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?

Subject

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

Why?

How?

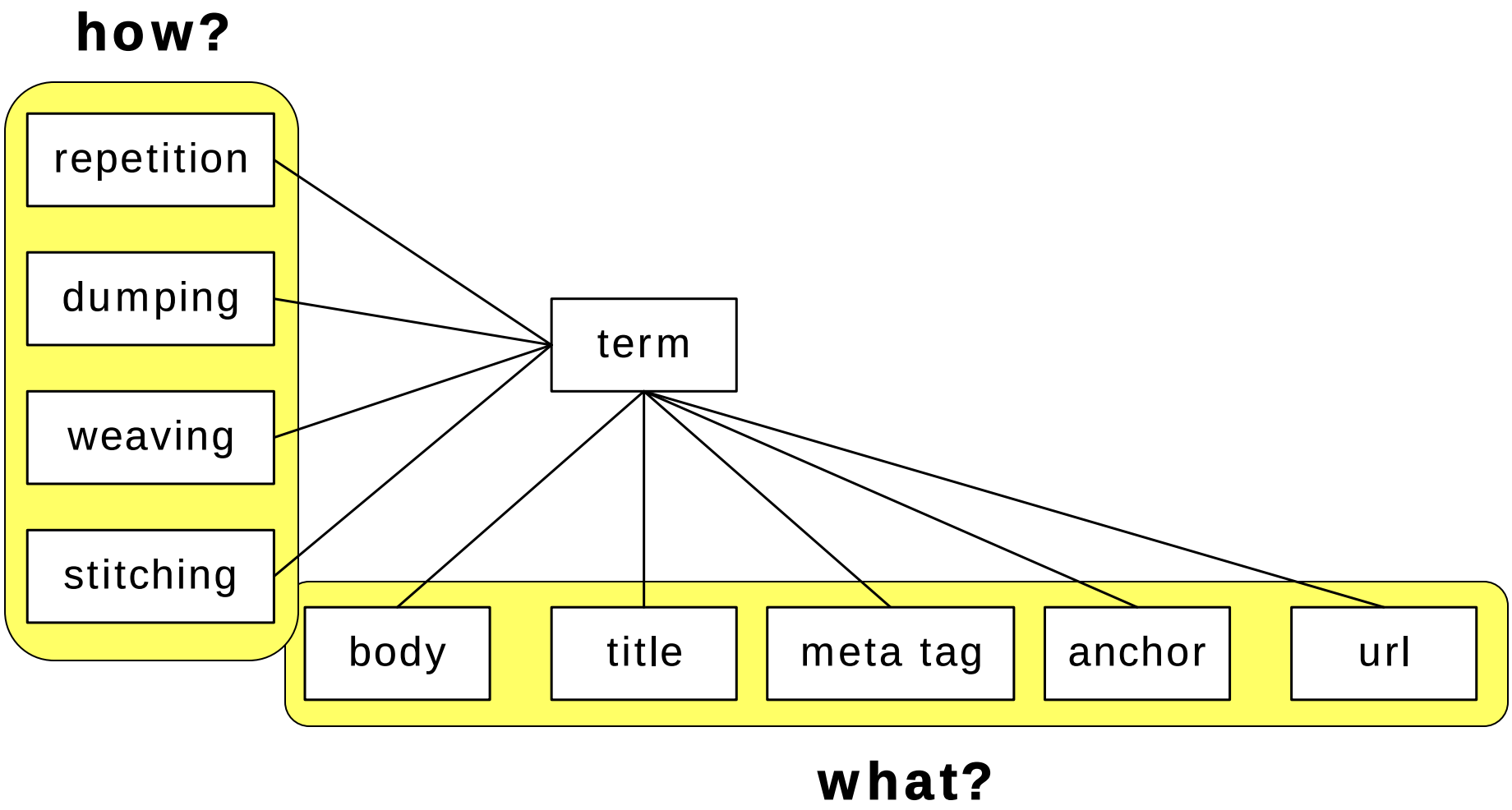
Roadmap

- 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



Techniques / Boosting / Term

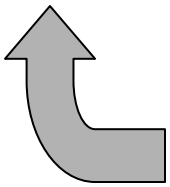
```
<html>
    <head>
        <meta name = “keywords” content = “teddy
        bears; plush bears; plus animals; gift bears; toy
        bears; stuffed bears”>
        <title>Teddy Bears</title>
    </head>
    <body>
        Our customers agree that we are the best online
        retailer of plush teddy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

What?

meta tag

title

body



