Link Spam Detection Based on Mass Estimation

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

Roadmap

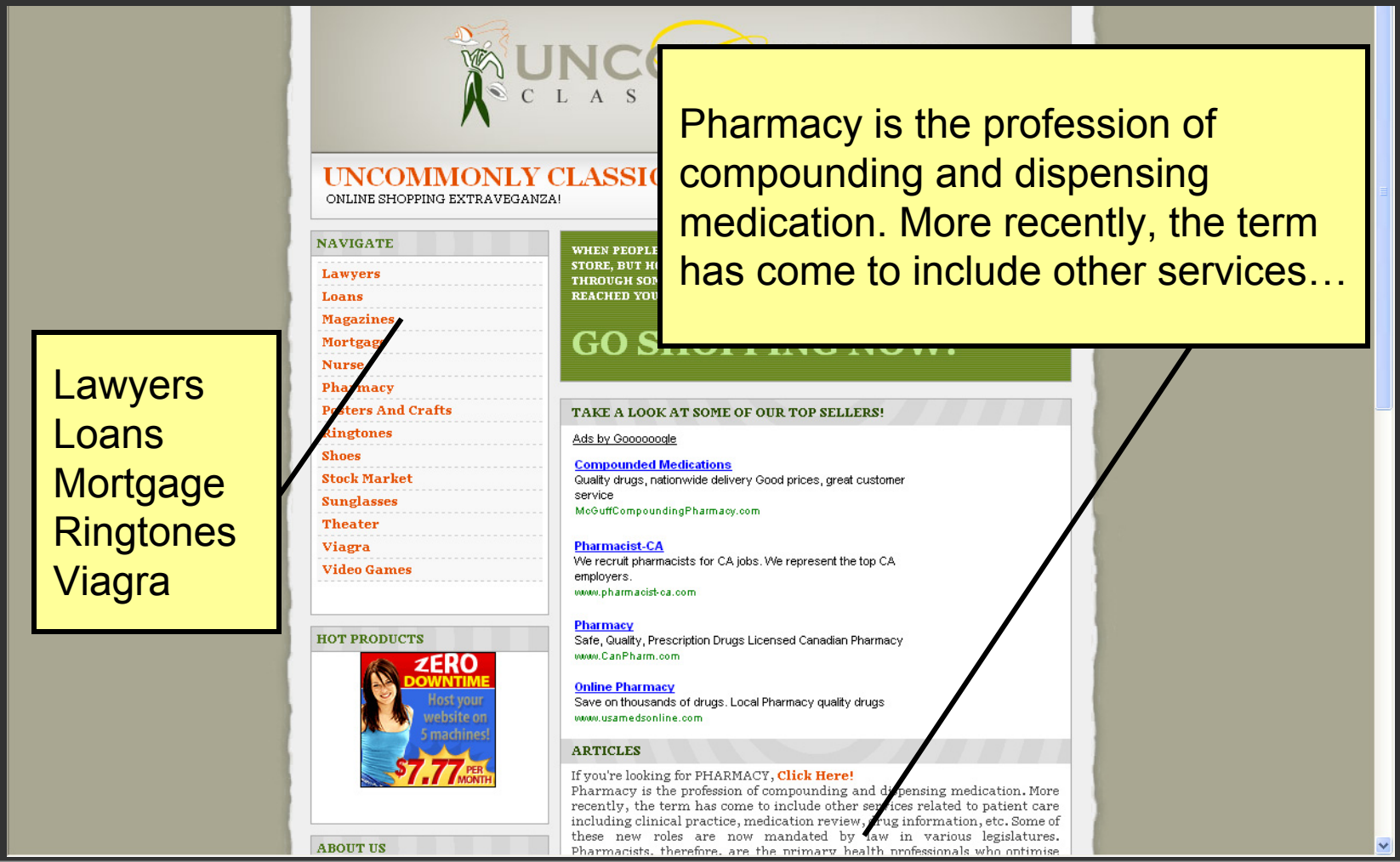
- Web spamming
- Link spamming
- PageRank contribution
- Spam mass
  - Definition
  - Estimation
  - Algorithm
- Experiments

Web Spamming: Example

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

Web Spamming: Example

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



Lawyers
Loans
Mortgage
Ringtones
Viagra

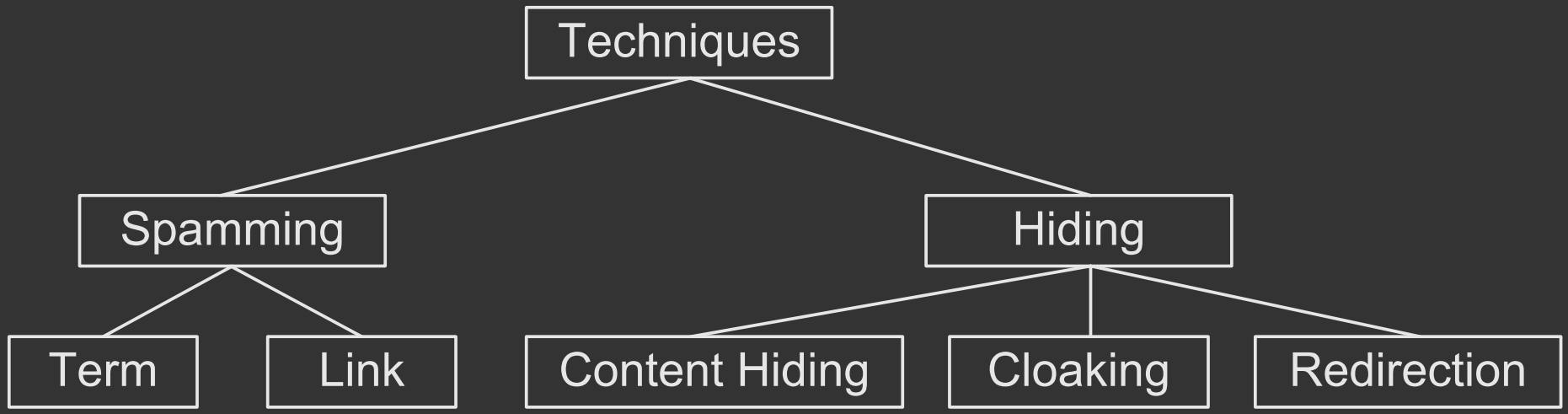
Pharmacy is the profession of compounding and dispensing medication. More recently, the term has come to include other services…

