Link Spam Detection Based on Mass Estimation

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

Roadmap

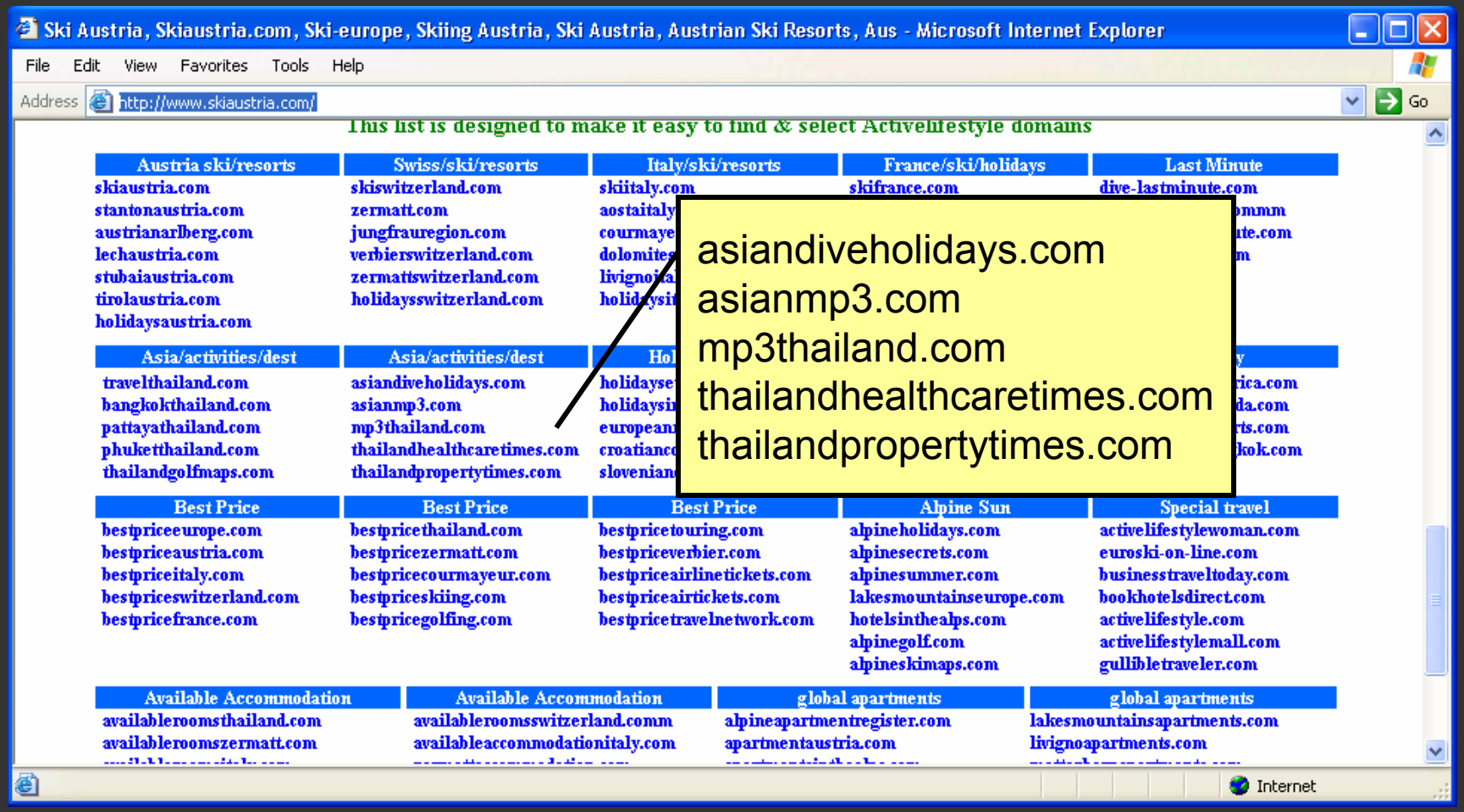
- Search engine spamming
- Link spamming
- PageRank contribution
- Spam mass
  - Definition
  - Estimation
  - Algorithm
- Experiments

Spamming: Example

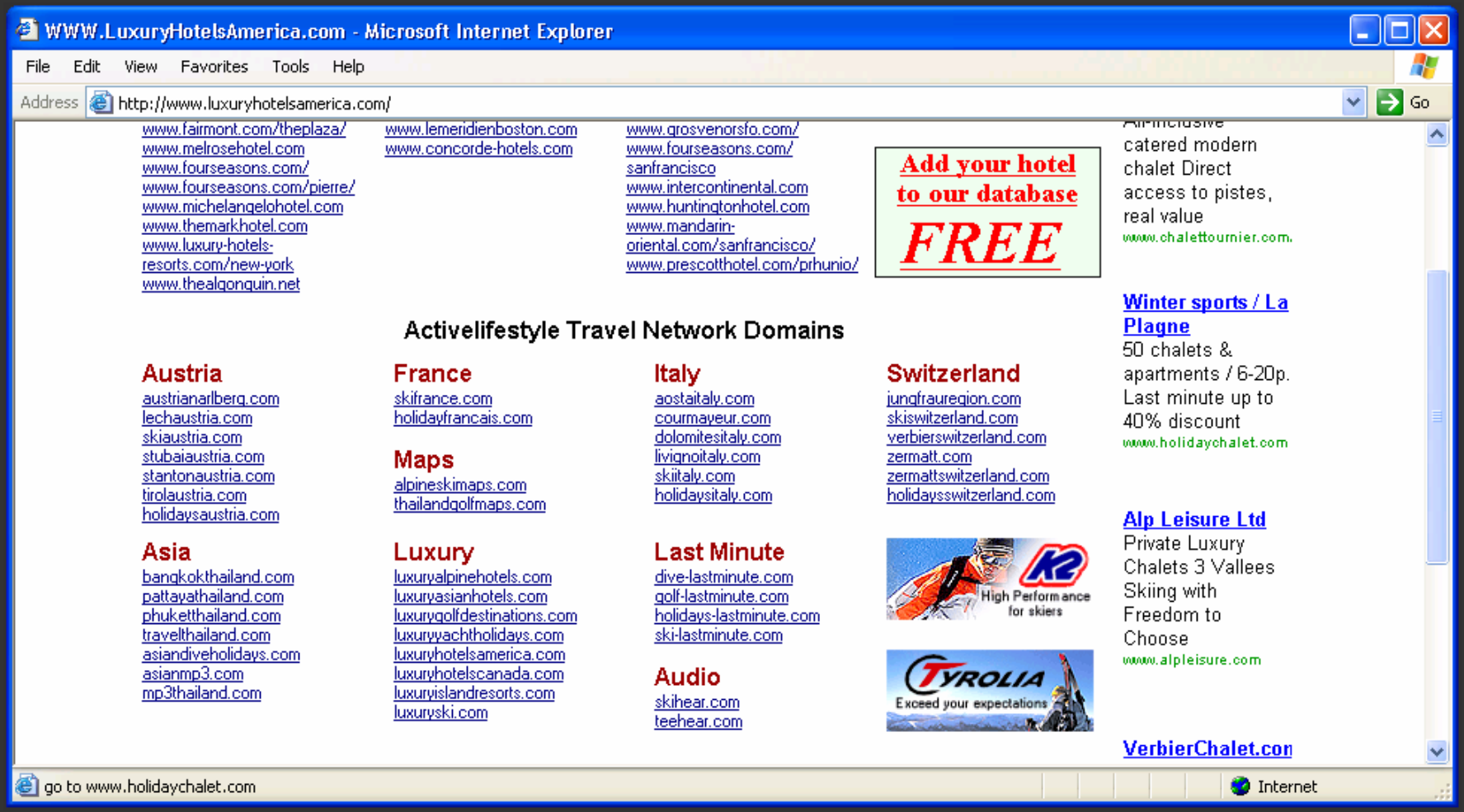
#1 search result for the query “austria ski”

Spamming: Example

#1 search result for the query “austria ski”



