

YAHOO!



# 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”

This list is designed to make it easy to find & select Activelifestyle domains

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.commm
austrianarlberg.com	jungfrauregion.com	courmayeur.com		holidays-lastminute.com
lechaustria.com	verbierswitzerland.com	dolomitesitaly.com		ski-lastminute.com
stubaiaustria.com	zermattswitzerland.com	livignoitaly.com		
tirolaustria.com	holidayswitzerland.com	holidaysitaly.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
thailandgolffmaps.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	alpineseconds.com	euroski-on-line.com
bestpriceitaly.com	bestpricecourmayeur.com	bestpriceairlinetickets.com	alpinsummer.com	businesstraveltoday.com
bestpriceswitzerland.com	bestpriceskiing.com	bestpriceairtickets.com	lakemountainseurope.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	lakemountainapartments.com
availableroomszermatt.com	availableaccommodationitaly.com	apartmentaustria.com	livignoapartments.com

# Spamming: Example

#1 search result for the query “austria ski”

This list is designed to make it easy to find & select Activelifestyle domains

Austria ski/resorts	Swiss/ski/resorts	Italy/ski/resorts	France/ski/holidays	Last Minute
skiaustria.com stantonaustria.com austrianarlberg.com lechaustria.com stubaiaustria.com tirolaustria.com holidaysaustria.com	skiswitzerland.com zermatt.com jungfrauregion.com verbierswitzerland.com zermattswitzerland.com holidayswitzerland.com	skiitaly.com aostaitaly.com courmayeur.com dolomites.com livignoitaly.com holidaysitaly.com	skifrance.com	dive-lastminute.com mmmm.com m
Asia/activities/dest	Asia/activities/dest	Holidays		
travelthailand.com bangkokthailand.com pattayathailand.com phuketthailand.com thailandgolfmaps.com	asiandiveholidays.com asianmp3.com mp3thailand.com thailandhealthcaretimes.com thailandpropertytimes.com	holidayswitzerland.com holidaysitaly.com european.com croatian.com slovenian.com		ica.com da.com rts.com kok.com
Best Price	Best Price	Best Price	Alpine Sun	Special travel
bestpriceeurope.com bestpriceaustria.com bestpriceitaly.com bestpriceswitzerland.com bestpricefrance.com	bestpricethailand.com bestpricezermatt.com bestpricecourmayeur.com bestpriceskiing.com bestpricegolfing.com	bestpricetouring.com bestpriceverbier.com bestpriceairlinetickets.com bestpriceairtickets.com bestpricetravelnetwork.com	alpineholidays.com alpineseconds.com alpinsummer.com lakemountainseurope.com hotelsinthealps.com alpinegolf.com alpineskimaps.com	activelifestylewoman.com euroski-on-line.com businesstraveltoday.com bookhotelsdirect.com activelifestyle.com activelifestylemall.com gullibletraveler.com
Available Accommodation	Available Accommodation	global apartments	global apartments	
availableroomsthailand.com availableroomszermatt.com	availableroomsswitzerland.com availableaccommodationitaly.com	alpineapartmentregister.com apartmentaustria.com	lakemountainapartments.com lignoapartments.com	

# Spamming: Example

WWW.LuxuryHotelsAmerica.com - Microsoft Internet Explorer

File Edit View Favorites Tools Help

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

[www.fairmont.com/theplaza/](http://www.fairmont.com/theplaza/)  
[www.melrosehotel.com](http://www.melrosehotel.com/)  
[www.fourseasons.com/](http://www.fourseasons.com/)  
[www.fourseasons.com/pierre/](http://www.fourseasons.com/pierre/)  
[www.michelangelohotel.com](http://www.michelangelohotel.com)  
[www.themarkhotel.com](http://www.themarkhotel.com)  
[www.luxury-hotels-resorts.com/new-york](http://www.luxury-hotels-resorts.com/new-york)  
[www.thealgonquin.net](http://www.thealgonquin.net)

[www.lemeridienboston.com](http://www.lemeridienboston.com)  
[www.concorde-hotels.com](http://www.concorde-hotels.com)

[www.grosvenorsfo.com/](http://www.grosvenorsfo.com/)  
[www.fourseasons.com/sanfrancisco](http://www.fourseasons.com/sanfrancisco)  
[www.intercontinental.com](http://www.intercontinental.com)  
[www.huntingtonhotel.com](http://www.huntingtonhotel.com)  
[www.mandarin-oriental.com/sanfrancisco/](http://www.mandarin-oriental.com/sanfrancisco/)  
[www.prescotthotel.com/prhunio/](http://www.prescotthotel.com/prhunio/)