Web Spamming: Techniques

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

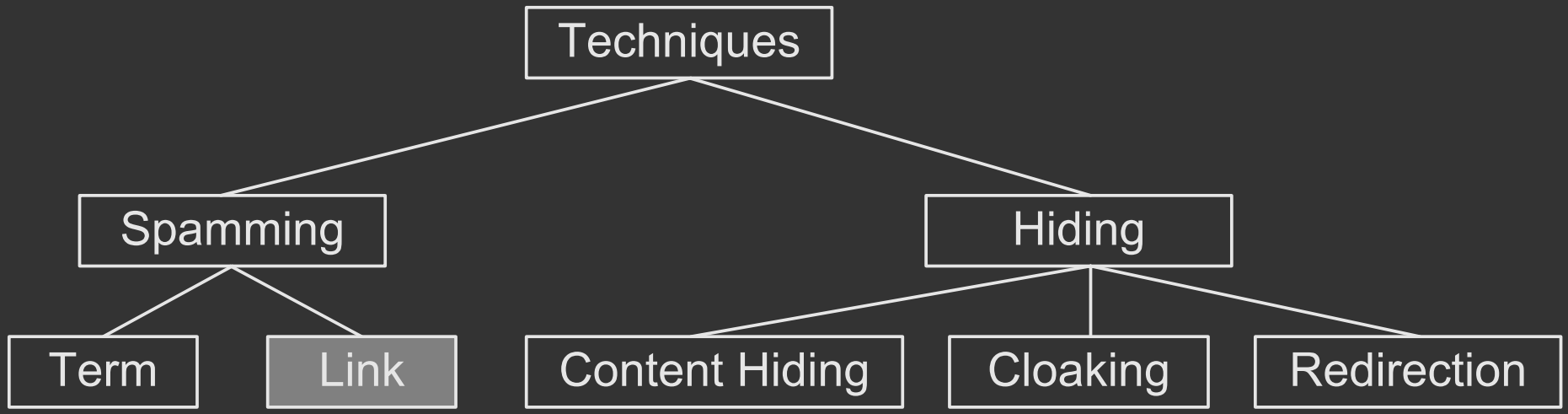
Web Spamming: Techniques

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



Web Spamming: Techniques

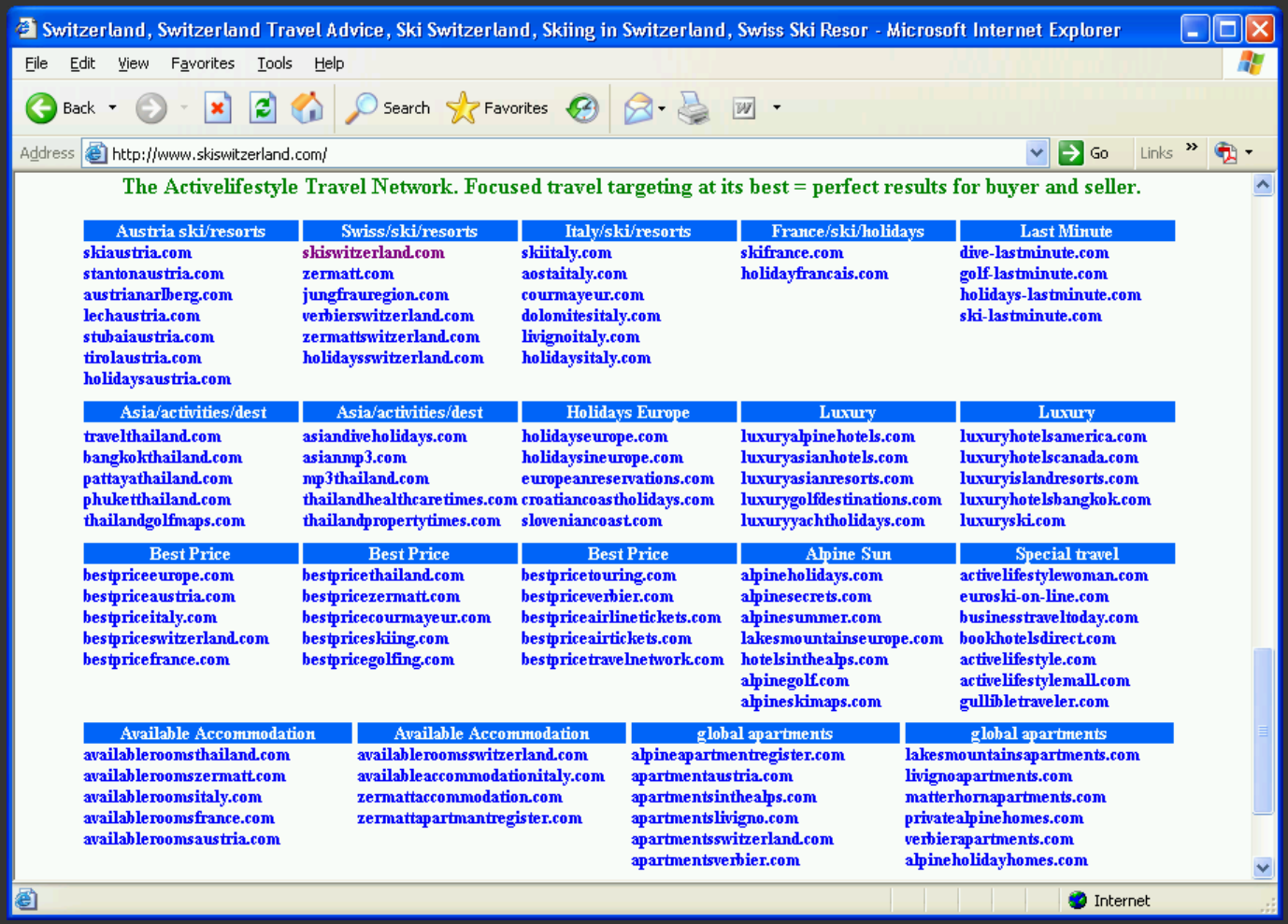
Spamming = misleading search engines to obtain higher-than-deserved ranking



