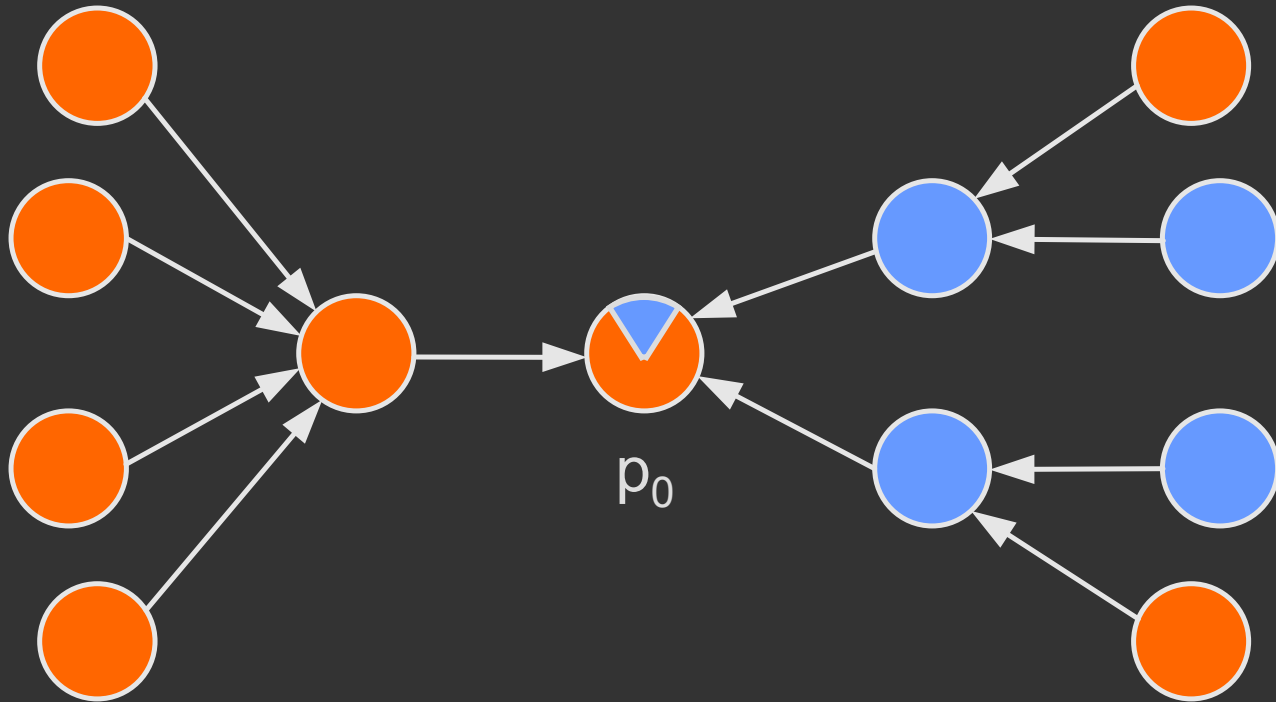
PageRank Contribution



PageRank Contribution



PageRank Contribution



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

$$p_0^- = c^2 \frac{6(1-c)}{n} + c \frac{(1-c)}{n}$$