**Add your hotel to our database**  
**FREE**

Activelifestyle Travel Network Domains

**Austria**  
[austrianaarlberg.com](http://austrianaarlberg.com)  
[lechaustria.com](http://lechaustria.com)  
[skiaustria.com](http://skiaustria.com)  
[stubaiaustria.com](http://stubaiaustria.com)  
[stantonaustria.com](http://stantonaustria.com)  
[tirolaustria.com](http://tirolaustria.com)  
[holidaysaustria.com](http://holidaysaustria.com)

**France**  
[skifrance.com](http://skifrance.com)  
[holidayfrancais.com](http://holidayfrancais.com)

**Italy**  
[aostaitaly.com](http://aostaitaly.com)  
[courmayeur.com](http://courmayeur.com)  
[dolomitesitaly.com](http://dolomitesitaly.com)  
[livignoitaly.com](http://livignoitaly.com)  
[skitaly.com](http://skitaly.com)  
[holidaysitaly.com](http://holidaysitaly.com)

**Switzerland**  
[jungfrauregion.com](http://jungfrauregion.com)  
[skiswitzerland.com](http://skiswitzerland.com)  
[verbierswitzerland.com](http://verbierswitzerland.com)  
[zermatt.com](http://zermatt.com)  
[zermattswitzerland.com](http://zermattswitzerland.com)  
[holidayswitzerland.com](http://holidayswitzerland.com)

**Maps**  
[alpineskimaps.com](http://alpineskimaps.com)  
[thailandgolfmaps.com](http://thailandgolfmaps.com)

**Asia**  
[bangkokthailand.com](http://bangkokthailand.com)  
[pattayathailand.com](http://pattayathailand.com)  
[phuketthailand.com](http://phuketthailand.com)  
[travelthailand.com](http://travelthailand.com)  
[asiandiveholidays.com](http://asiandiveholidays.com)  
[asianmp3.com](http://asianmp3.com)  
[mp3thailand.com](http://mp3thailand.com)

**Luxury**  
[luxuryalpinehotels.com](http://luxuryalpinehotels.com)  
[luxuryasianhotels.com](http://luxuryasianhotels.com)  
[luxurygolfdestinations.com](http://luxurygolfdestinations.com)  
[luxuryyachtholidays.com](http://luxuryyachtholidays.com)  
[luxuryhotelsamerica.com](http://luxuryhotelsamerica.com)  
[luxuryhotelscanada.com](http://luxuryhotelscanada.com)  
[luxuryislandresorts.com](http://luxuryislandresorts.com)  
[luxuryrski.com](http://luxuryrski.com)

**Last Minute**  
[dive-lastminute.com](http://dive-lastminute.com)  
[golf-lastminute.com](http://golf-lastminute.com)  
[holidays-lastminute.com](http://holidays-lastminute.com)  
[ski-lastminute.com](http://ski-lastminute.com)

**Audio**  
[skihear.com](http://skihear.com)  
[teehear.com](http://teehear.com)

**Winter sports / La Plagne**  
50 chalets & apartments / 6-20p.  
Last minute up to 40% discount  
[www.holidaychalet.com](http://www.holidaychalet.com)

**Alp Leisure Ltd**  
Private Luxury Chalets 3 Vallees Skiing with Freedom to Choose  
[www.alpleisure.com](http://www.alpleisure.com)

**VerbierChalet.com**

go to [www.holidaychalet.com](http://www.holidaychalet.com)

