



# Web Spam Taxonomy

Zoltán Gyöngyi  
Hector Garcia-Molina

# Roadmap

- Subject
- Observed behavior
  - Boosting
    - Term-based
    - Link-based
  - Hiding
- Statistics
- Challenges

# Roadmap

- Subject
- Observed behavior
  - Boosting
    - Term-based
    - Link-based
  - Hiding
- Statistics
- Challenges

# Subject



So... who does what?

## Spamming

deliberate human action

meant to trigger unjustifiably  
high **ranking**

importance  
*(global)*

relevance  
*(query-dependent)*

# Subject

- Monetization

- Better ranking = higher click-through rate
- Search engine optimization
- Affiliate spam

Why?

[Compare Canon Prices](#) - search.ActivShopper.com  
Compare Prices For Canon Cameras In The USA Before You Buy!  
Ads by Goooooogle

[Study online at MusicianUniversity.com](#)  
Study audio, music and instruments courses  
Start your course with 25% off ONLY Today

### Canon Digital Rebel XT 8MP Digital SLR Camera with EF-S 18-55mm f3.5-5.6 Lens (Black)

**Features**

- Powered by rechargeable Lithium-ion battery (included, with charger)
- 8.0-megapixel CMOS sensor captures enough detail for photo-quality 16 x 22-inch prints
- Includes Canon's EF-S 18-55mm, f3.5-5.6 zoom lens
- DIGIC II Image Processor provides fast, accurate image processing; captures images at a rate of up to 3 frames per second
- Fast start-up time--2 seconds

**Canon Digital Rebel XT 8MP Digital SLR Camera with EF-S 18-55mm f3.5-5.6 Lens (Black) Reviews**

**This is one awesome camera**

🎵 🎵 🎵 🎵 🎵 🎵 🎵 🎵 🎵 🎵 (10 out of 10)

I bought from amazon after being let down by Dell. It arrived next day and on the weekend i gave it a good workout at a local airshow. I took just under 1000 photos and was amazed by how well the shots came out in sport mode. Very quick focus and spot on. excellent camera and well built , not small as some people have noted, works for me...

**Ready To Buy Online Now?**

List Price: ~~\$1,499.99~~  
Online Price at this moment: **Too Low To Display \$1011.67**  
Buy used: \$915.00  
\$440.00

[Buy, See Prices](#)

Canada | UK | Germany/Europe

CRMAV.com Pro Audio  
MusiciansNews.com  
DJMusicEquipment.com  
GuitarsMusicGear.com  
GuitarSheetMusicTabs.com  
Digital Pianos Keyboards  
Live Sound PA Systems  
RecordingMixingMastering.com  
DigitalCamerasPhotography.com  
DigitalVideoCameraEditing.com  
ShoesClothesFashion.com  
CarsMotorsTrucks.com  
StayLimy.com Health/Sports  
PC4D.com  
Computers/Technology  
House Home Furniture

# Subject

- Monetization

- Better ranking = higher click-through rate
- Search engine optimization
- Affiliate spam

Why?

Compare Canon Prices - search.ActivChopper.com  
Compare Prices For Canon Cameras In The USA Before You Buy!

Ads by Google

Study online at MusicianUniversity.com  
Study audio, music and instruments courses  
Start your course with 25% off ONLY Today

Canon Digital Rebel XT BMP Digital SLR Camera with EF-S 18-55mm f3.5-5.6 Lens (Black)

Features

- Powered by rechargeable Lithium-ion battery (included)
- 8.0-megapixel CMOS sensor captures enough detail for
- Includes Canon's EF-S 18-55mm, f3.5-5.6 zoom lens
- DIGIC II Image Processor provides fast, accurate image
- Fast start-up time--2 seconds

Canon Digital Rebel XT BMP Digital SLR Camera with EF-S 18-55mm f3.5-5.6 Lens (Black) Reviews

This is one awesome camera

10 out of 10

I bought from amazon after being let down by Dell. It arrived next day and on the weekend i gave it a good workout at a local airshow. I took just under 1000 photos and was amazed by how well the shots came out in sport mode. Very quick focus and spot on. excellent camera and well built, not small as some people have noted, works for me...