Spam Mass: Definition

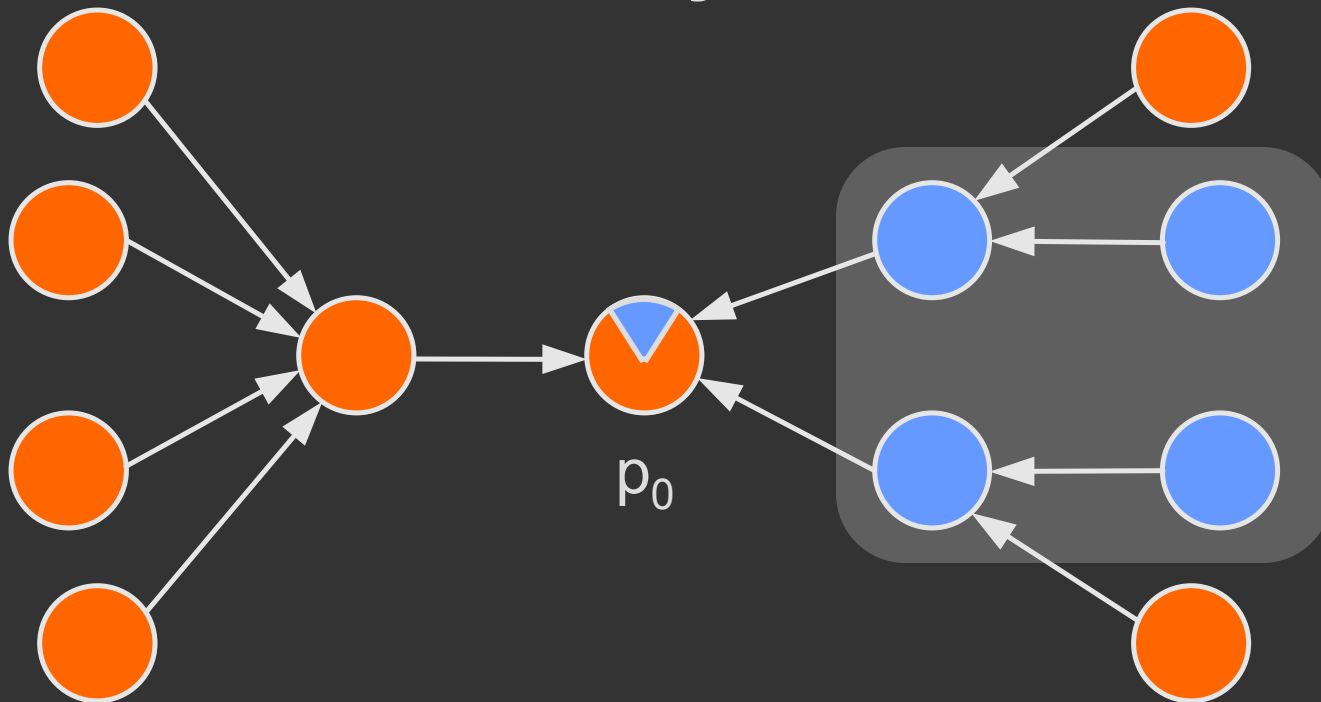
- Absolute mass
 - **Amount** of PageRank coming from spam

- Relative mass
 - **Fraction** of PageRank coming from spam
 - More useful in practice