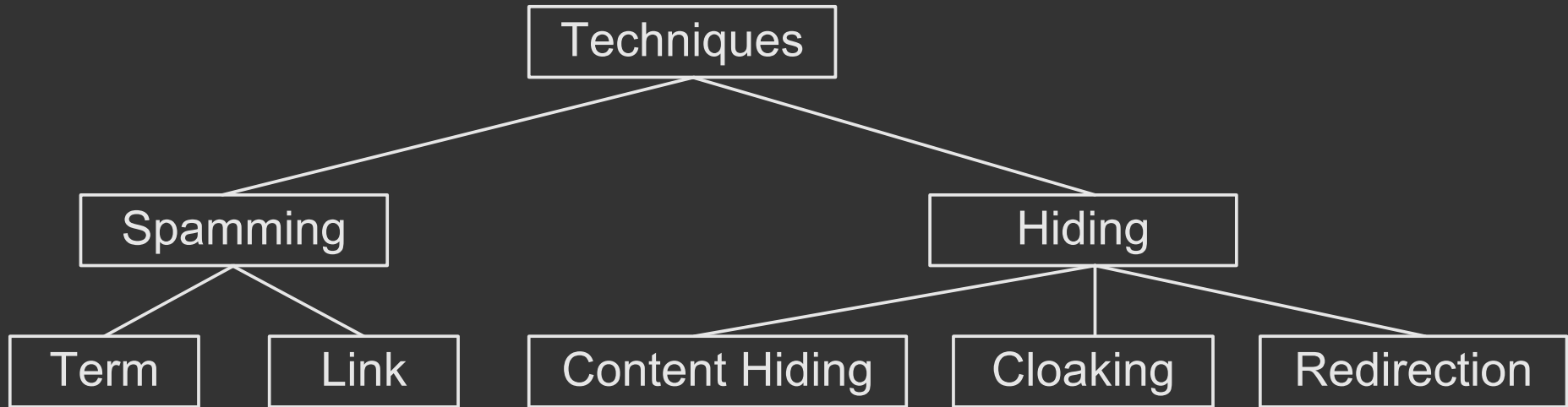
Internet

# Spamming: Introduction

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

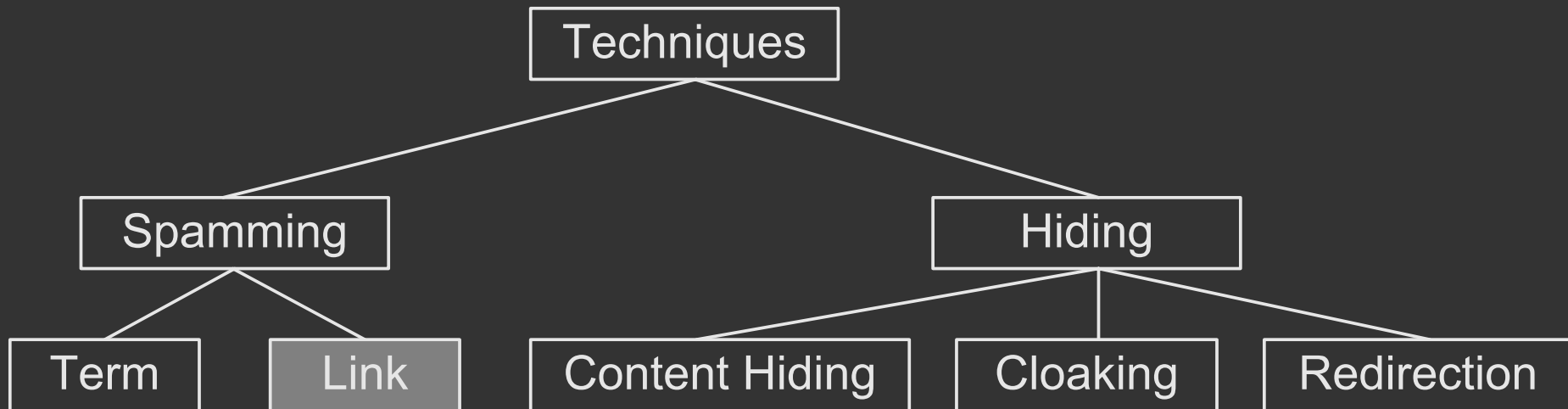
# Spamming: Introduction