Link spamming = building link structures that boost PageRank score

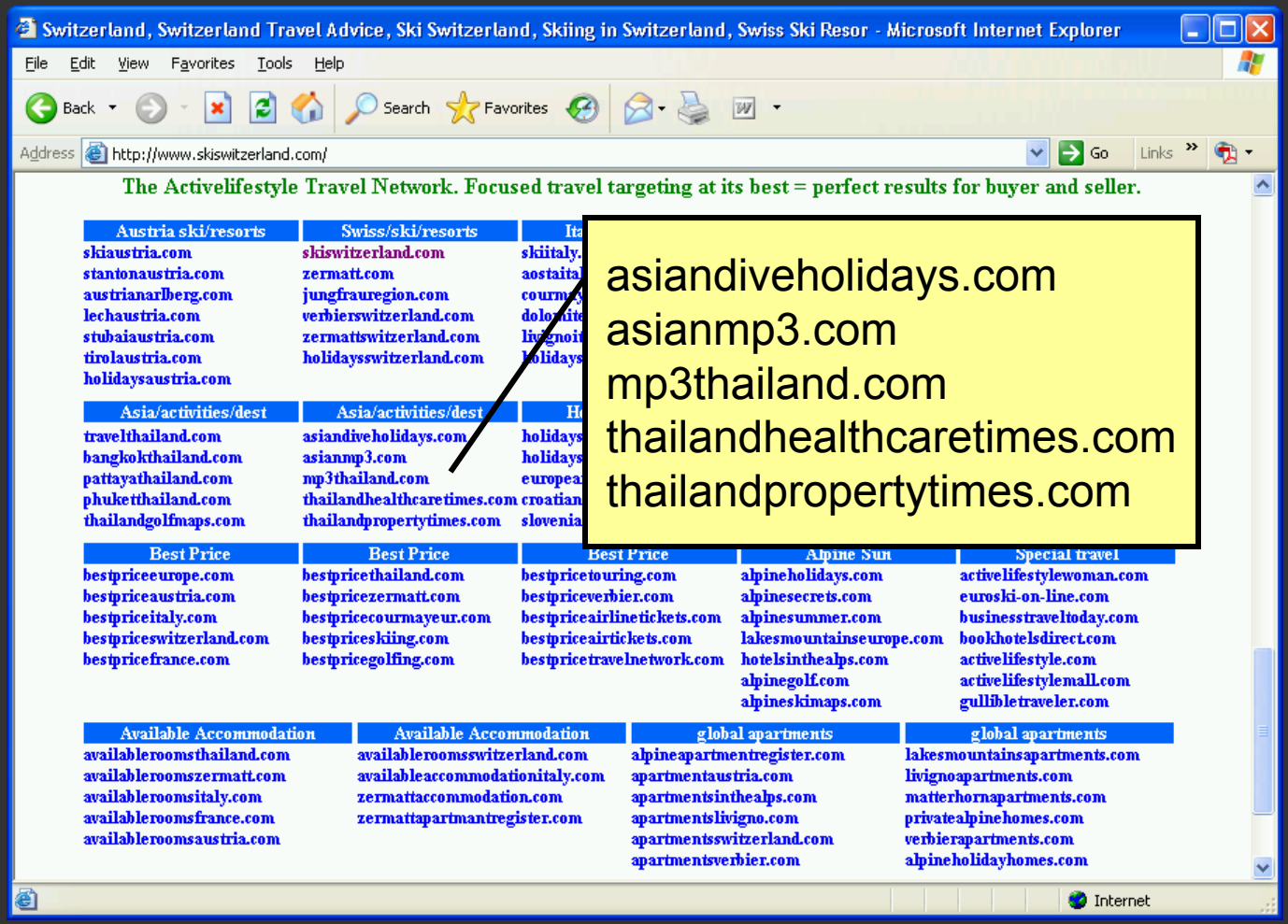
Link Spamming: Example

#1 search result for the query “switzerland ski”