Spamming: Example

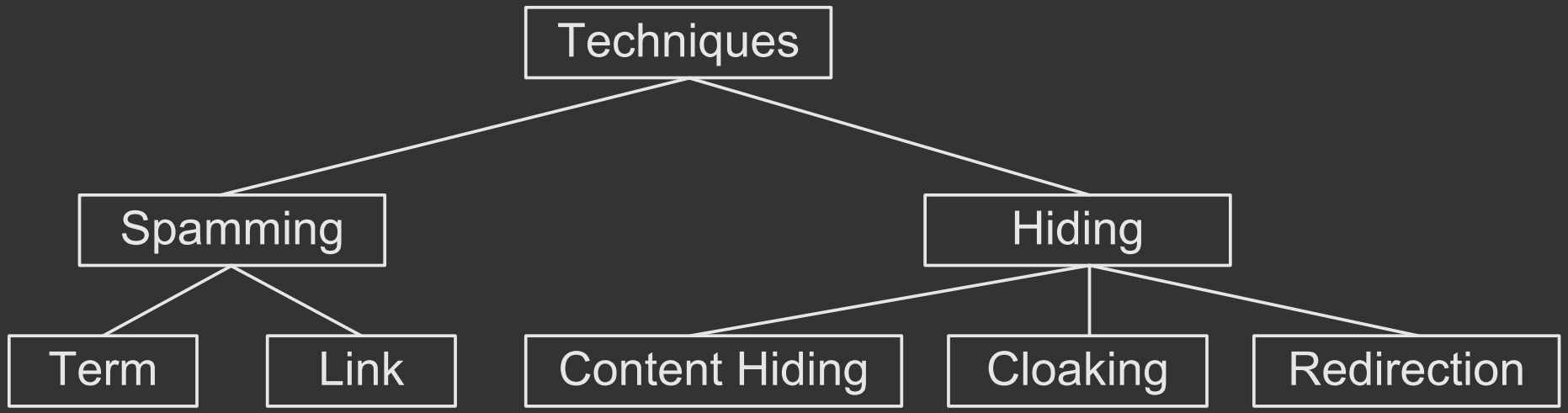


Spamming: Introduction

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

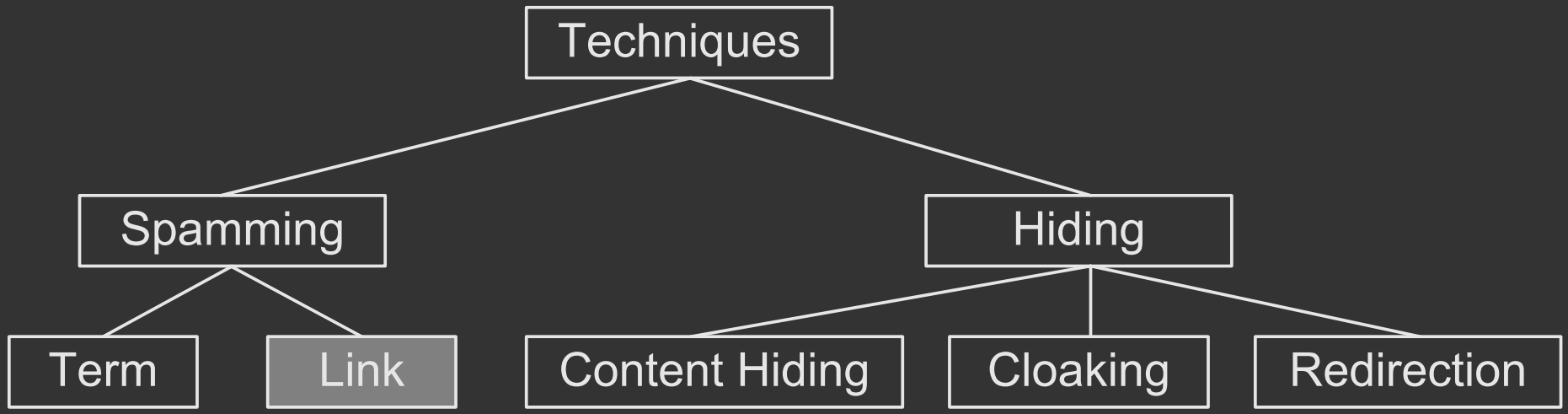
Spamming: Introduction

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



Spamming: Introduction

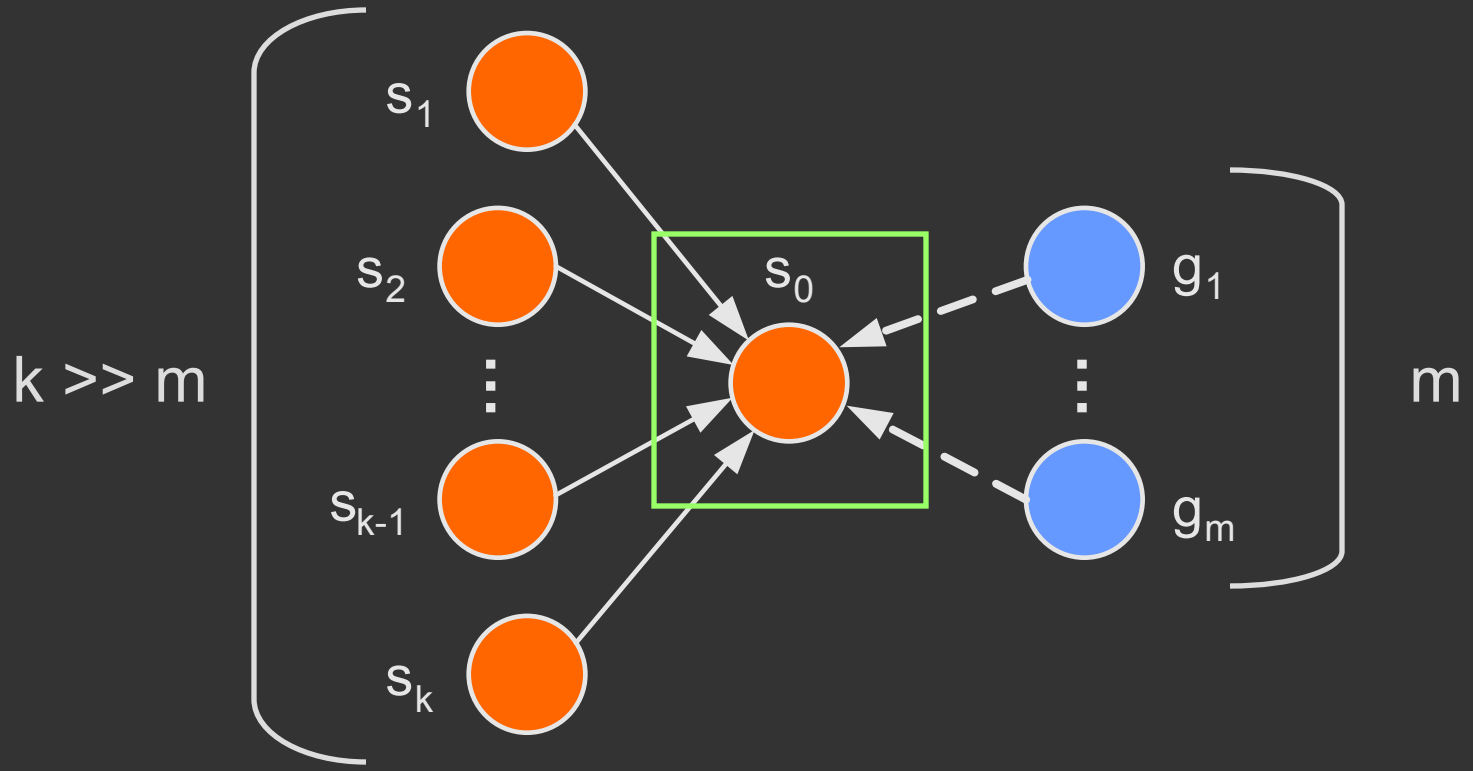
Spamming = misleading search engines to obtain higher-than-deserved ranking