List Price: \$1,499.99  
Online Price at this moment: **Too Low To Display \$1011.67**  
Buy Used: \$915.00 \$440.00

Buy, See Prices

Germany | FR | Germany/Europe

CRMV.com Pro Audio  
MusiciansNews.com  
DJMusicEquipment.com  
GuitarsMusicGear.com  
GuitarShee#MusicTabs.com  
Digital Piano Keyboards  
Live Sound PA Systems  
RecordingMixingMastering.com  
DigitalCamerasPhotography.com  
DigitalVideoCameraEditing.com  
ShoesClothesFashion.com  
CarsMotorsTrucks.com  
StayLimy.com Health/Sports  
FC4D.com  
Computers/Technology  
House Home Furniture

How?

# Roadmap

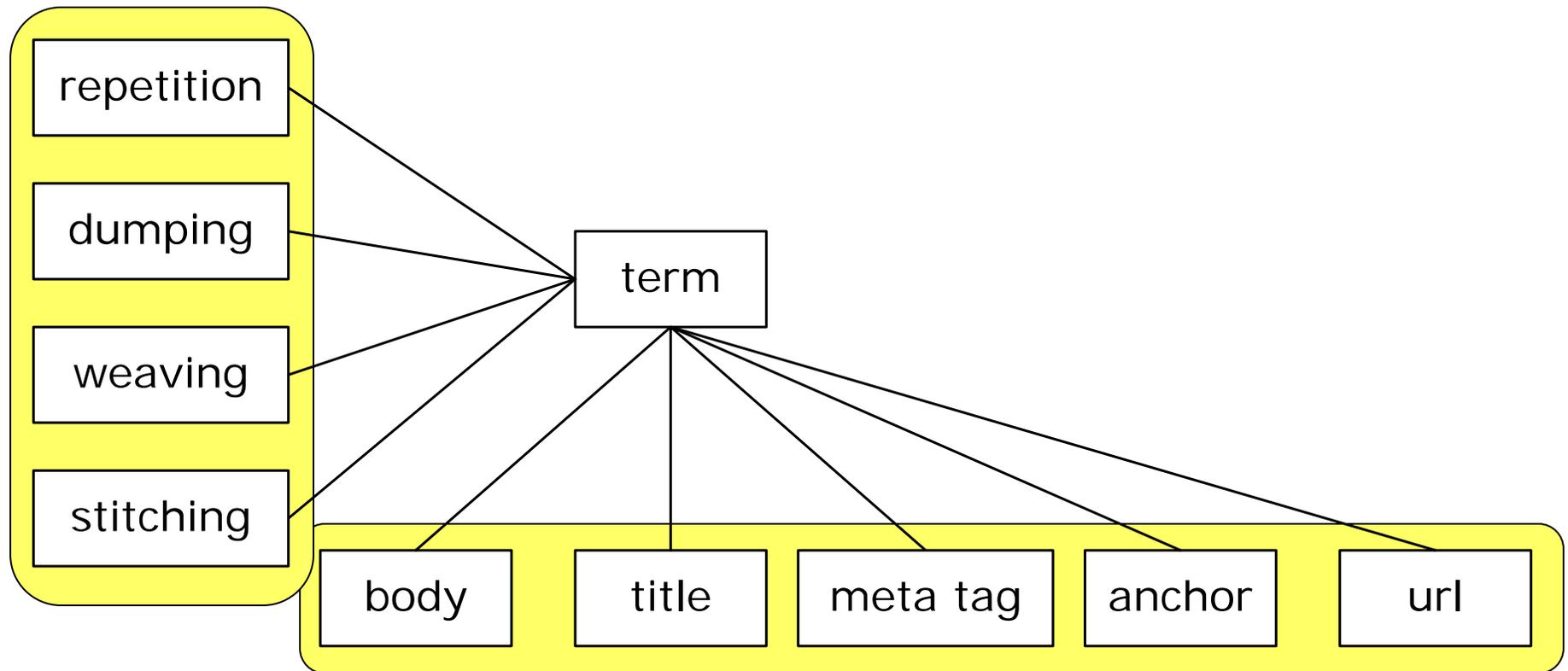
- Subject
- Observed behavior
  - Boosting
    - Term-based
    - Link-based
  - Hiding
- Statistics
- Challenges