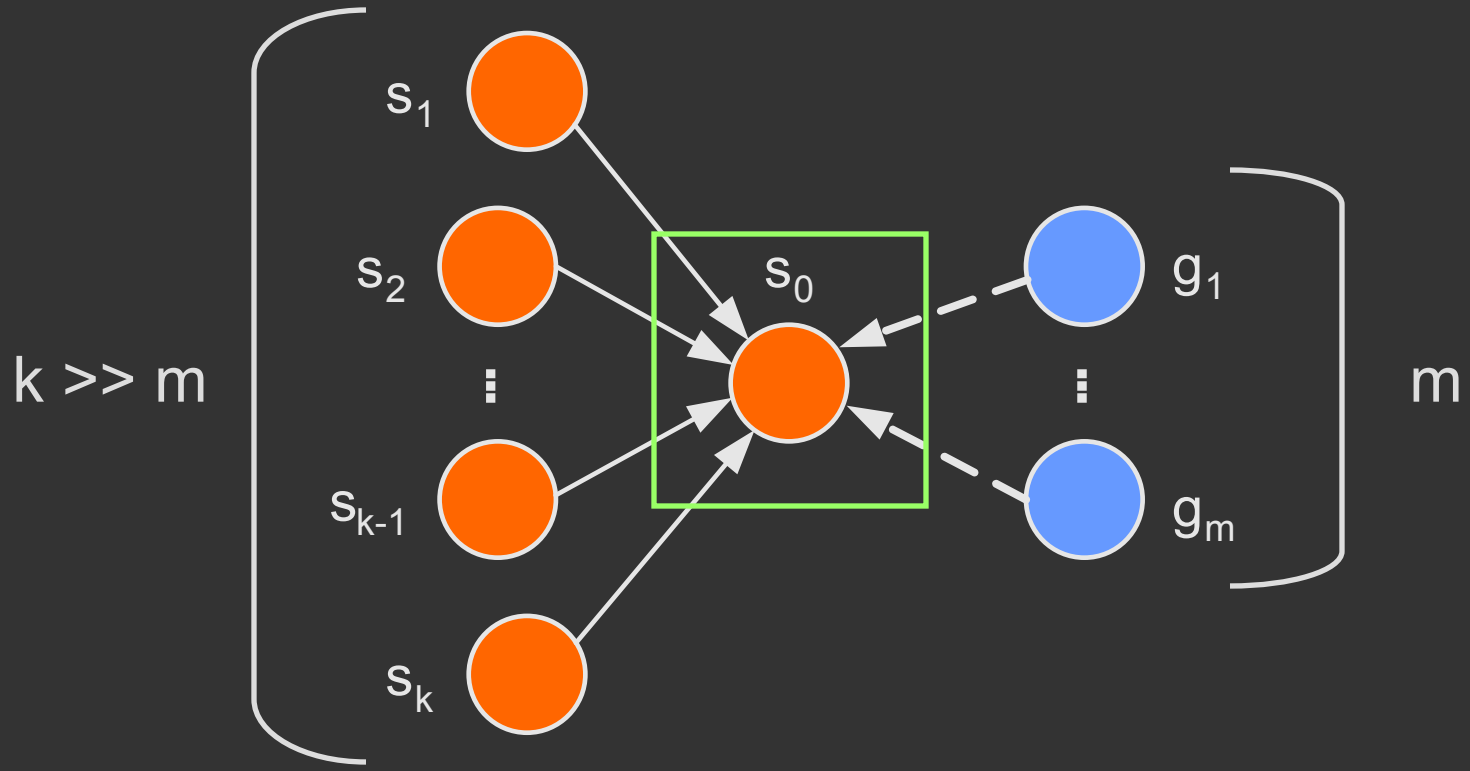
Link Spamming: Example

#1 search result for the query “switzerland ski”