```
<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

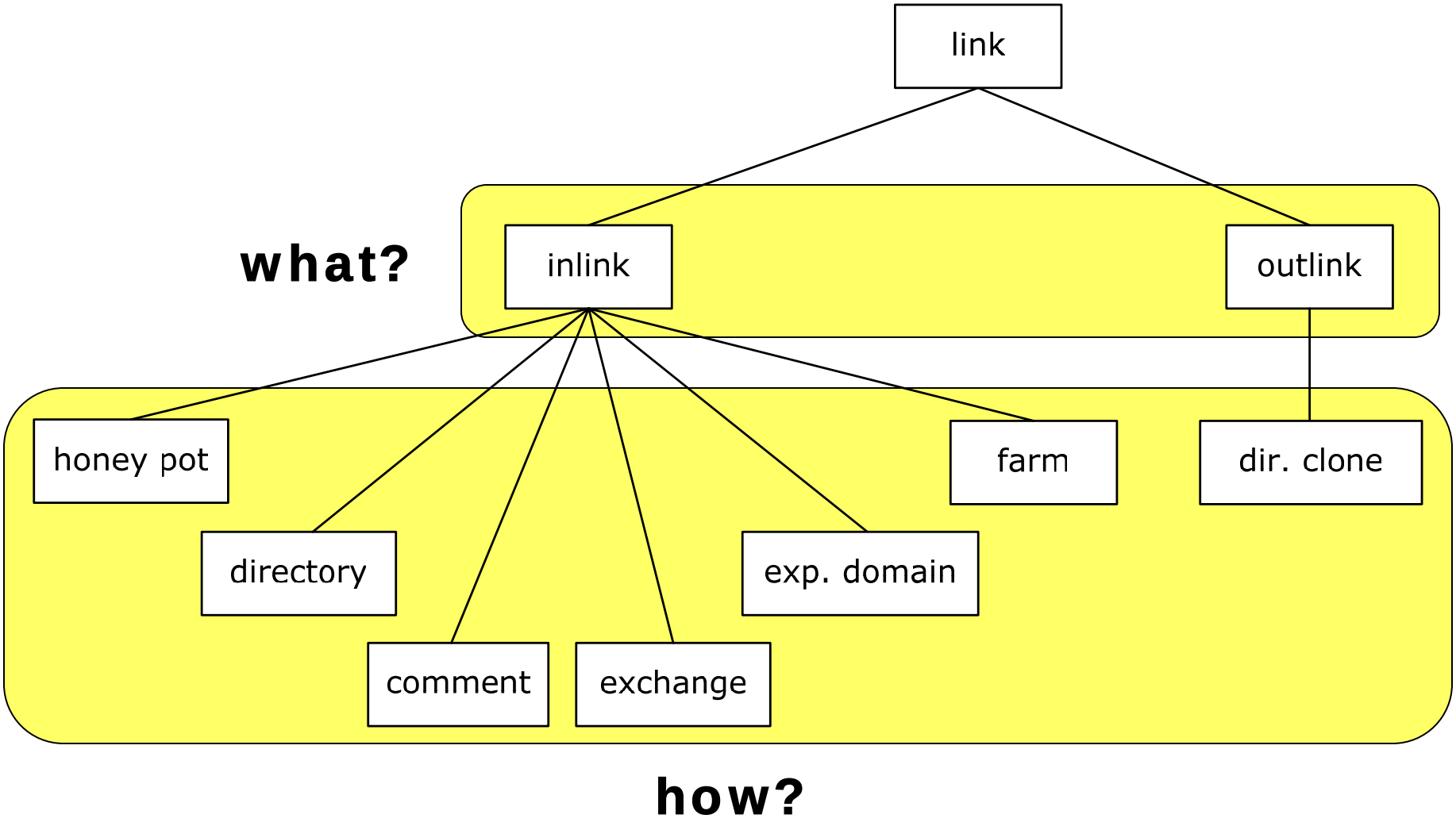
Techniques / Boosting / Term

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** woun 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