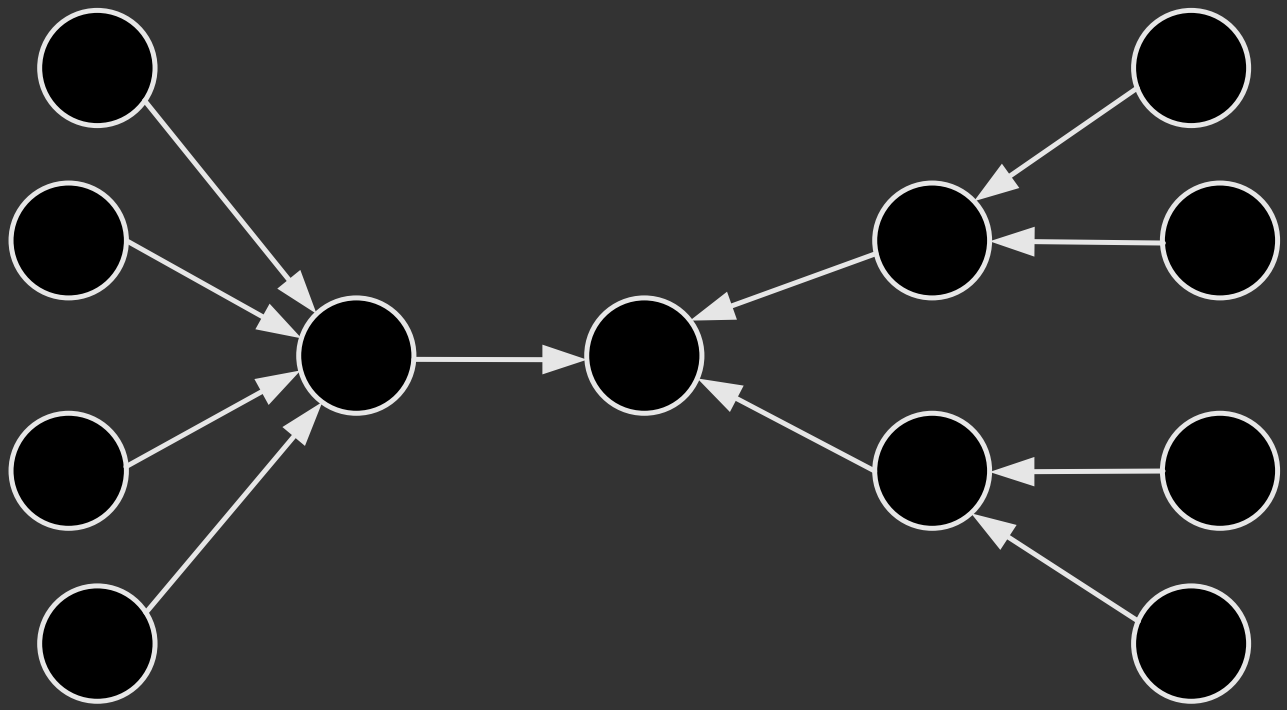
Link spamming = building link structures that boost PageRank score

Spamming: Our Target

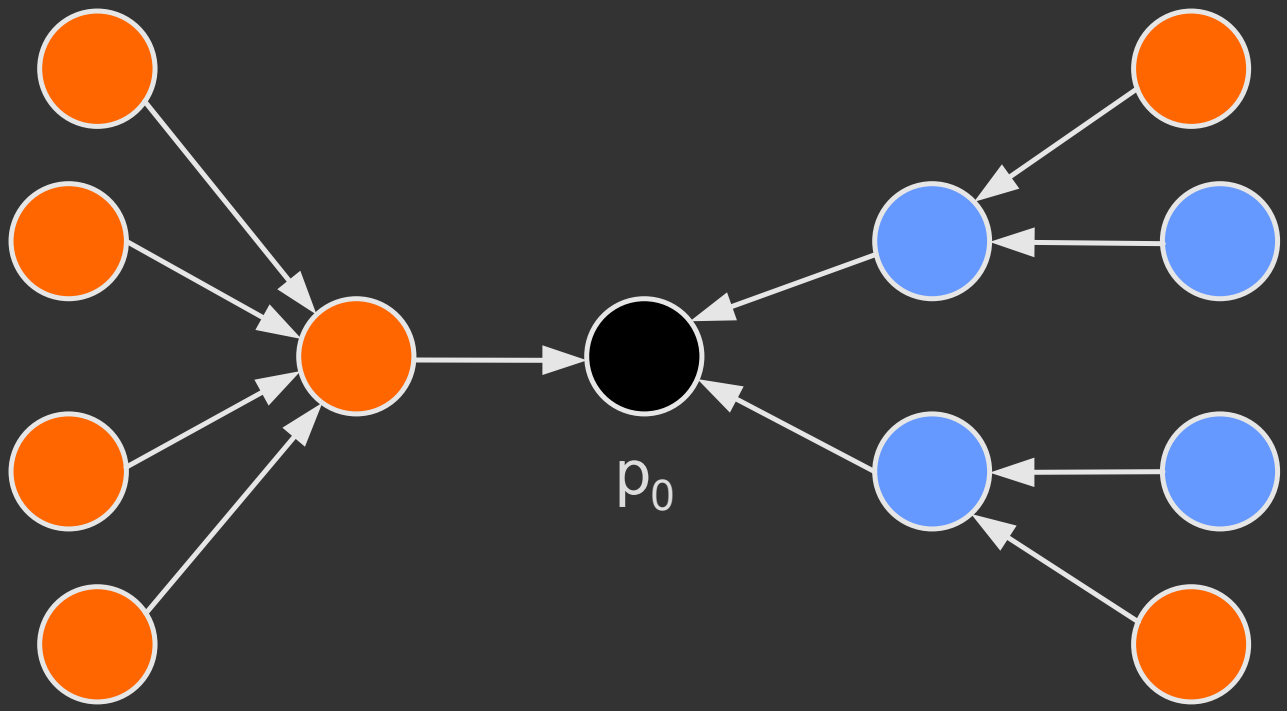
Detect pages that achieve high PageRank through link spamming