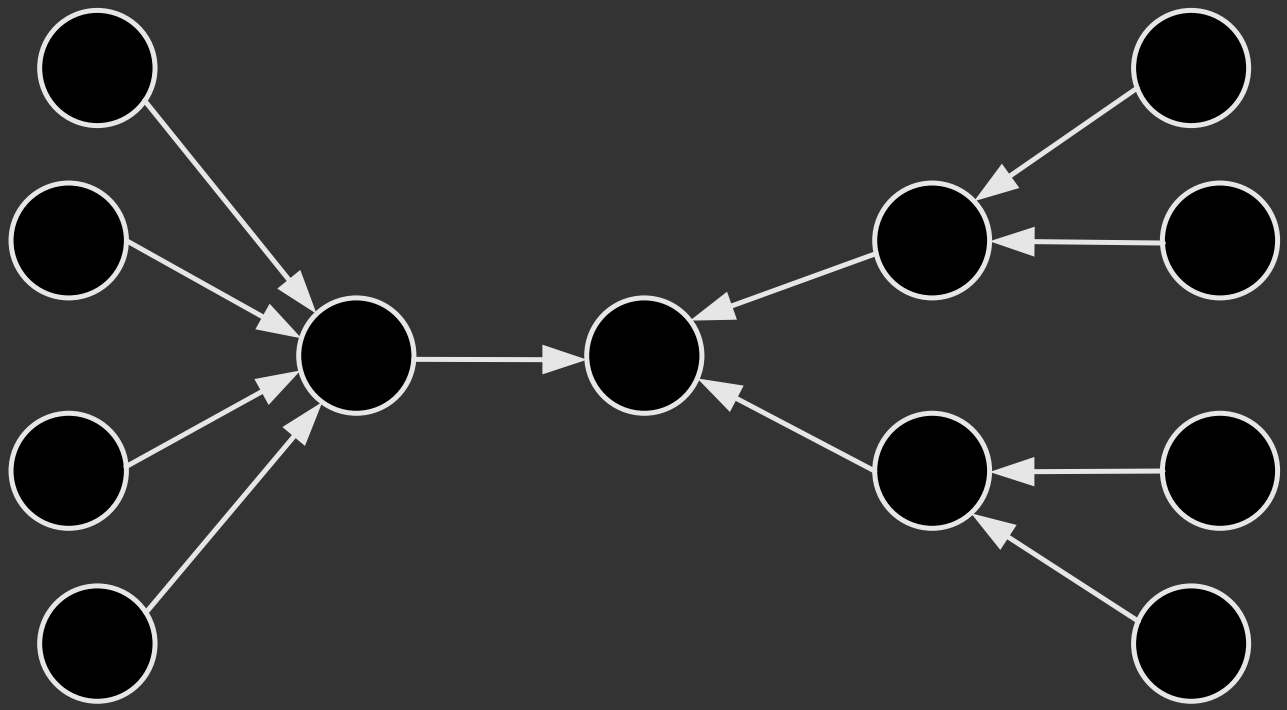


Link Spamming: Our Goal

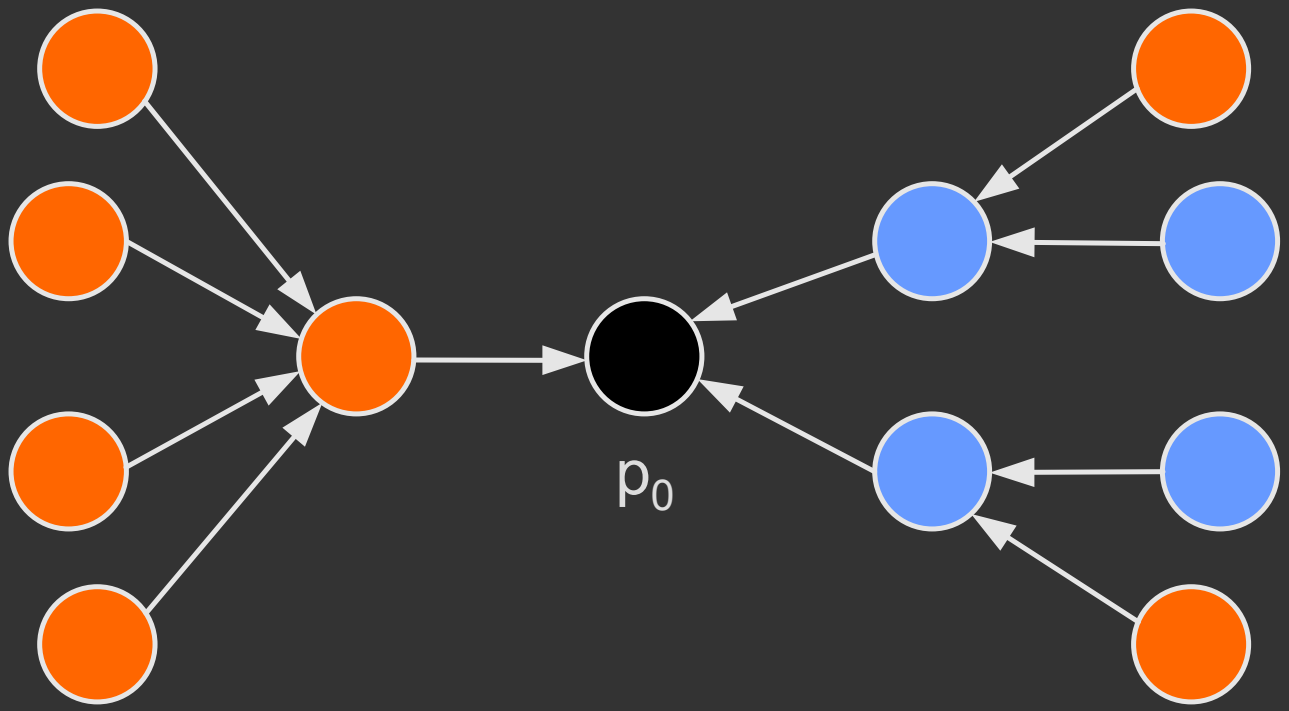
Detect pages that achieve high PageRank through link spamming