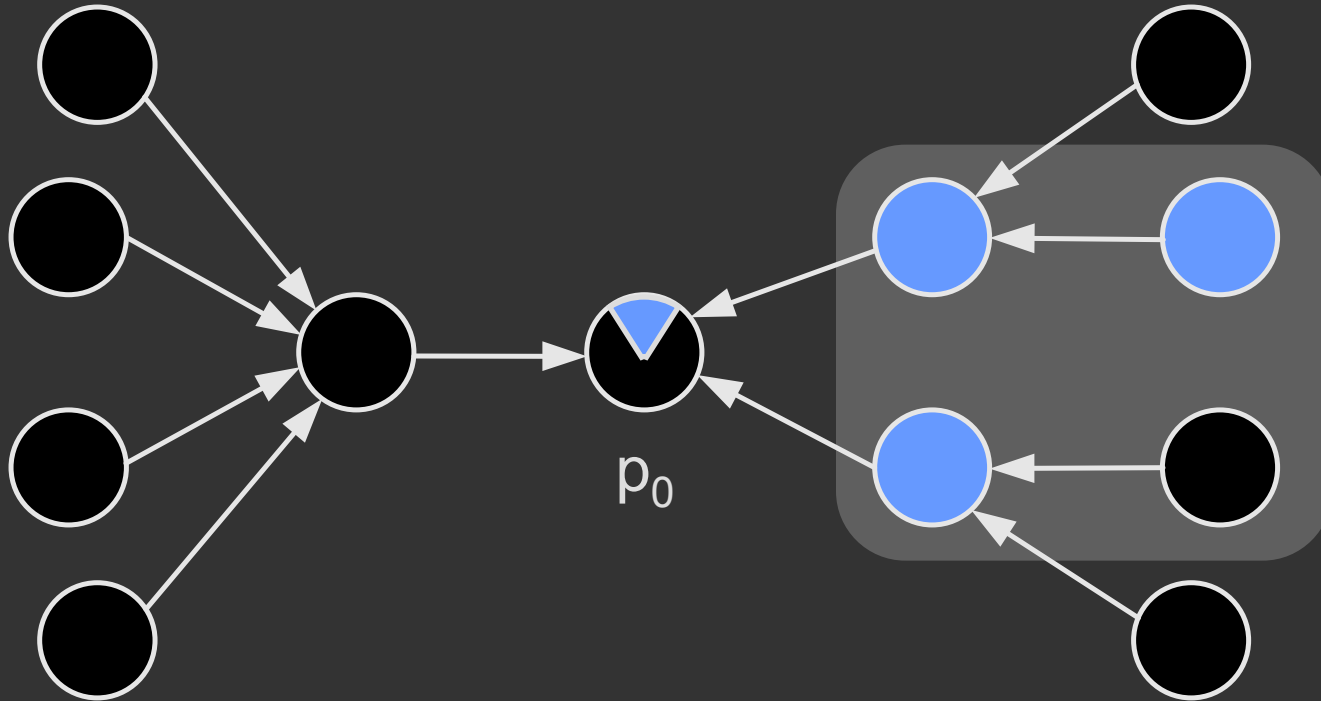
Spam Mass: Estimation

Ideally...



Spam Mass: Estimation

In practice...



- Approximate the set of good nodes by a subset called **good core**

Spam Mass: Algorithm

1. Create good core
2. Compute PageRank scores p_i and p_i^+
3. Compute estimated relative mass m_i as $(p_i - p_i^+) / p_i$
4. For all pages i with large PageRank
Mark page as spam if $m_i > \text{threshold}$