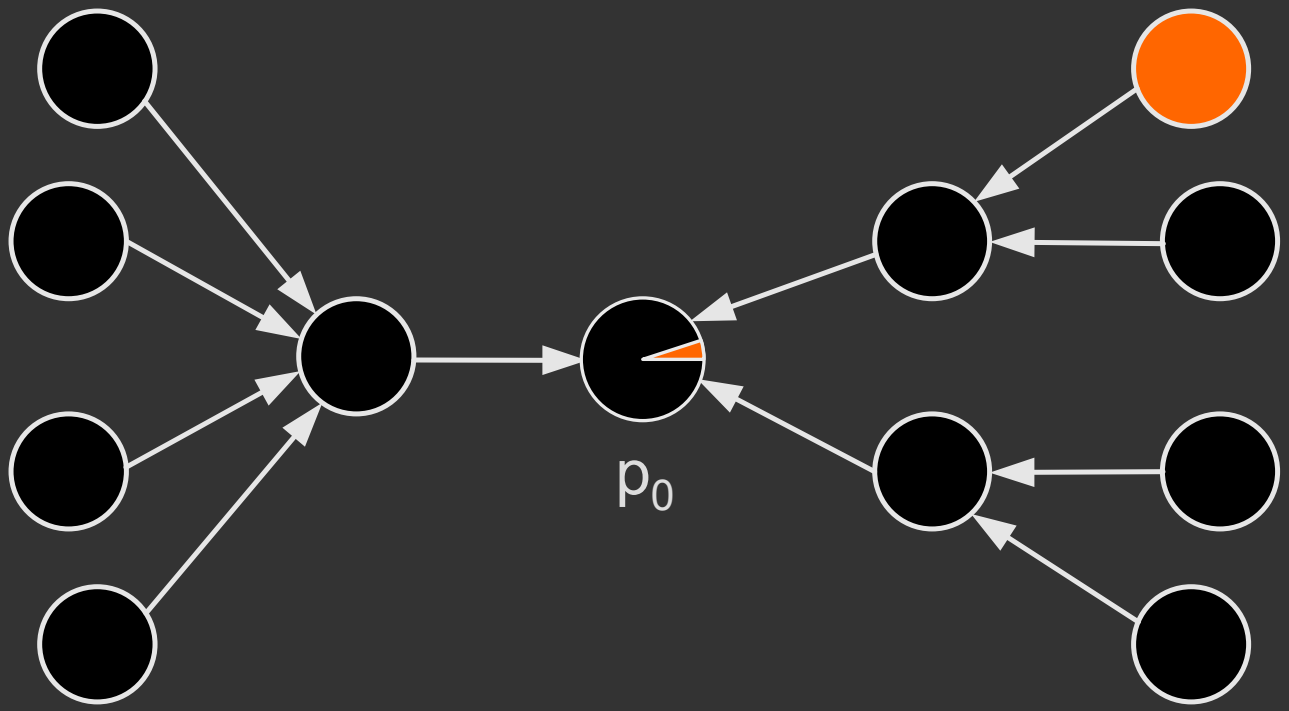
PageRank Contribution



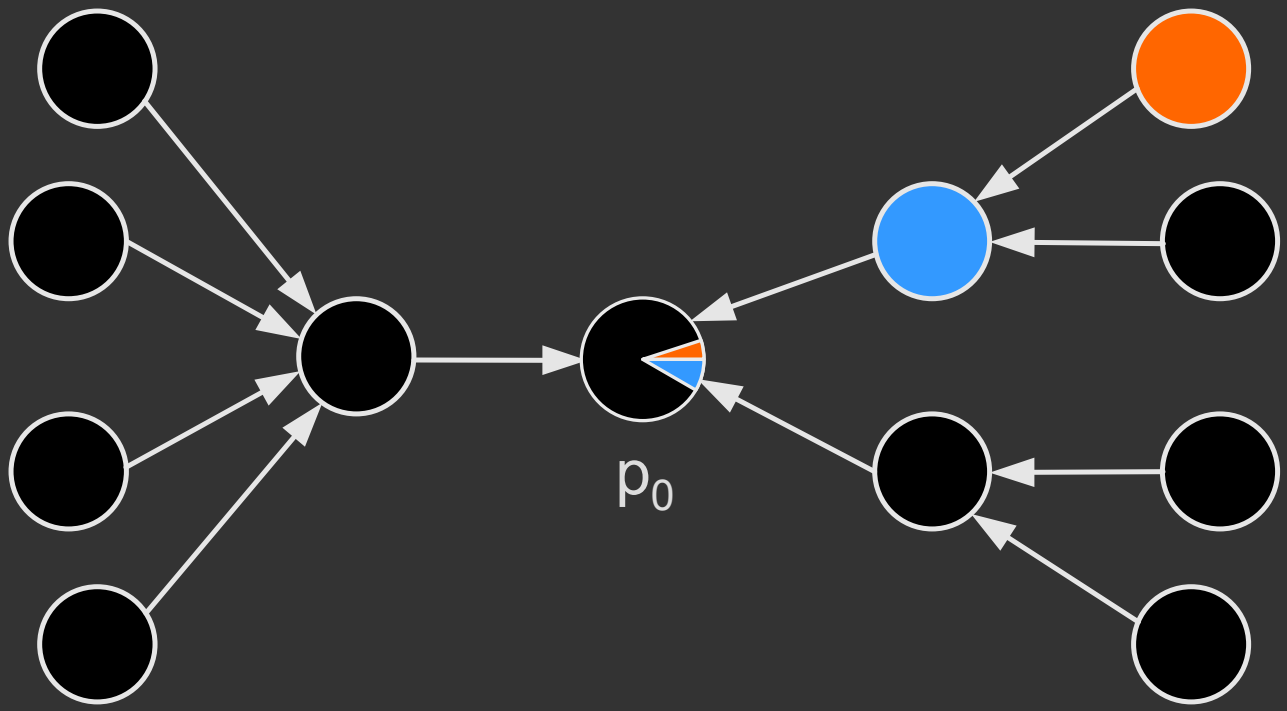
PageRank Contribution



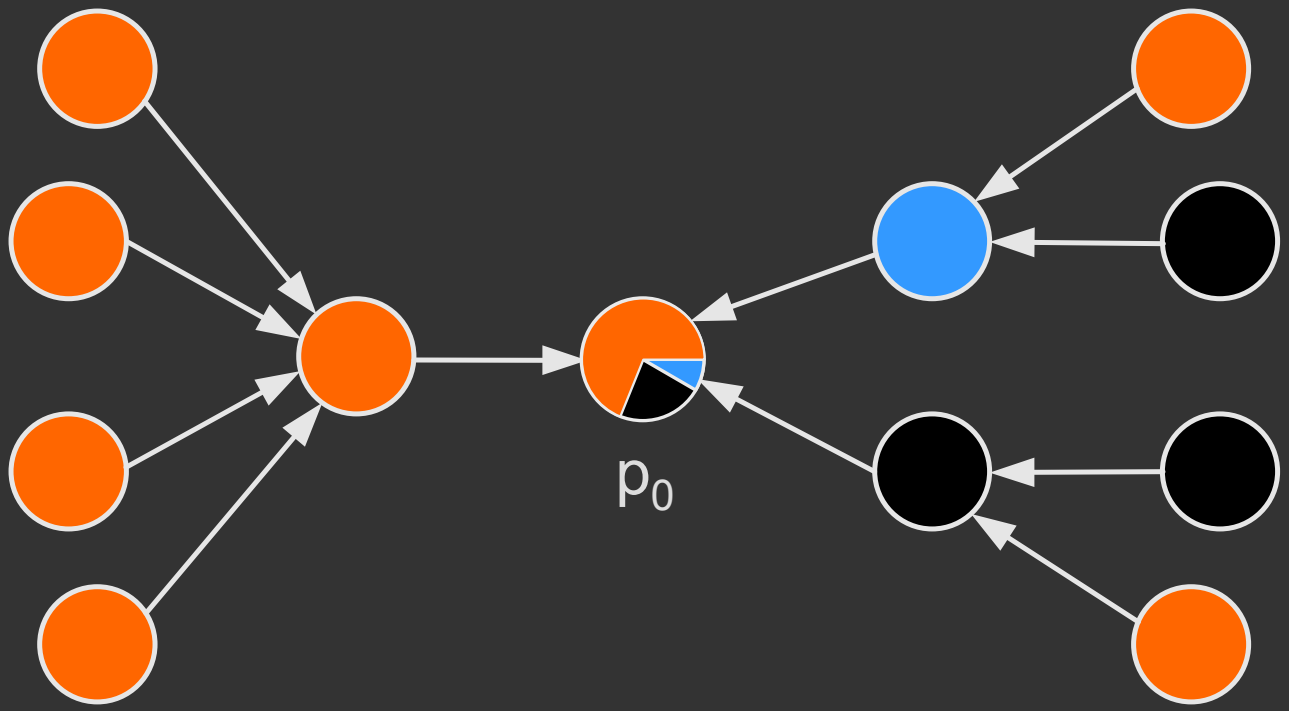
PageRank Contribution



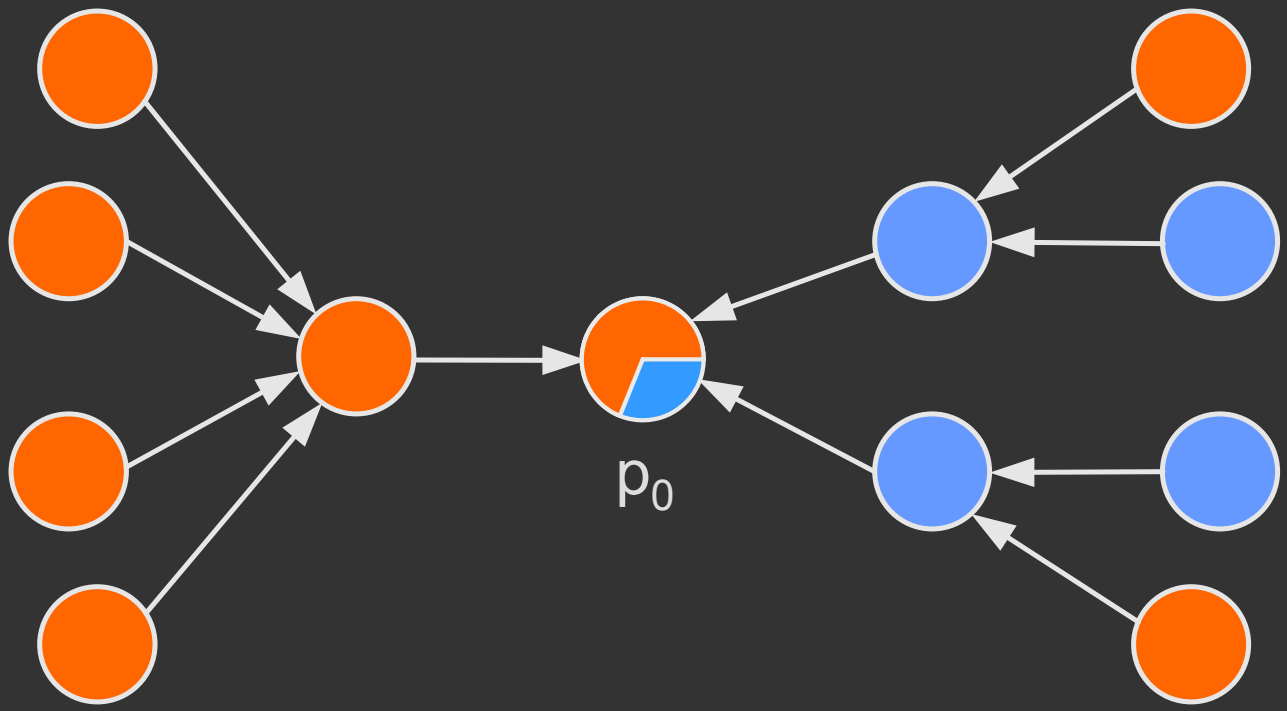
PageRank Contribution



PageRank Contribution



PageRank Contribution



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