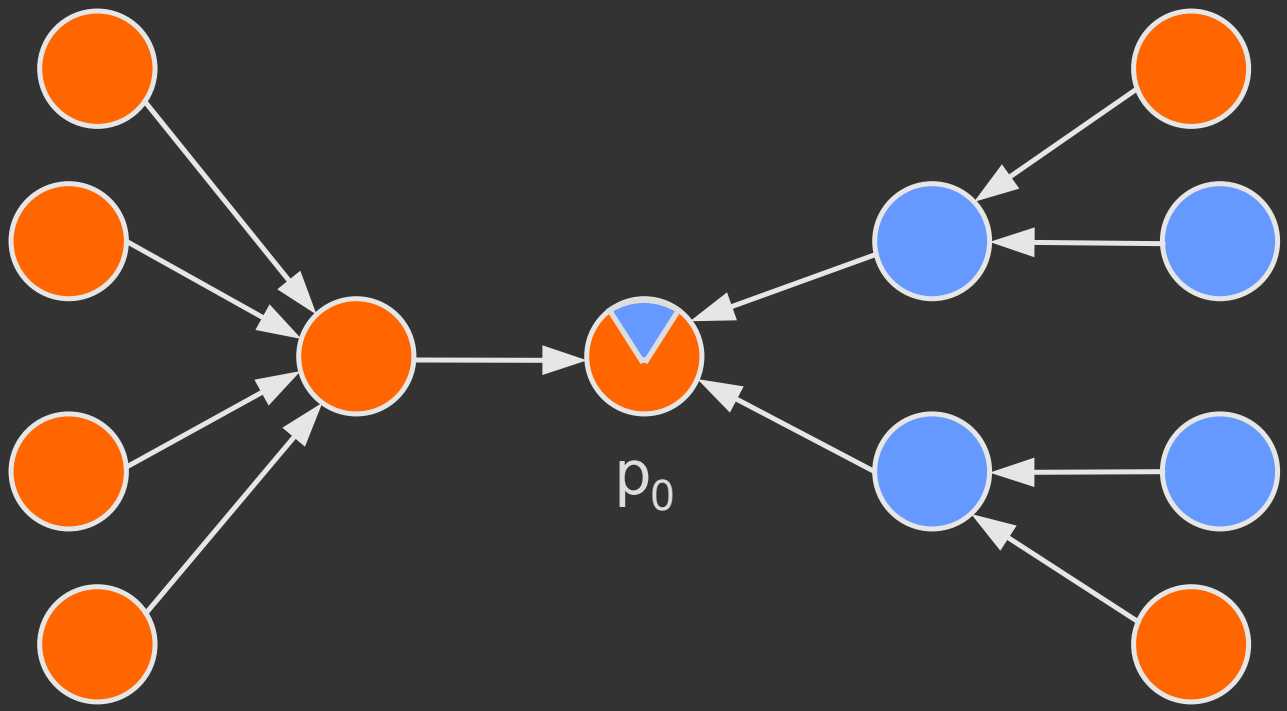
PageRank Contribution



PageRank Contribution



PageRank Contribution



$$p_0^+ = c^2\, 2\,(1 - c) / n + c\, 2\,(1 - c) / n$$

$$p_0^- = c^2\, 6\,(1 - c) / n + c\,(1 - c) / n$$

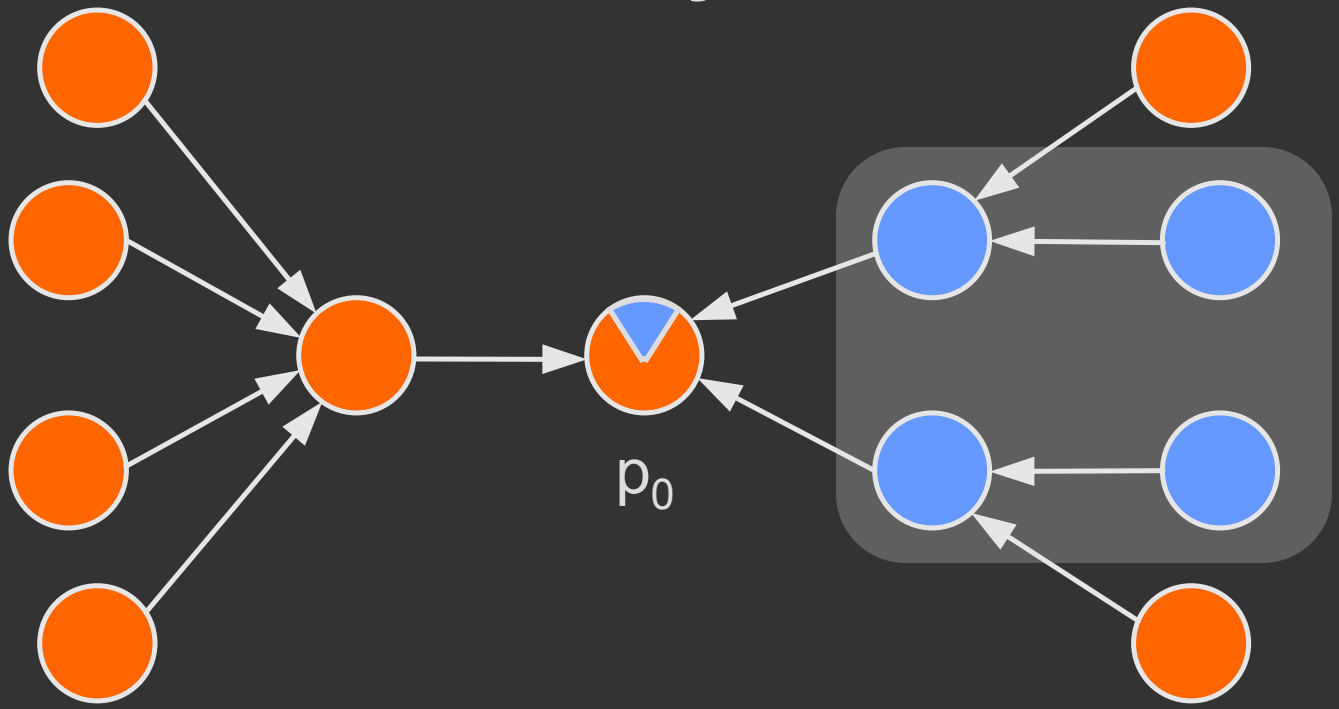
Spam Mass: Definition

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

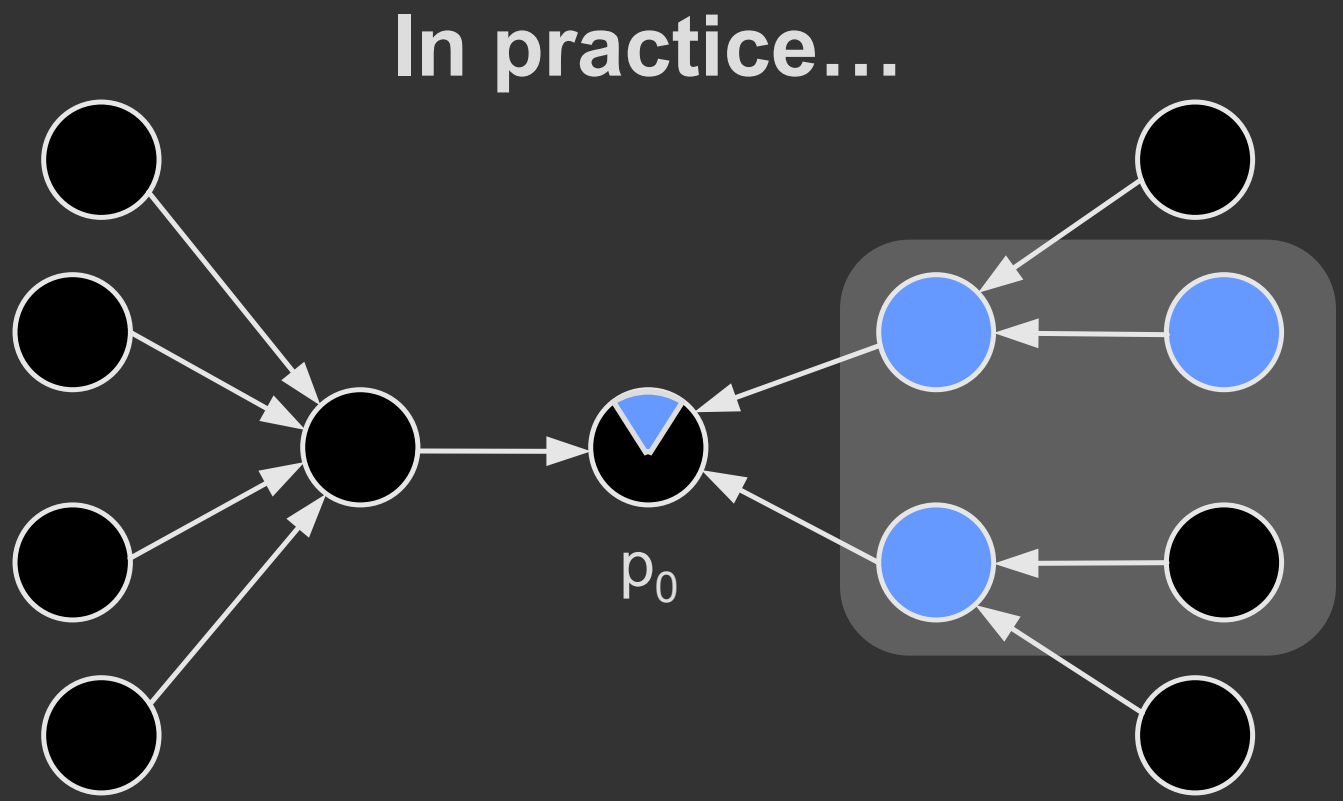
- Relative mass
  - **Fraction** of PageRank coming from spam
  - More useful in practice