# Techniques / Boosting

- Used to increase ranking
- **Hypertext** boosting
  - Term
    - Relevance (one/many queries)
    - Target: TF-IDF variants
  - Link
    - Importance
    - Target: inlink/outlink count, HITS, PageRank

# Techniques / Boosting / Term

**how?**



**what?**

# Techniques / Boosting / Term

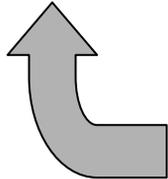
What?

**meta tag**

**title**

**body**

```
<html>
  <head>
    <meta name = "keywords" content = "teddy
    bears; plush bears; plus animals; gift bears; toy
    bears; stuffed bears" >
    <title>Teddy Bears</title>
  </head>
  <body>
    Our customers agree that we are the best online
    retailer of plush teddy bears!
    ...
  </body>
</html>
```



```
<html>
  ...
  A great <a href = "plush.com">stuffed plush bear</a>
  store.
</html>
```

**url**      **anchor text**

# Techniques / Boosting / Term

How?

- repetition repetition **repetition**  
repetition repetition repetition
- dumortierite dumose dumous dump  
dumpage dumper dumpily dumpiness  
**dumping** dumpish dumpishly
- *work in **weaving** three-women teams*  
*is an ancient textile art on looms*
- *please refrain from using the **phrase***  
**stitching** wounds located on the lower  
limbs

# Techniques / Boosting / Term

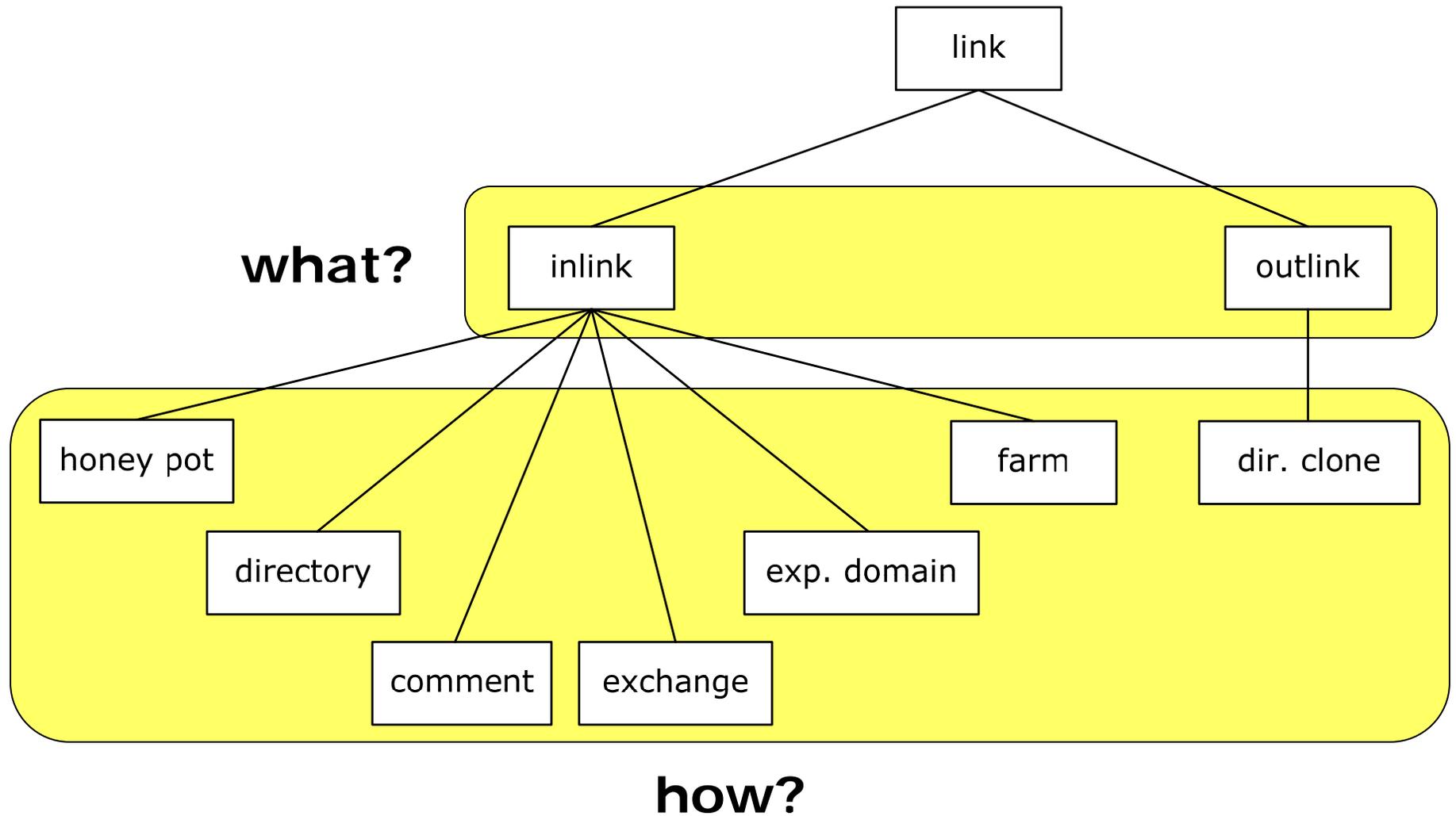
How?