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



# Spamming: Introduction

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

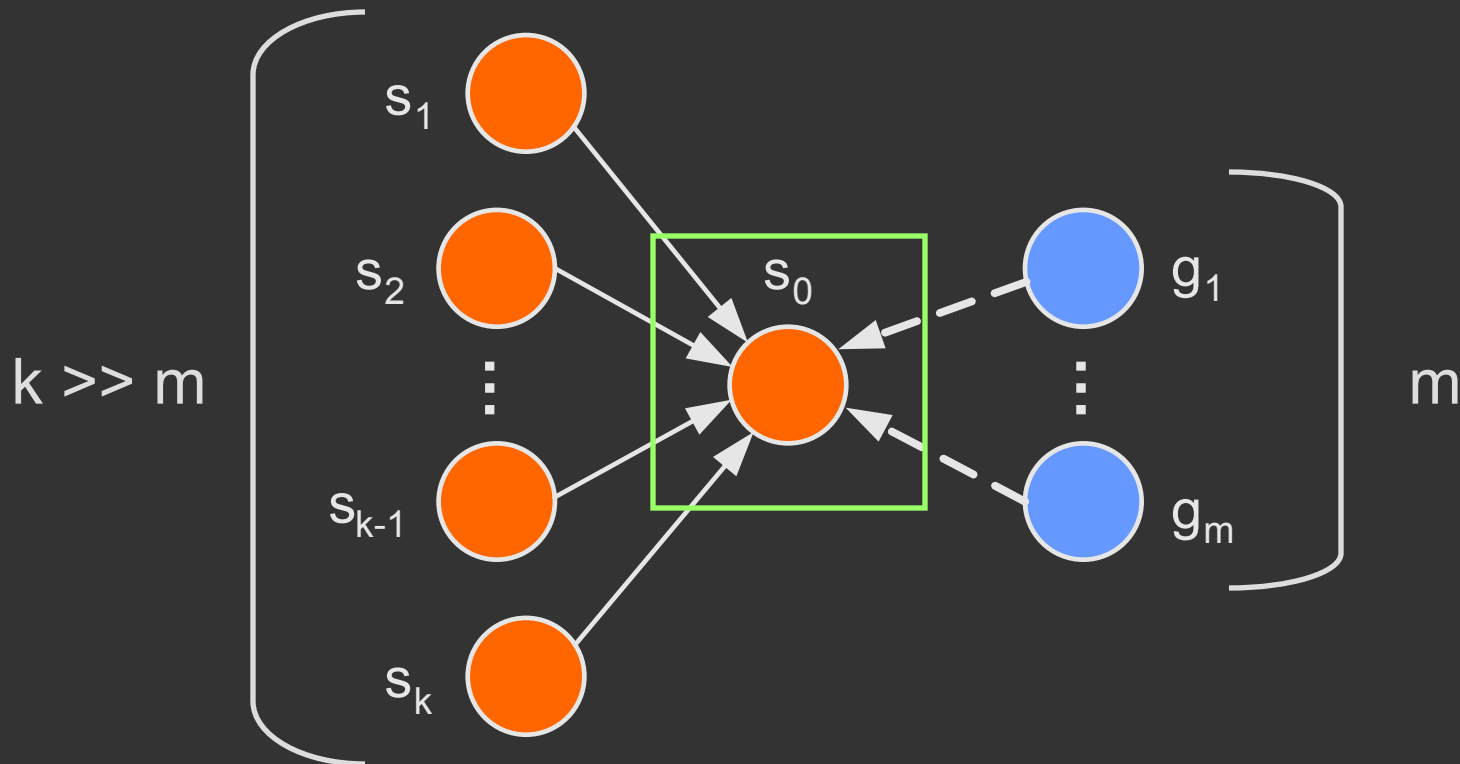


**Link spamming** = building link