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

Spam Mass: Definition

- Absolute mass
  - **Amount** (part) of PageRank coming from spam

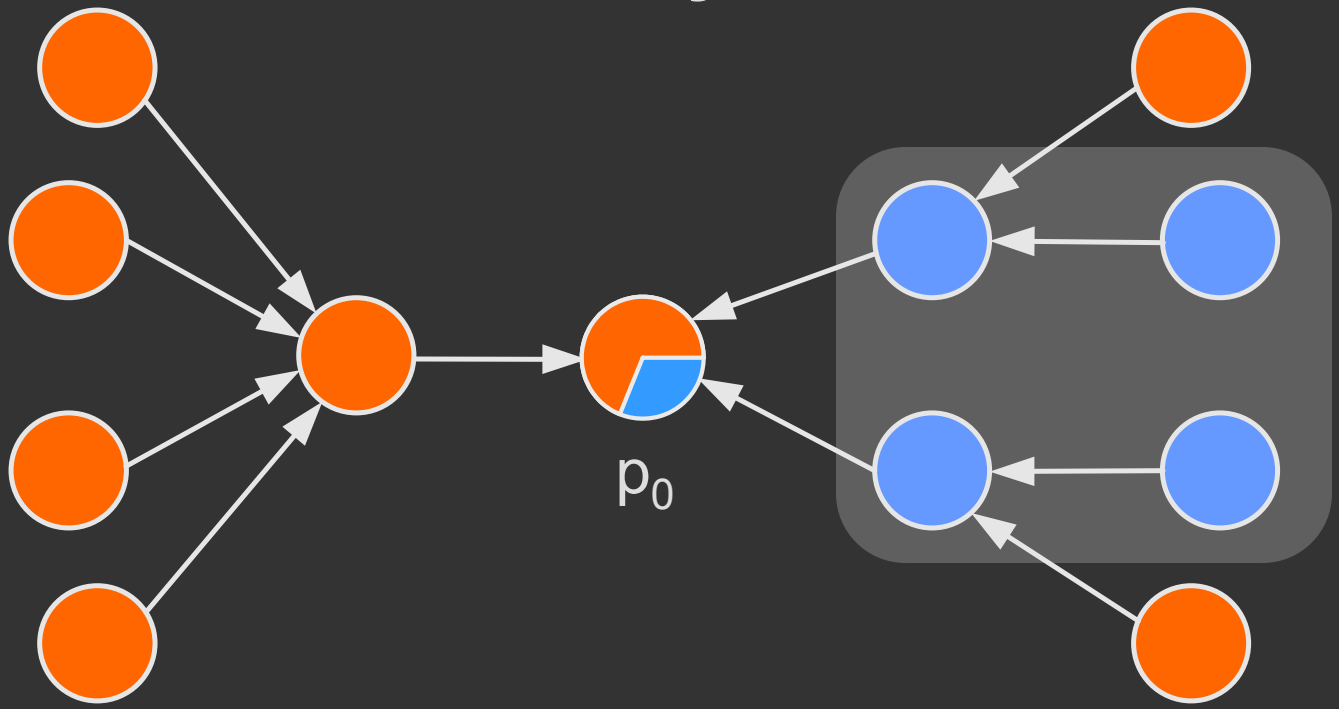
$$a.m. = p_0^- = 5$$

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

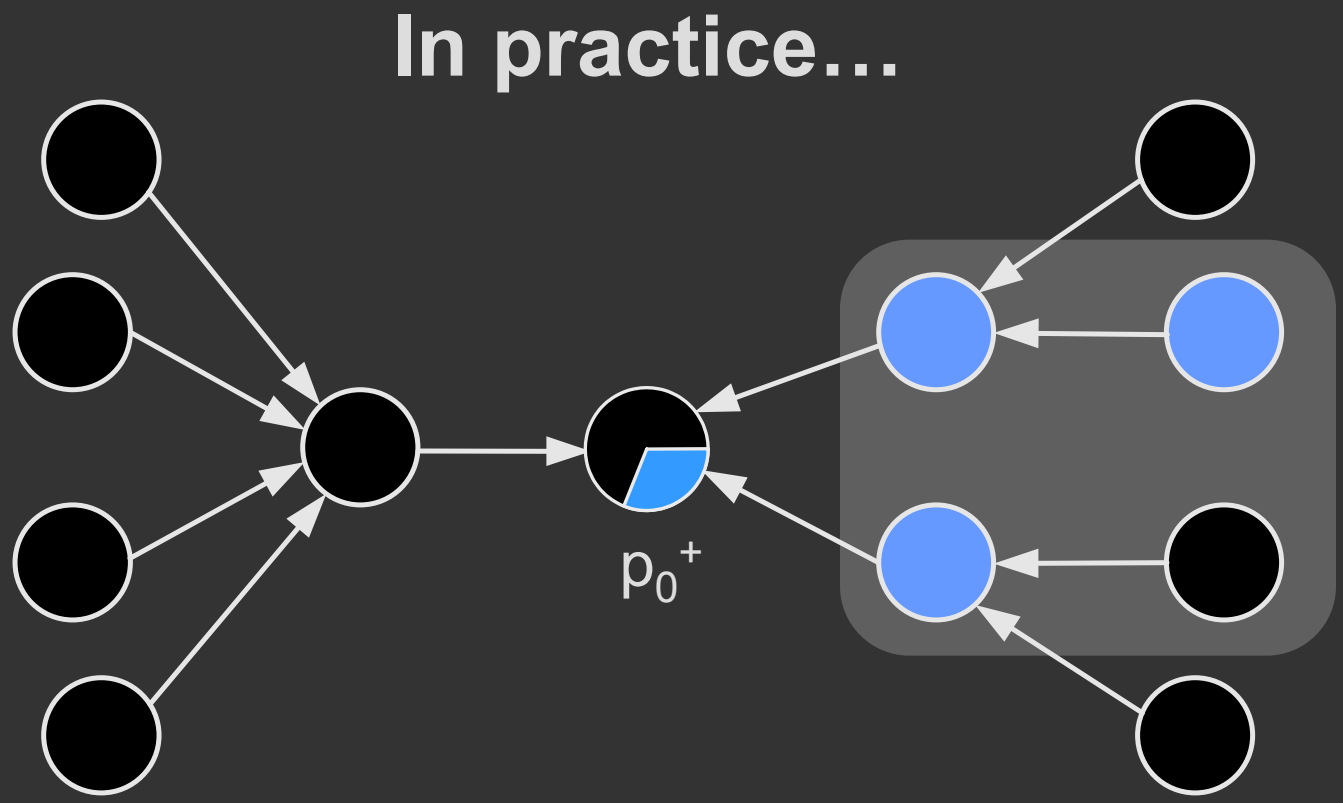
$$r.m. = \frac{p_0^-}{p_0} = \frac{5}{7}$$