- repetition repetition **repetition**  
repetition repetition repetition
- dumortierite dumose dumous dump  
dumpage dumper dumpily dumpiness  
**dumping** dumpish dumpishly
- *work in **weaving** three-women teams*  
*is an ancient textile art on looms*
- *please refrain from using the phrase*  
**stitching** wound  
limbs



- heuristics
- statistical analysis

# Techniques / Boosting / Link



# Techniques / Boosting / Link

How?

- Directory clones
  - Duplicate (parts of) DMOZ
- Comment spam
  - Post messages (containing links) to
    - Blogs
    - (Unmoderated) forums
    - Wikis
- Link spam farms
  - Increase size
  - Increase collusion



[MCL'05]

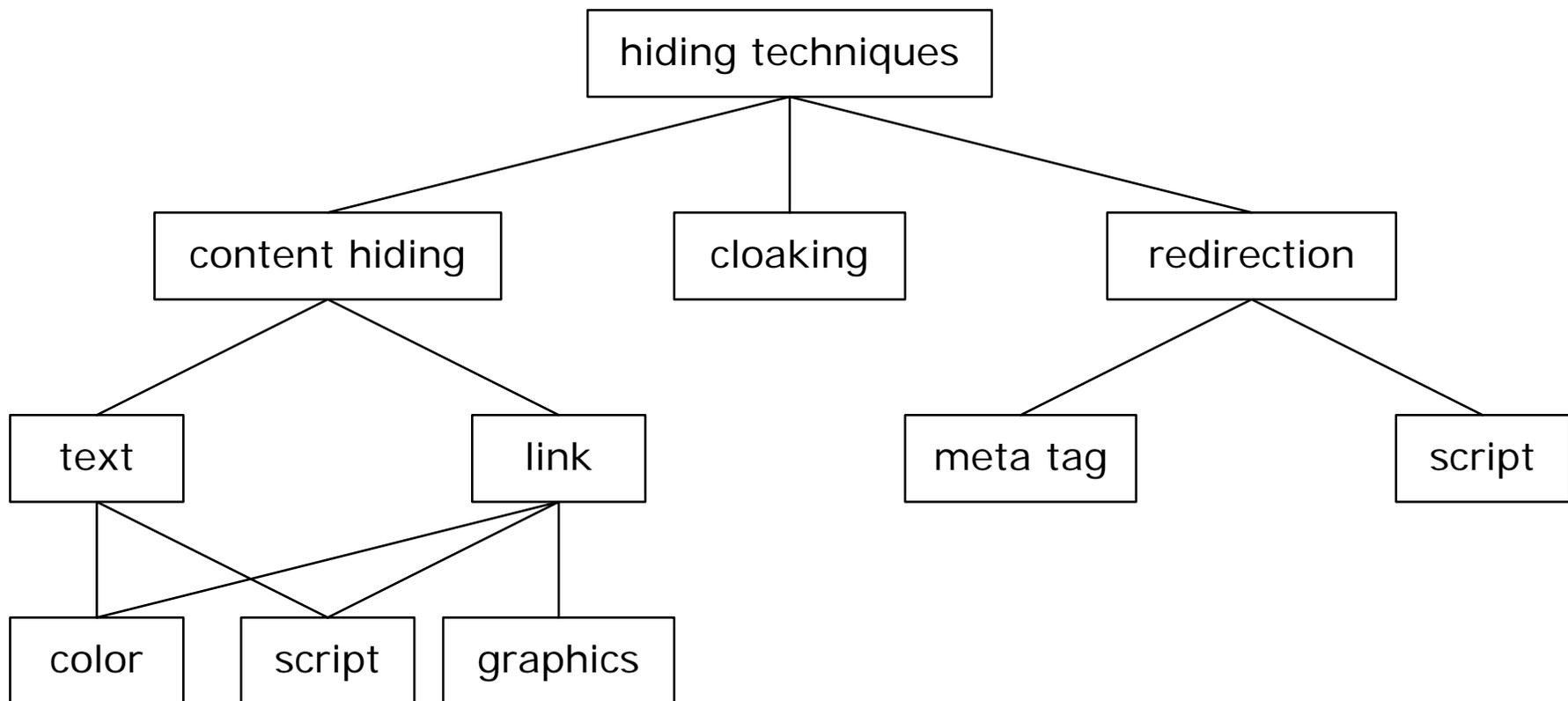


[BYCL'05]

[BCSU'05]

# Techniques / Hiding

- Used to conceal boosting



# Techniques / Hiding

- Content hiding

```
<style type = "text/css" >  
  body {  
    background-color: white;  
    color: white; }  
</style>
```

```
<div style = "visibility: hidden" >You  
can't see me!</div>
```

```
<a href = "..." ><img src  
= "1x1.gif" ></img></a>
```

- Cloaking

- Identify web crawlers
- Serve a different version of the page

# Techniques / Hiding

- Redirection
  - Redirect on load from a heavily spammed page to the true target

```
<meta http-equiv = "refresh" content =  
"0; url=plush.com" >
```

```
<script type = "text/javascript" > <!--  
    eval(window.location = "plush.com");  
//-->  
</script >
```



[WD'05]