structures that boost PageRank score

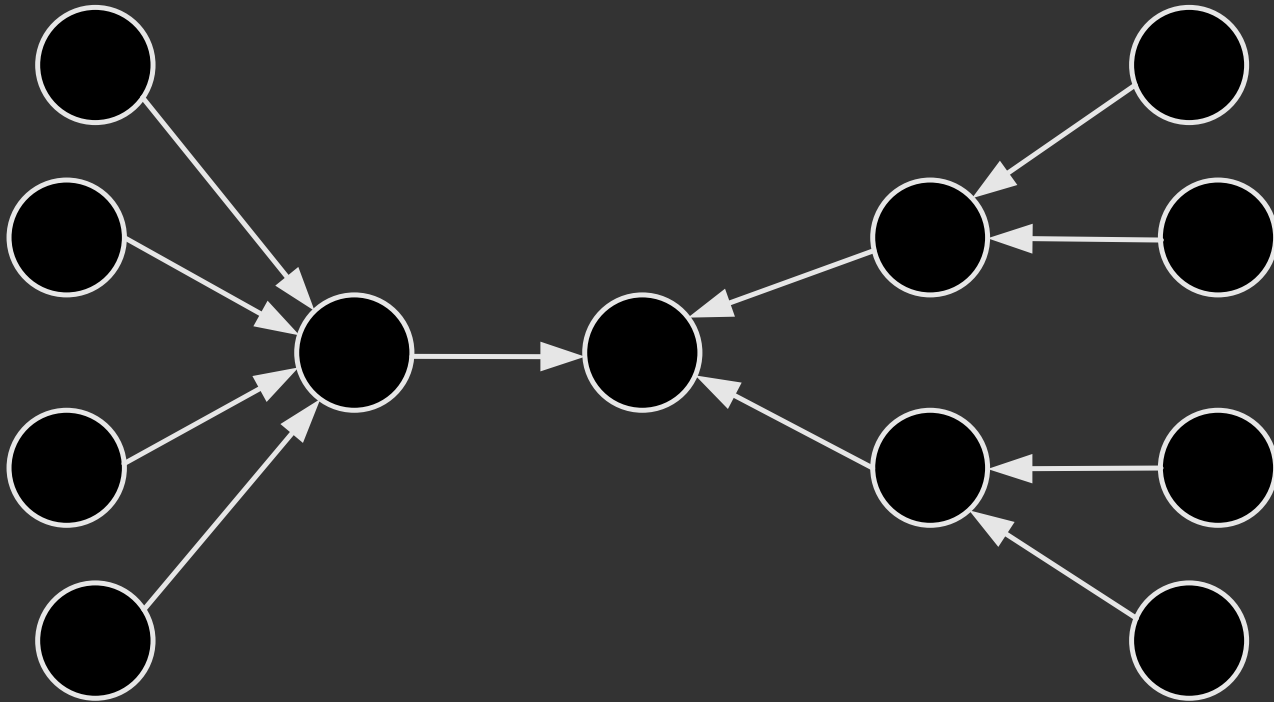


# Spamming: Our Target

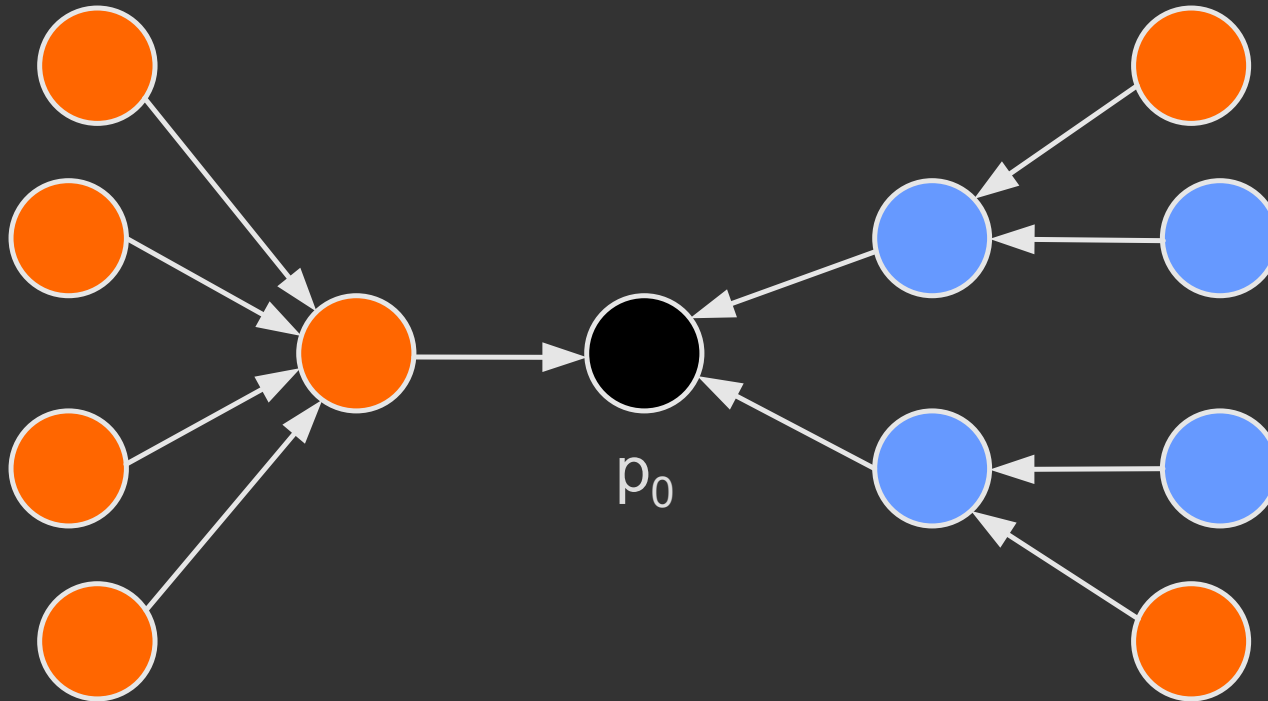
Detect pages that achieve high PageRank through link spamming