Spam Mass: Estimation

Ideally…



Spam Mass: Estimation



- 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 > threshold

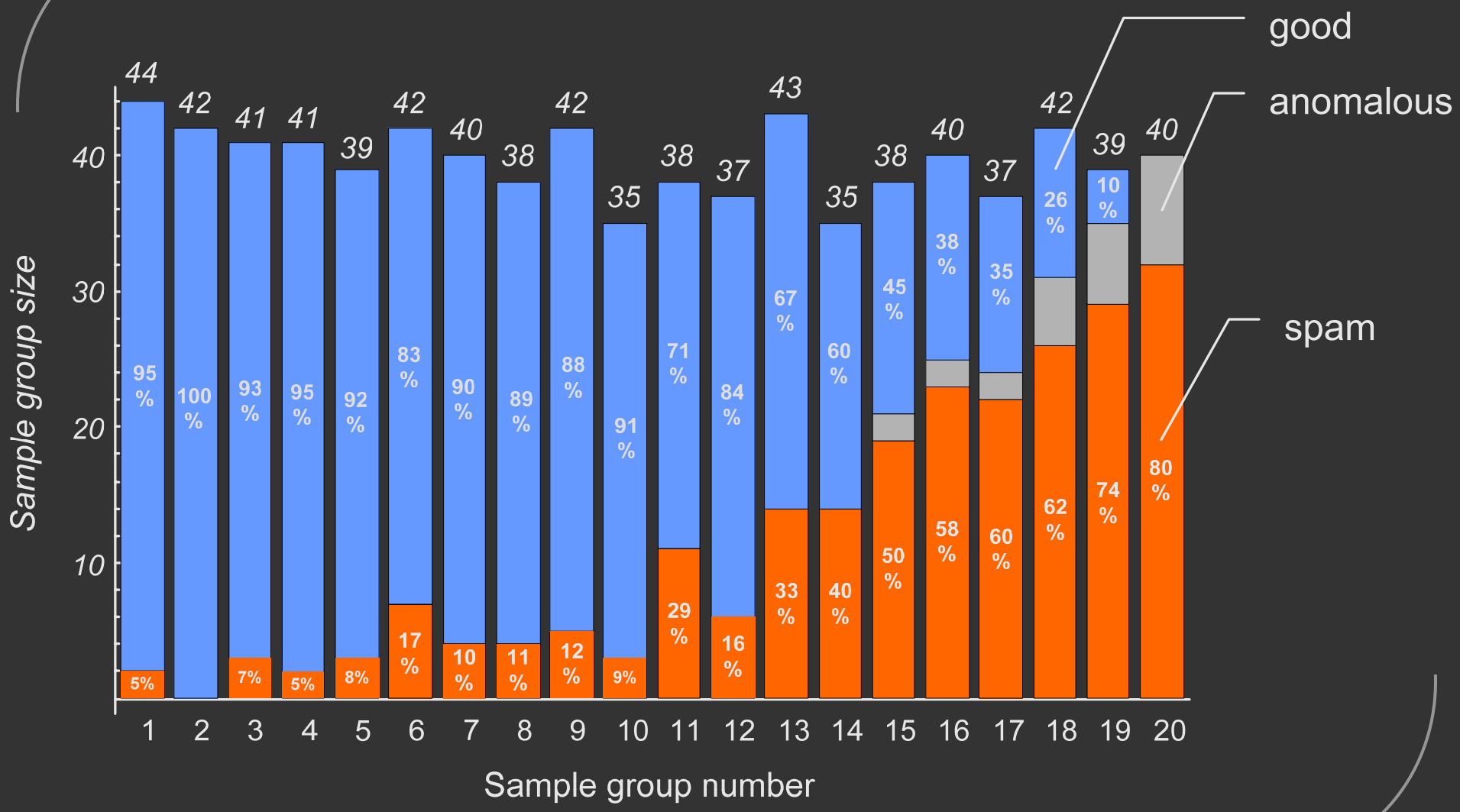
Experiments: Data

- 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