# Roadmap

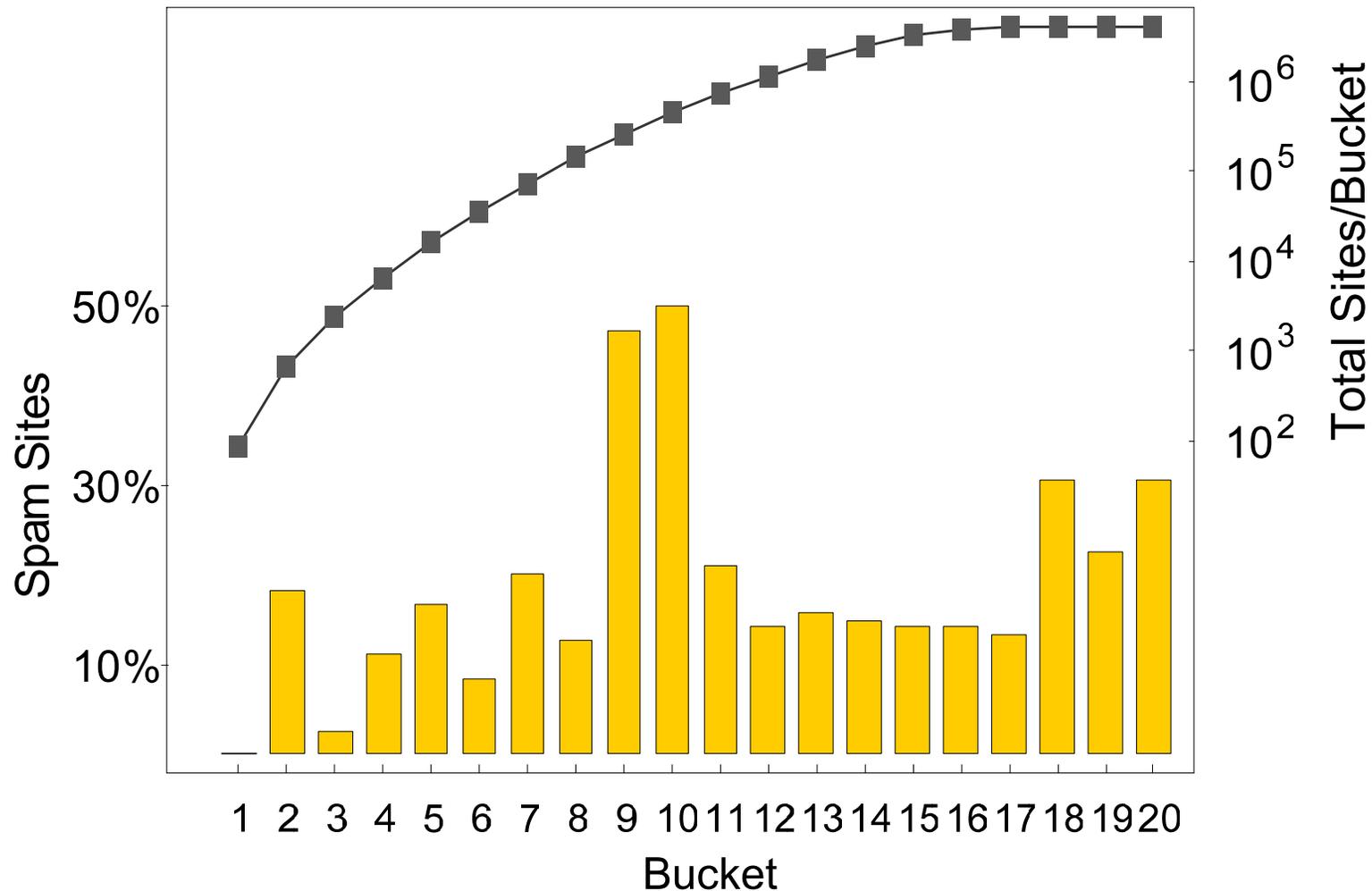
- Subject
- Observed behavior
  - Boosting
    - Term-based
    - Link-based
  - Hiding
- **Statistics**
- Challenges

# Statistics

- [FMN'04]/1
  - Beginning of 2003
  - 150M total / 751 sample pages
  - **8.1%** spam
- [FMN'04]/2
  - Summer of 2002
  - 429M total / 535 sample pages
  - **6.9%** spam
- [GGMP'04]
  - August 2003
  - 31M total / 748 sample sites
  - **18%** spam

# Statistics

- PageRank of spam



# Roadmap

- Subject
- Observed behavior
  - Boosting
    - Term-based
    - Link-based
  - Hiding
- Statistics
- Challenges

# Challenges

- Spam prevalence statistics
  - Per type
  - At various levels of granularity
  - In index vs. in results
- Spam neutralization
  - Spam-proof ranking algorithms (?)
  - Better use of human judgment
    - Exploitation of implicit feedback
    - Better semantic separation
  - Economy/game-theory + ads

# Conclusions

- Spamming techniques
  - Term-based or link-based
  - Of various complexity/efficiency
- Spam detection techniques
  - Wide scale
  - Work in progress
- Challenges
  - Statistics
- Contact: [zoltan@cs.stanford.edu](mailto:zoltan@cs.stanford.edu)