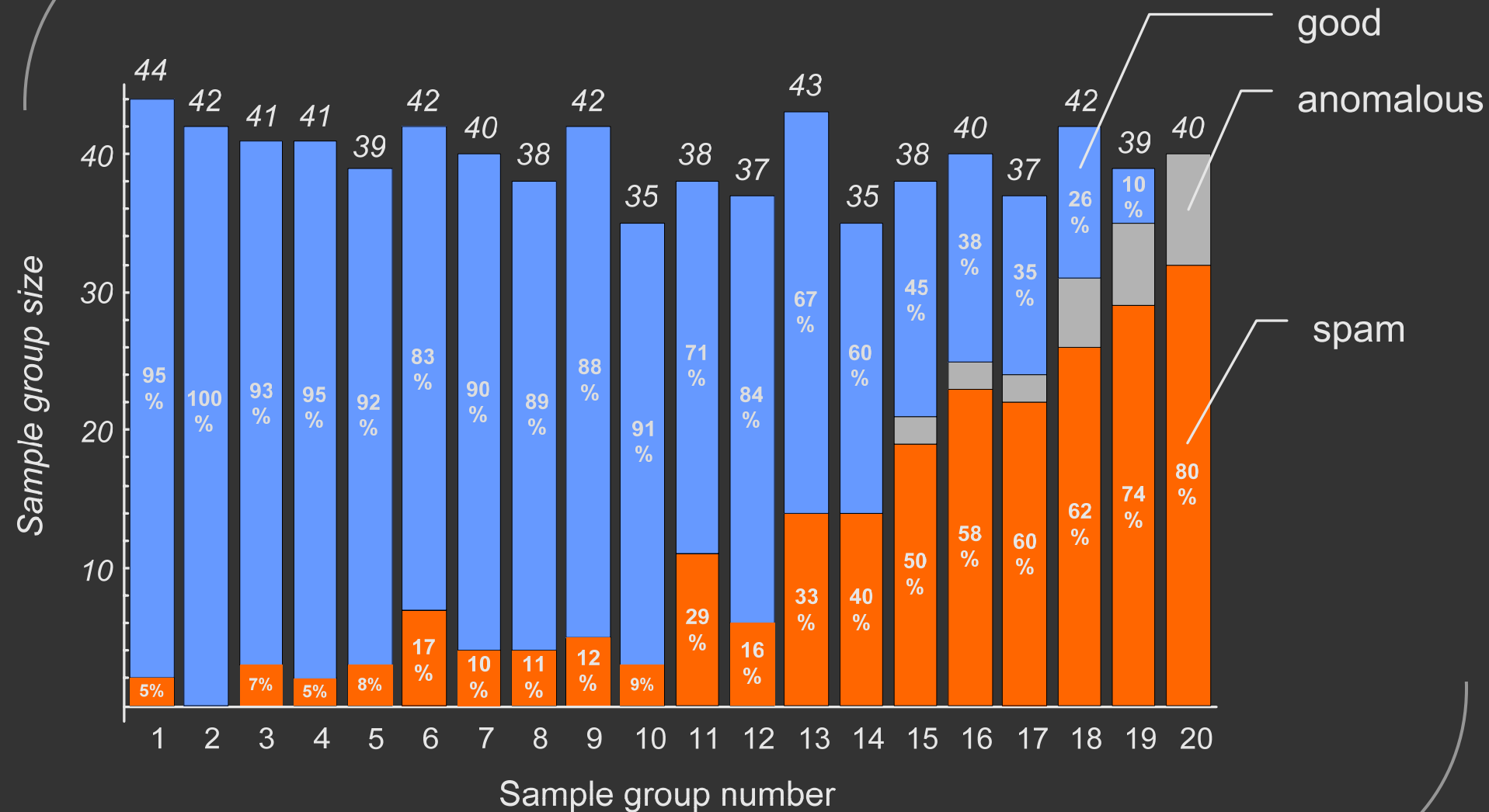
Experiments: Data

- Yahoo! web index → host graph
 - 73.3M nodes
 - 979M links
- Good core
 - High-quality web directory: 16,780
 - Governmental hosts: 55,320
 - **Educational hosts: 434,000**

Experiments: Data

- Sample
 - 0.1% of nodes with PageRank $>$ 10x minimum
 - 892 nodes
 - Manually labeled good, spam
- Relative mass groups (approx. same size)
 - Group 1: 44 samples with smallest rel. mass
 - ...
 - Group 20: 40 samples with largest rel. mass