Spam Mass: Estimation

Ideally…

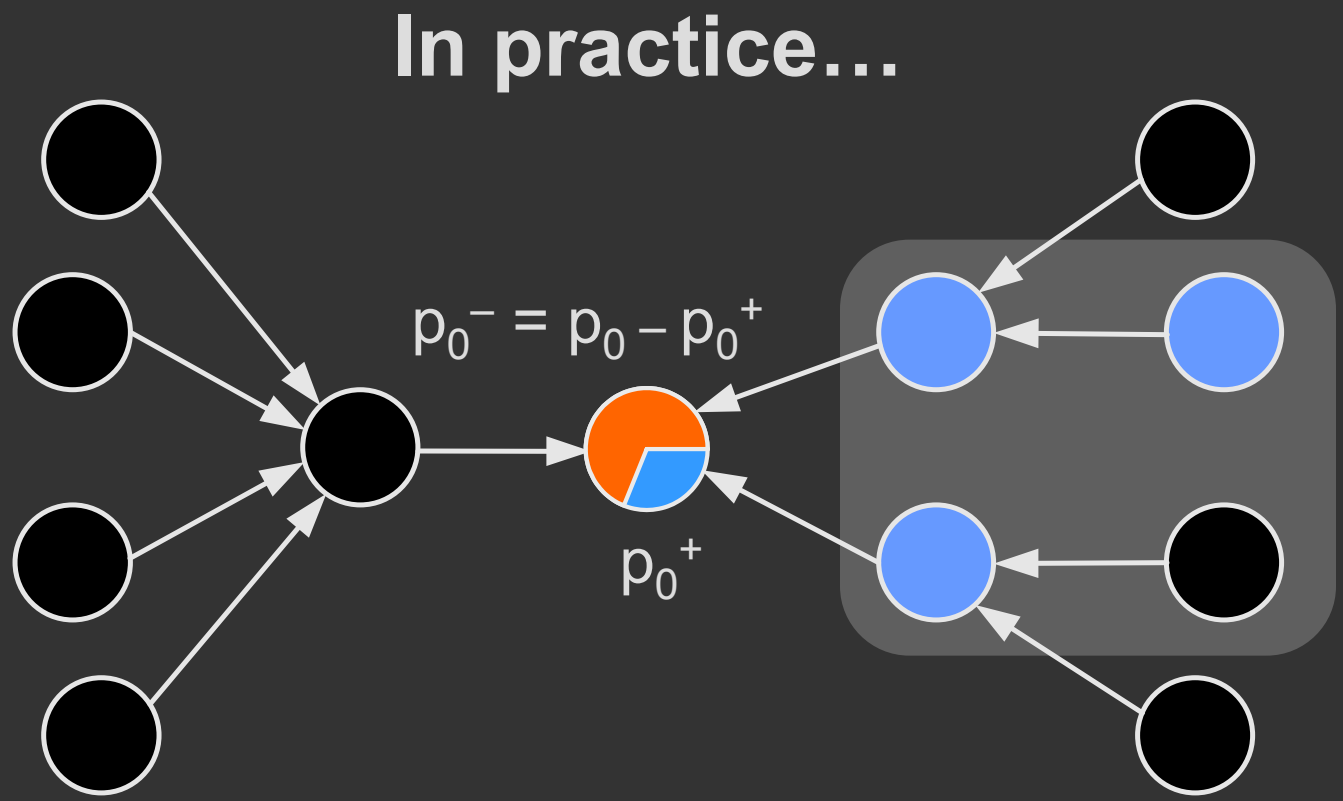


Spam Mass: Estimation



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

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

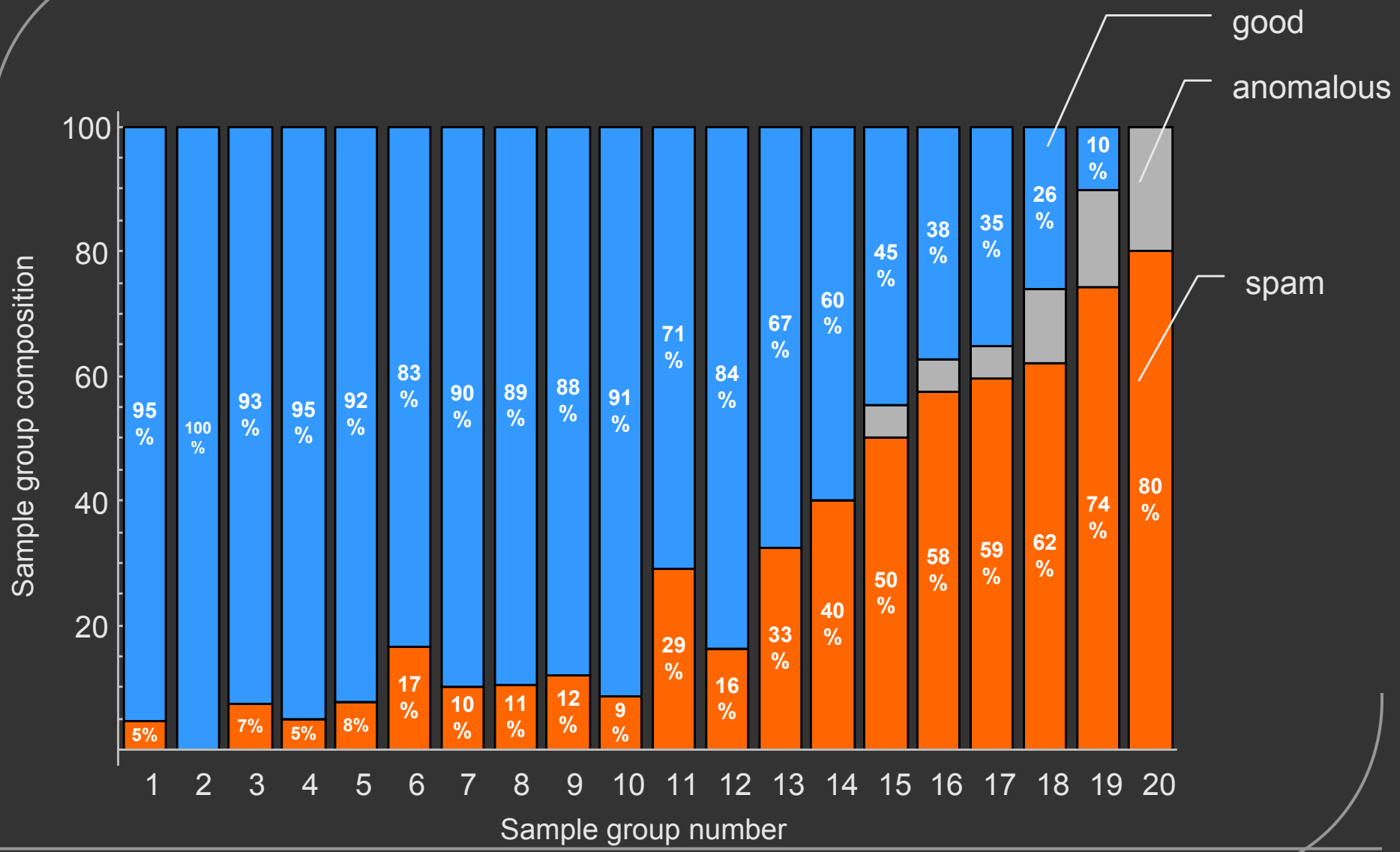
Experiments: Data

- 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

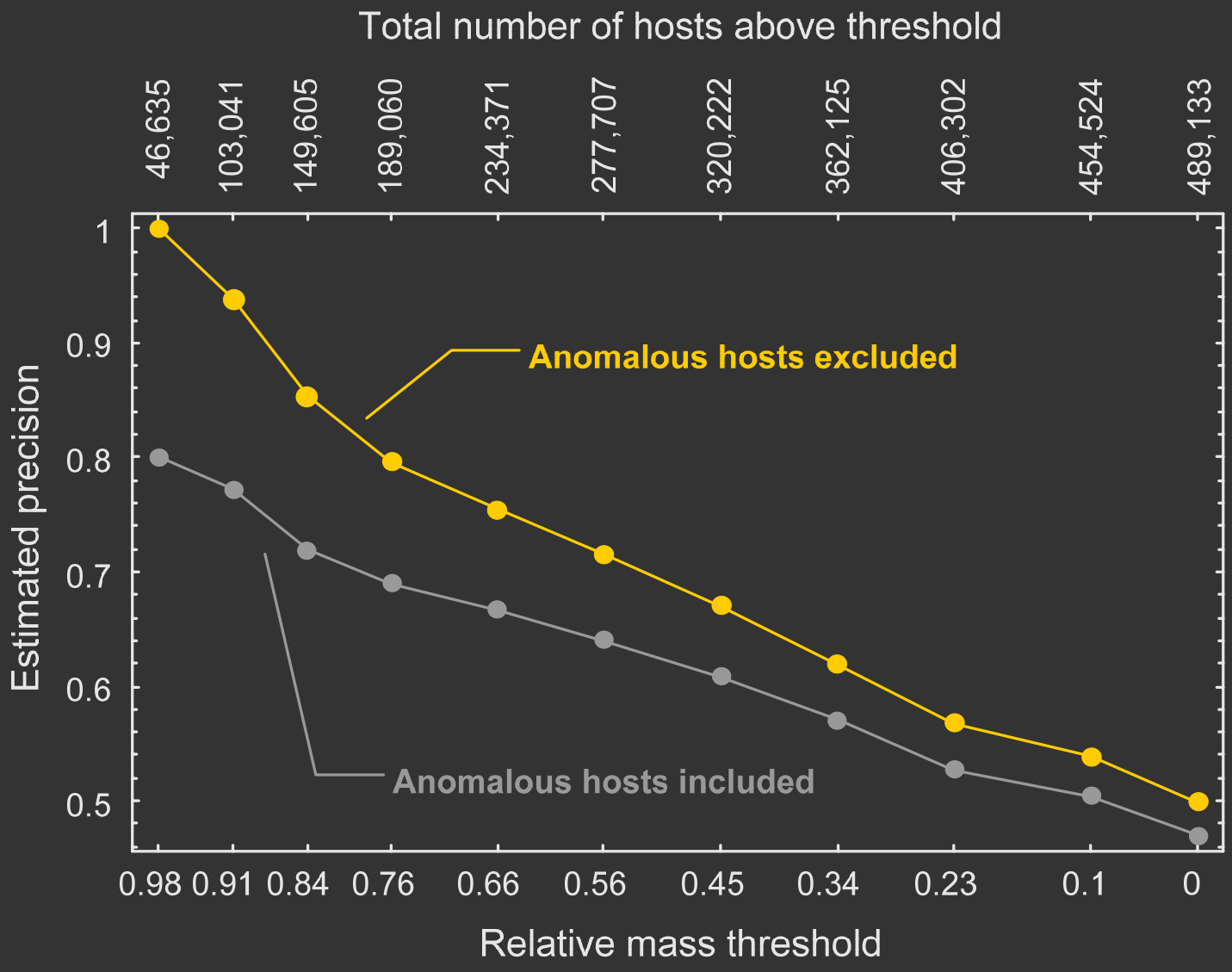
Experiments: Relative Mass



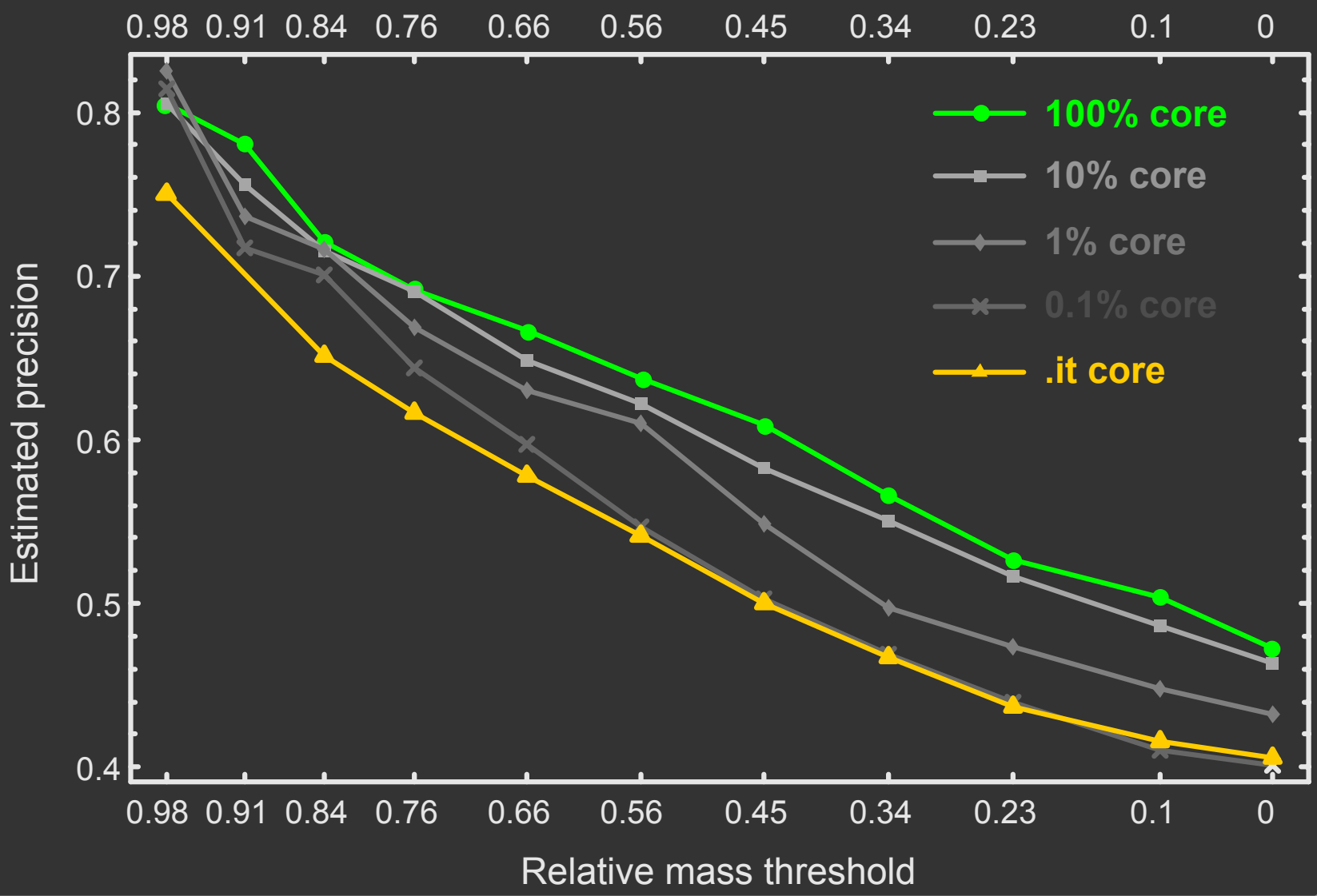
Experiments: Relative Mass

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

Experiments: Relative Mass



Experiments: Core Size



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], [Wu+2006]

Conclusions

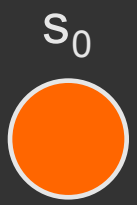
- Search engine spamming
  - Manipulation of search engine ranking
  - Focus on link spamming
- Spam mass
  - ~ PageRank contribution of spam
  - Useful in link spam detection
- Strong experimental results
  - Virtually 100% of top 47K nodes spam
  - 94% of top 105K nodes spam