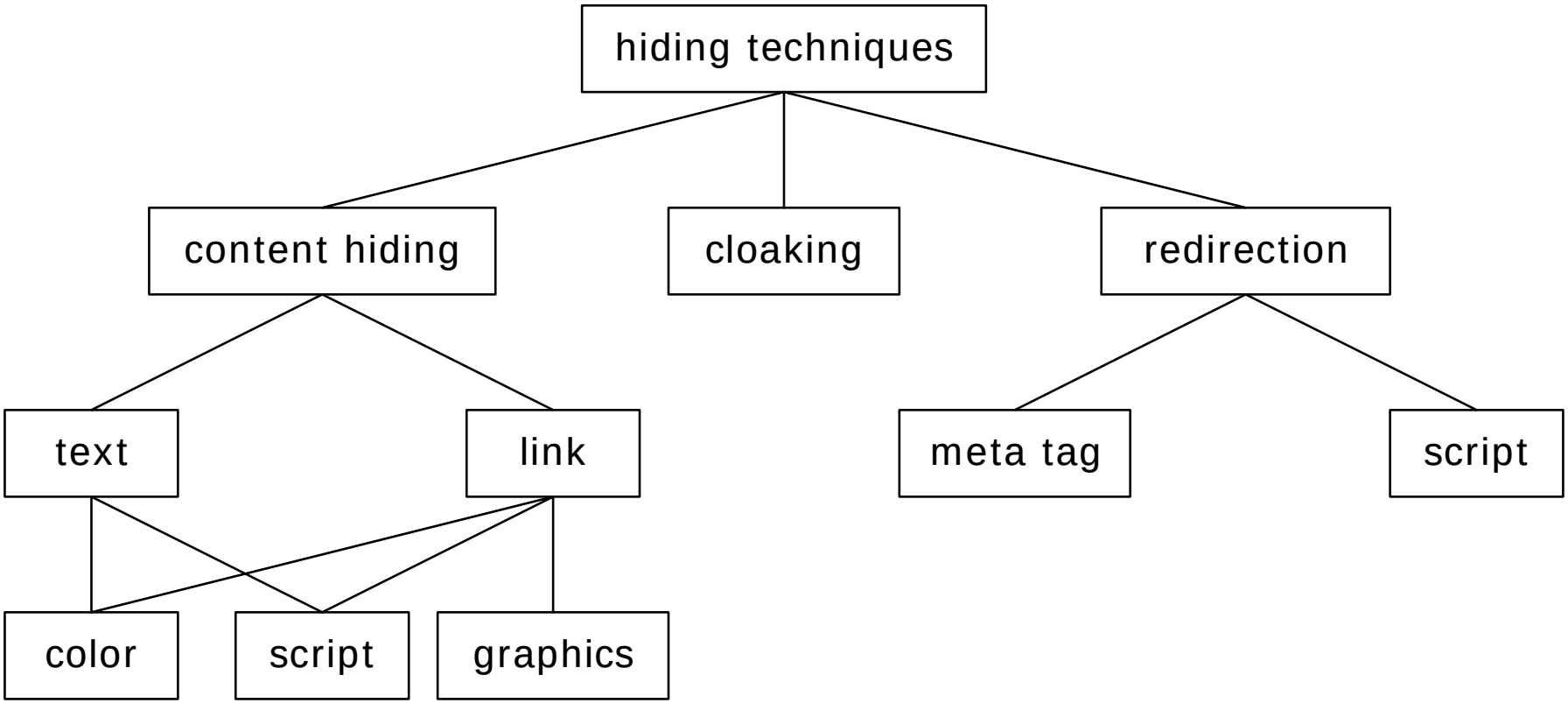
[MCL'05]

[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>
```

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