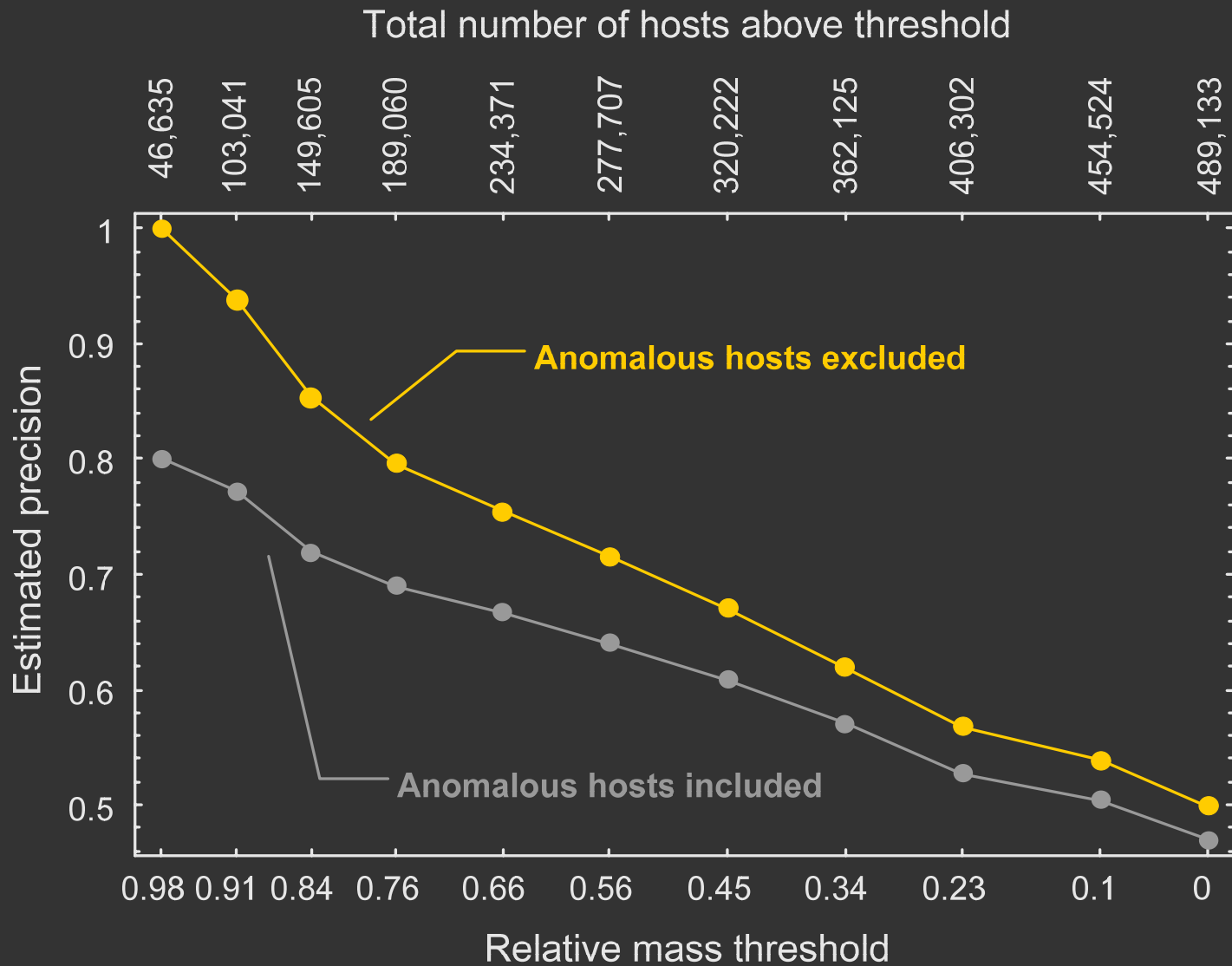
Experiments: Relative Mass



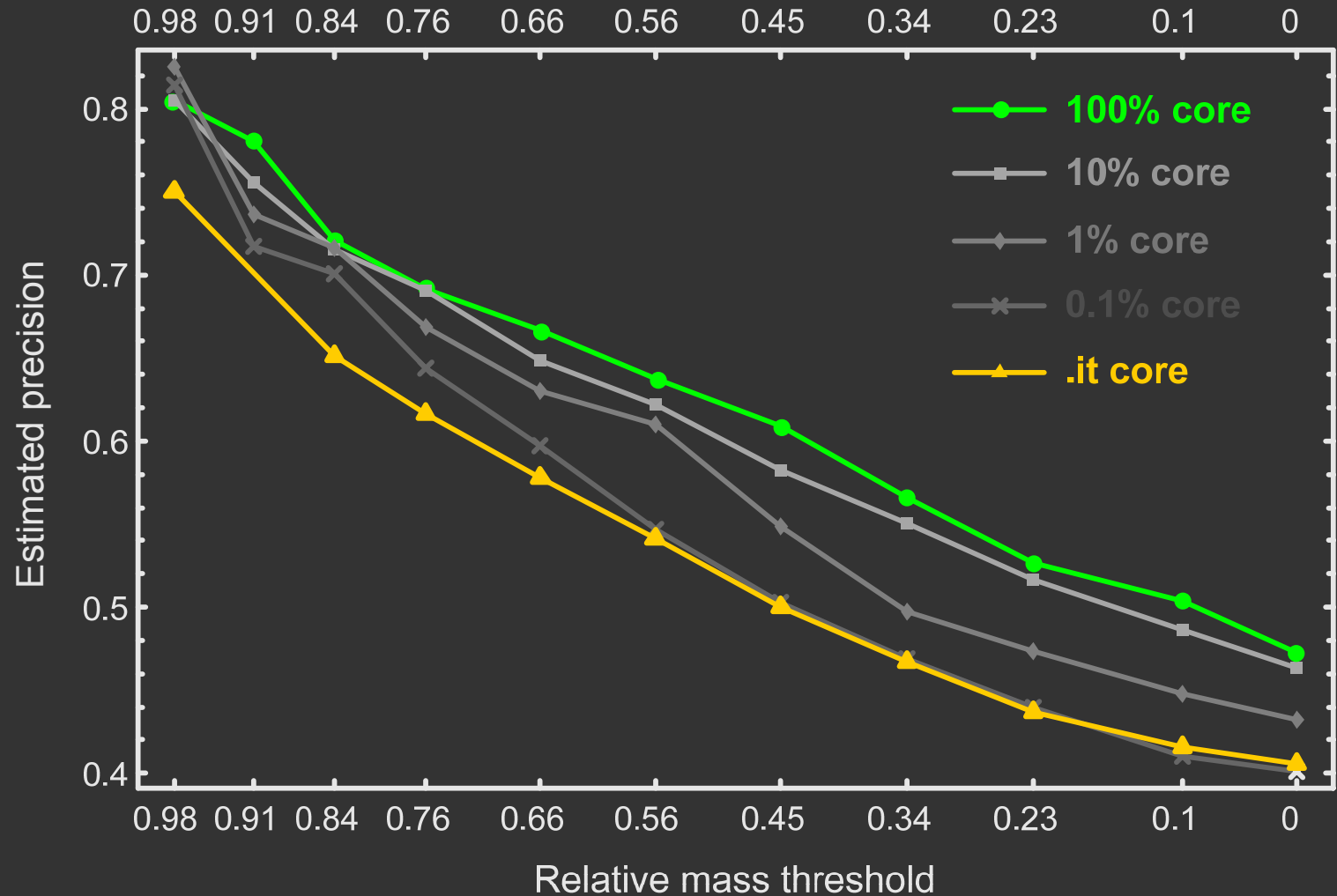
Experiments: Relative Mass

- Anomalies
 - *.alibaba.com
 - *.blogger.com.br
 - Polish hosts → only 12 .pl in good core

Experiments: Relative Mass



Experiments: Relative Mass



Related Work

- PageRank analyses
 - [Bianchini+2005], [Langville+2004]
- Link spam analyses
 - [Baeza+2005], [Gyöngyi+2005]
- Link spam detection
 - Statistics: [Fetterly+2004], [Benczúr+2005]
 - Collusion detection: [Zhang+2004], [Wu+2005]
- TrustRank
 - [Gyöngyi+2004]

Conclusions

- Web spamming
 - Manipulation of search engine ranking
 - Focus on link spamming
- Spam mass
 - ~ PageRank contribution
 - Useful in link spam detection
- Strong experimental results
 - Virtually 100% of top 47K nodes spam
 - 94% of top 105K nodes spam