Link Spamming: Model

- Spam farm

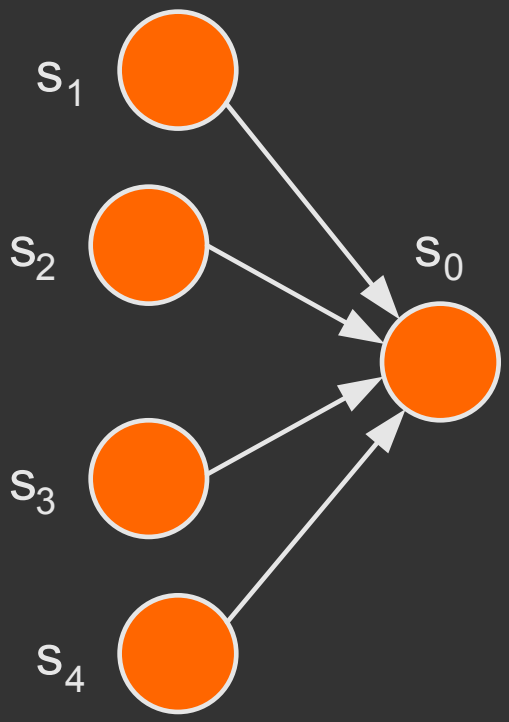
Link Spamming: Model

- Spam farm
  1. Target node



Link Spamming: Model

- Spam farm
  1. Target node
  2. Boosting nodes

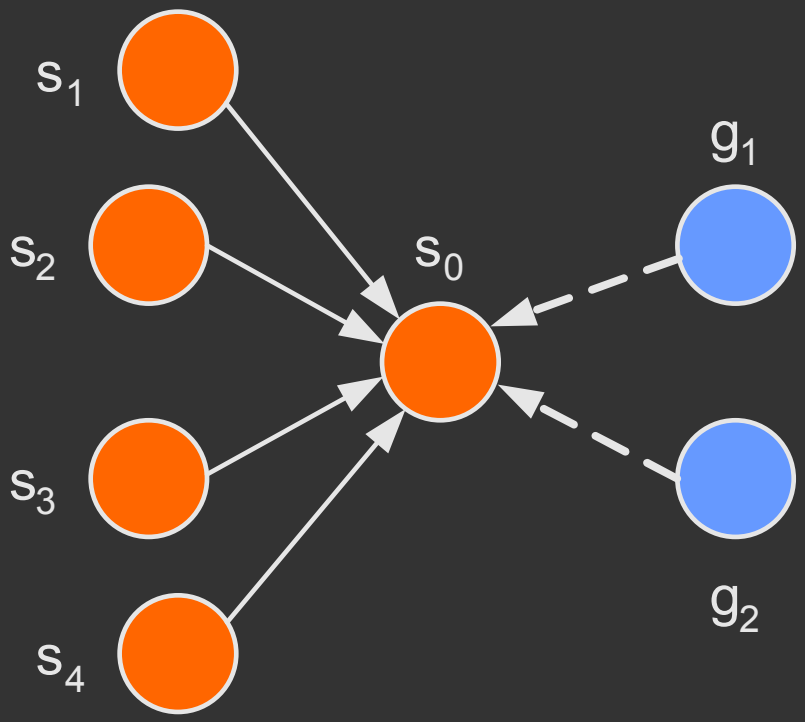


Ski Austria travel…

Great cheap ski Switzerland Italy travel best rates winter sports hotels

Link Spamming: Model

- Spam farm
  1. Target node
  2. Boosting nodes
  3. Hijacked links from good nodes



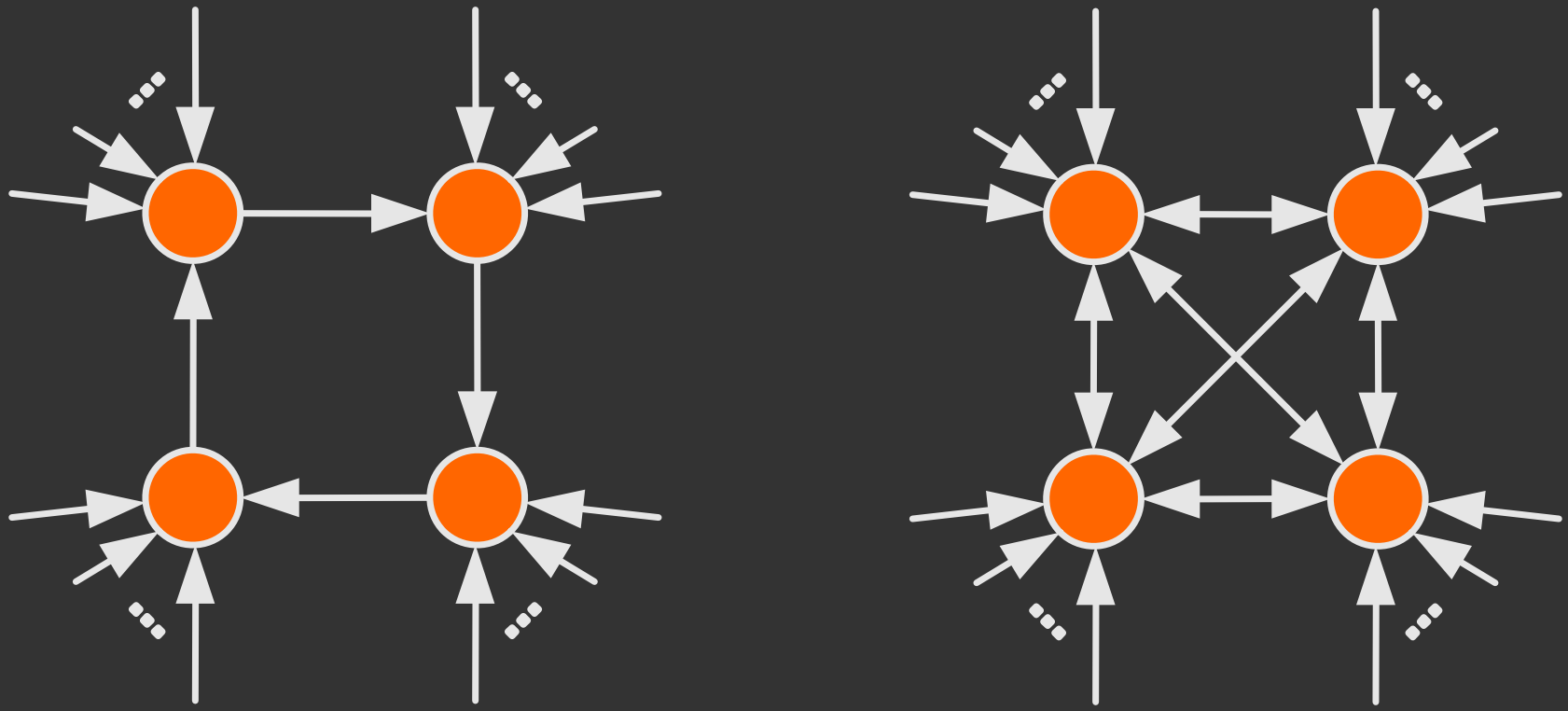
Joe's Blog

Comments

Great pictures! See my Austria ski vacation. (by as7869)

Link Spamming: Model

- Spam farm alliances



PageRank

- Probabilistic model: $p = c\, U^T p + (1 - c)\, v$
  - $U = U(T, v)$ stochastic transition matrix
  - $|v| = 1$
- Linear model: $(I - c\, T^T)\, p = (1 - c)\, v$
  - No adjustment for nodes without outlinks (transition matrix T has all-zero rows)
  - Advantages
    - For $p = PR(v)$ and $v = v_1 + v_2$, $p = p_1 + p_2$ where $p_1 = PR(v_1)$ and $p_2 = PR(v_2)$
    - Faster to compute

PageRank Contribution

- Walk W from x to y: $x = x_0, x_1, \ldots, x_k = y$
  - Weight $\pi(W) = \text{out}(x_0)^{-1} \cdot \cdots \cdot \text{out}(x_{k-1})^{-1}$
- Contribution of x to y over W:
  $c^k \pi(W) (1 - c) / n$
- PageRank contribution p_y^x of x to y—over all walks
  - Possibly infinite # of walks if there are cycles
  - p_y^x = PR(random jump to x only)
- See also [Jeh+2003]