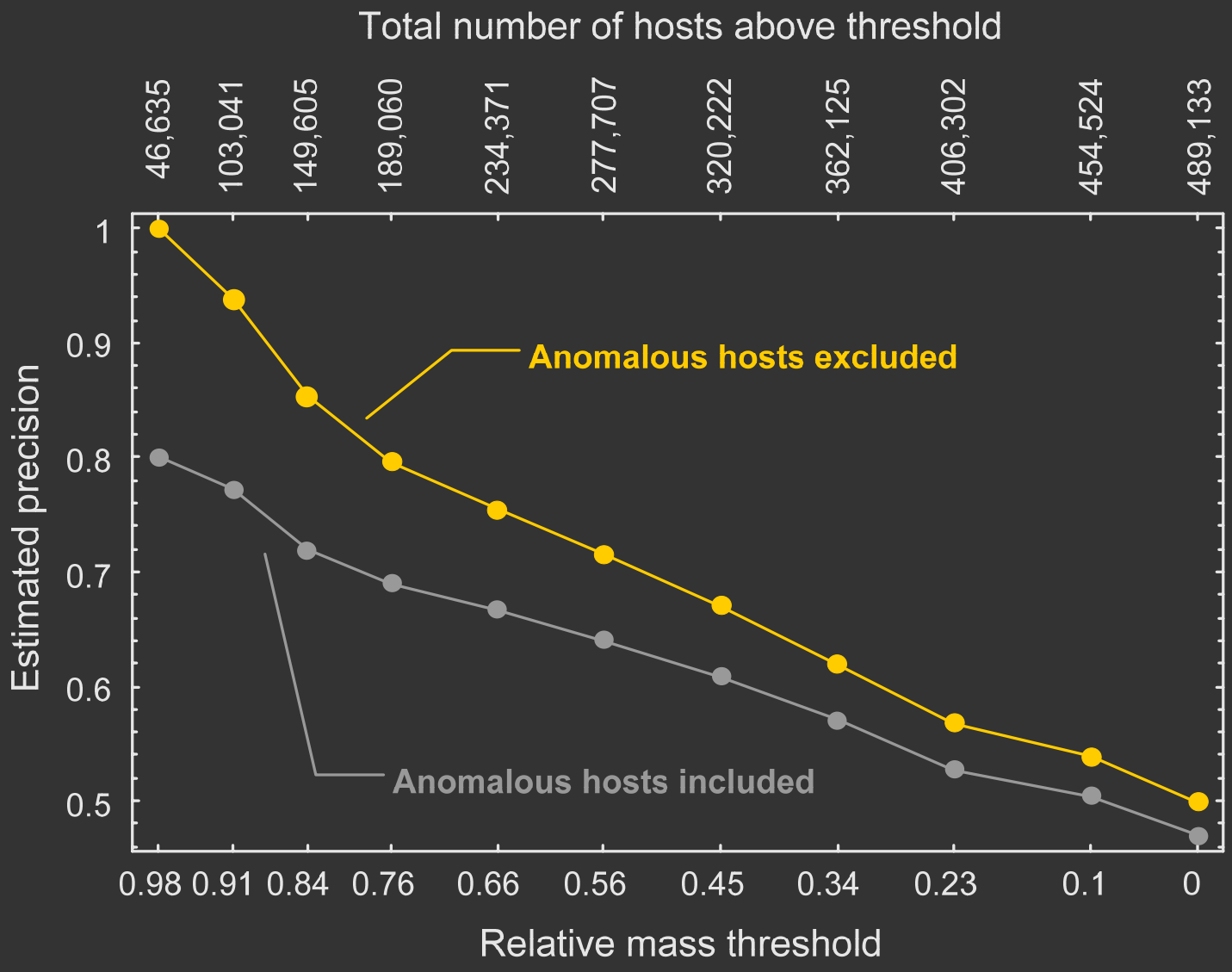
Experiments: Relative Mass



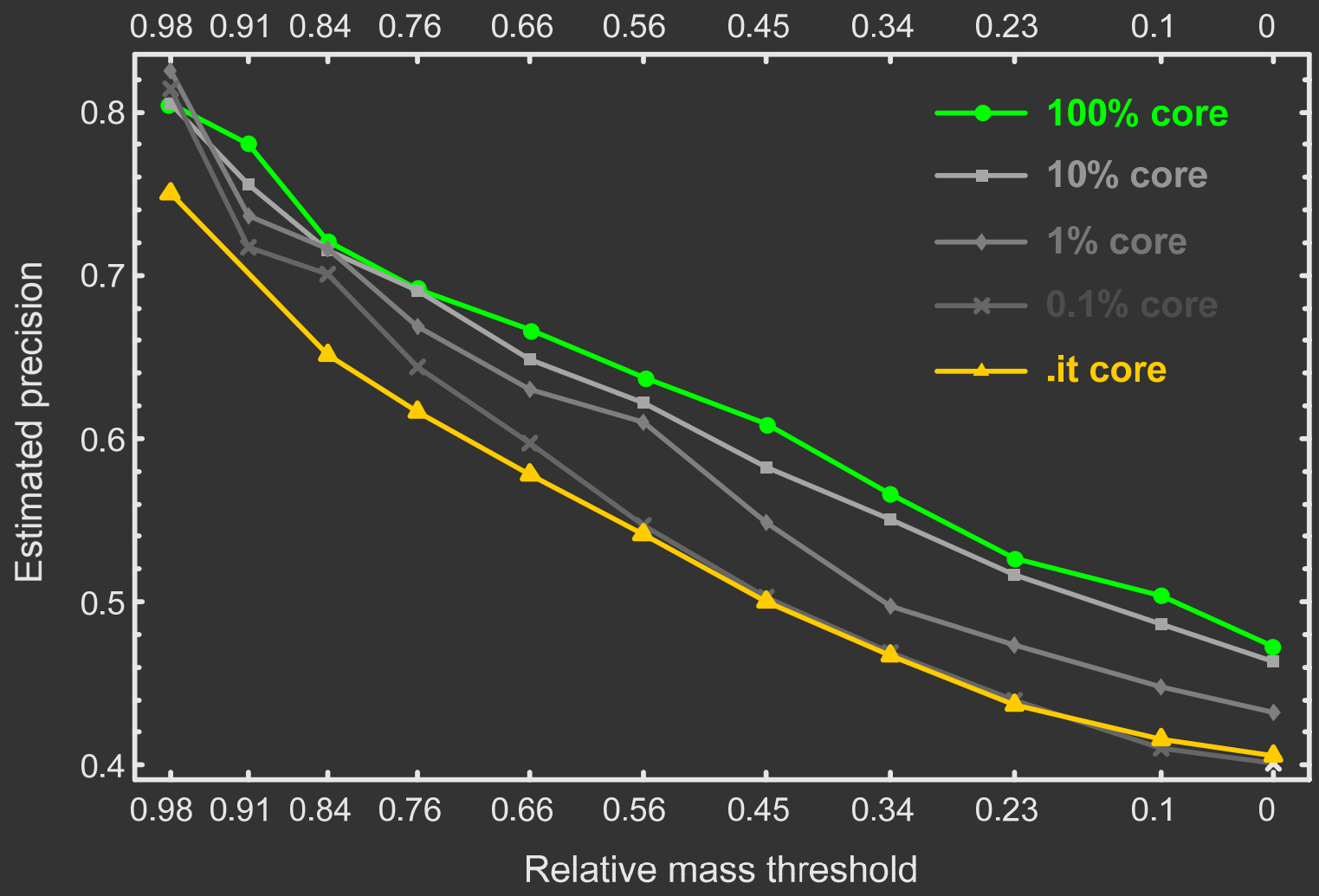
Experiments: Relative Mass

- 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