```
<div style = “visibility: hidden”>You
can’t see me!</div>
```

- 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

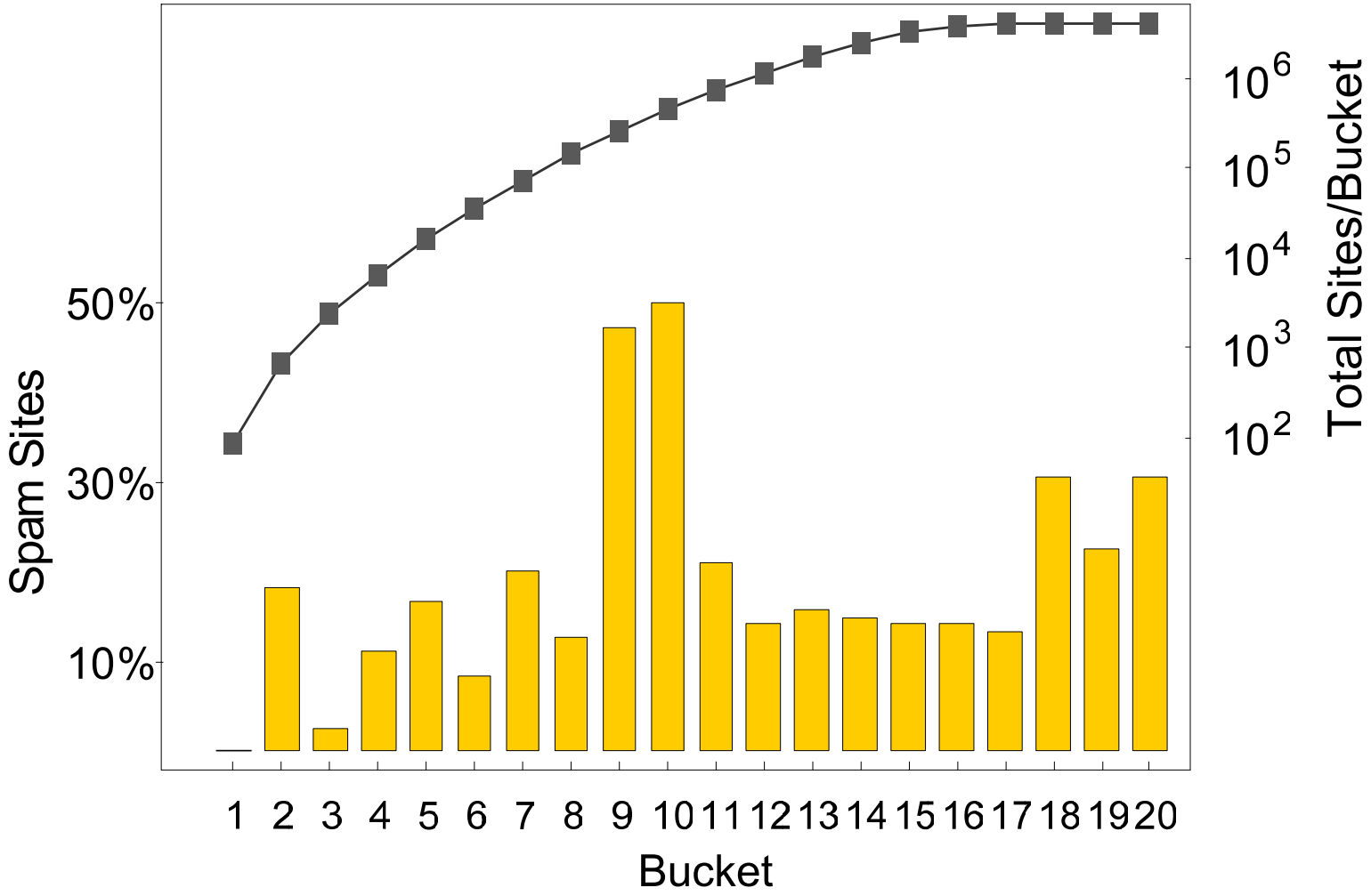
- Subject
- Observed behavior
  - Boosting
    - Term-based
    - Link-based
  - Hiding
- Statistics
- Challenges

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