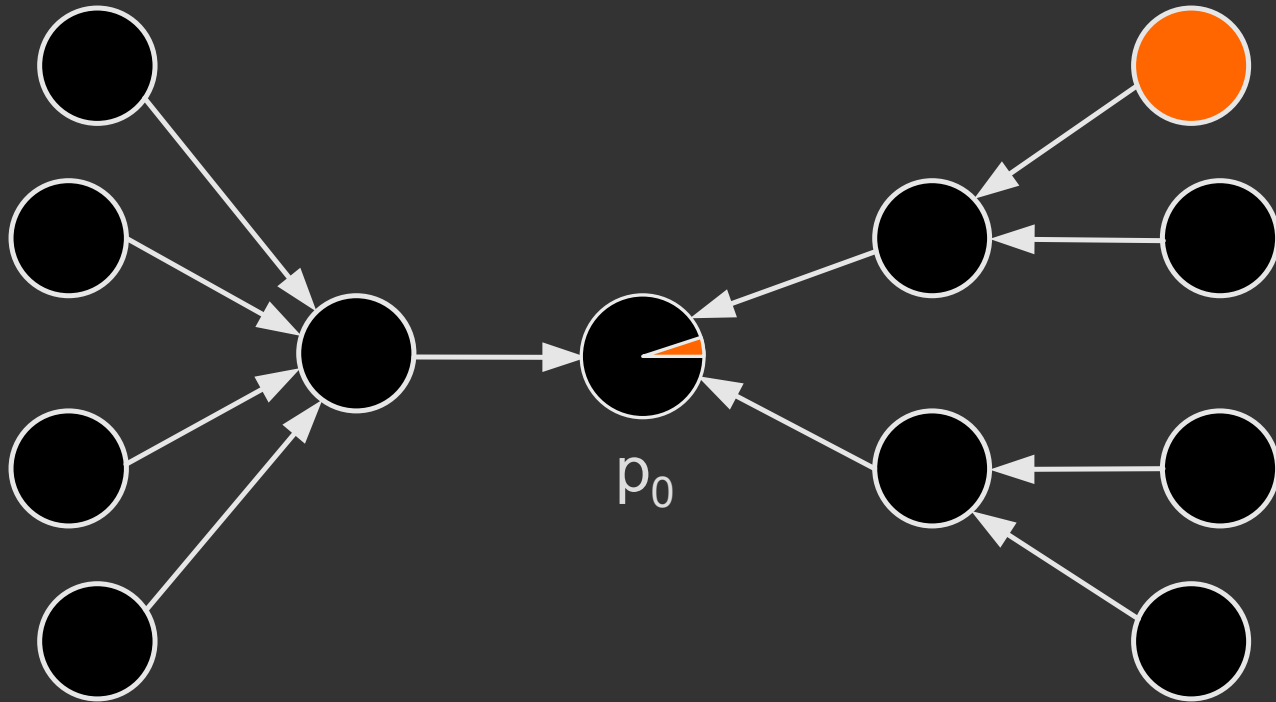
# PageRank Contribution



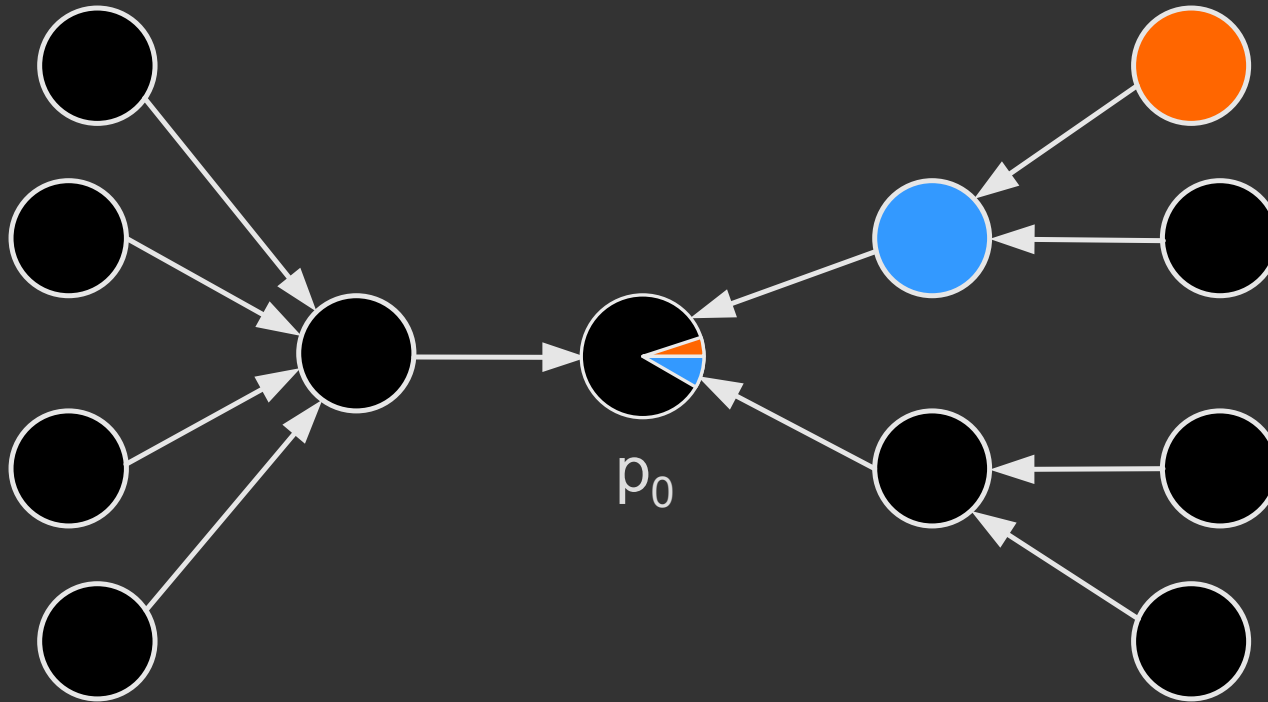
# PageRank Contribution



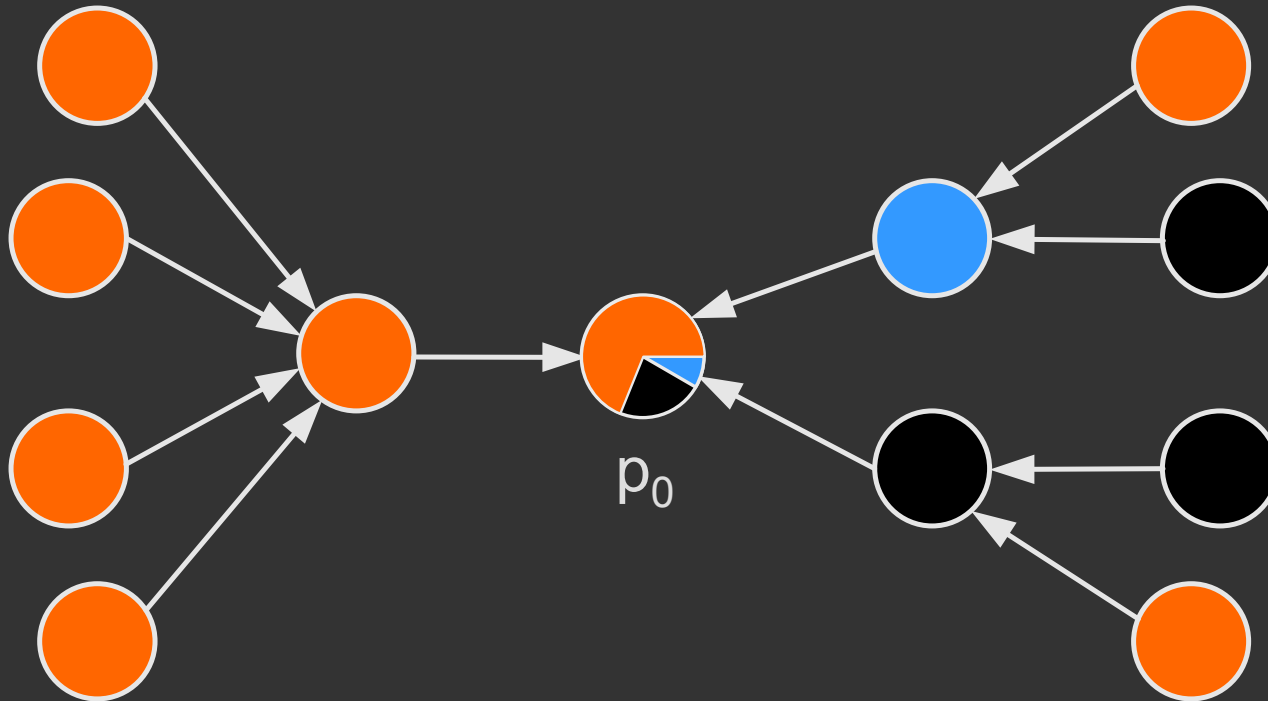
# PageRank Contribution



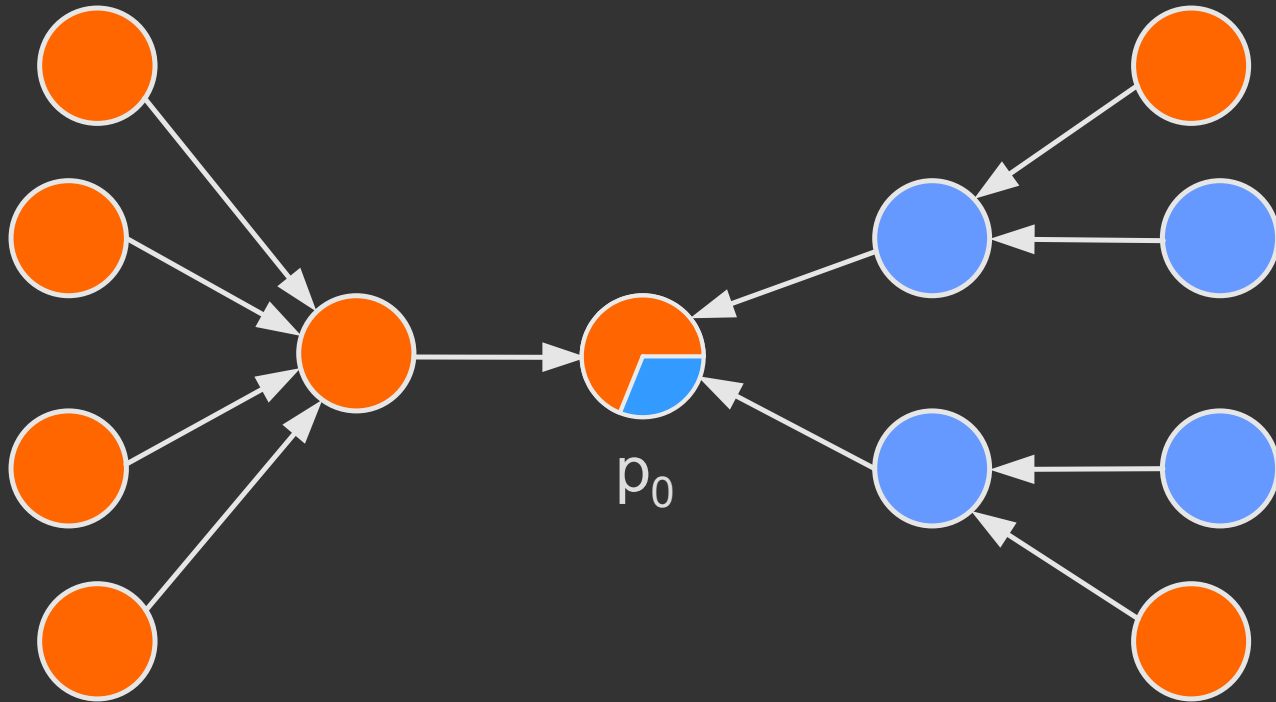
# PageRank Contribution



# PageRank Contribution



# PageRank Contribution



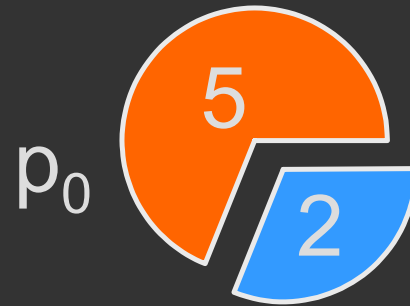
$$p_0^+ = 2c^2(1-c)/n + 2c(1-c)/n$$

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

# Spam Mass: Definition

- Absolute mass
  - **Amount** (part) of PageRank coming from spam
- Relative mass
  - **Fraction** of PageRank coming from spam
  - More useful in practice

$$\text{a.m.} = p_0^- = 5$$

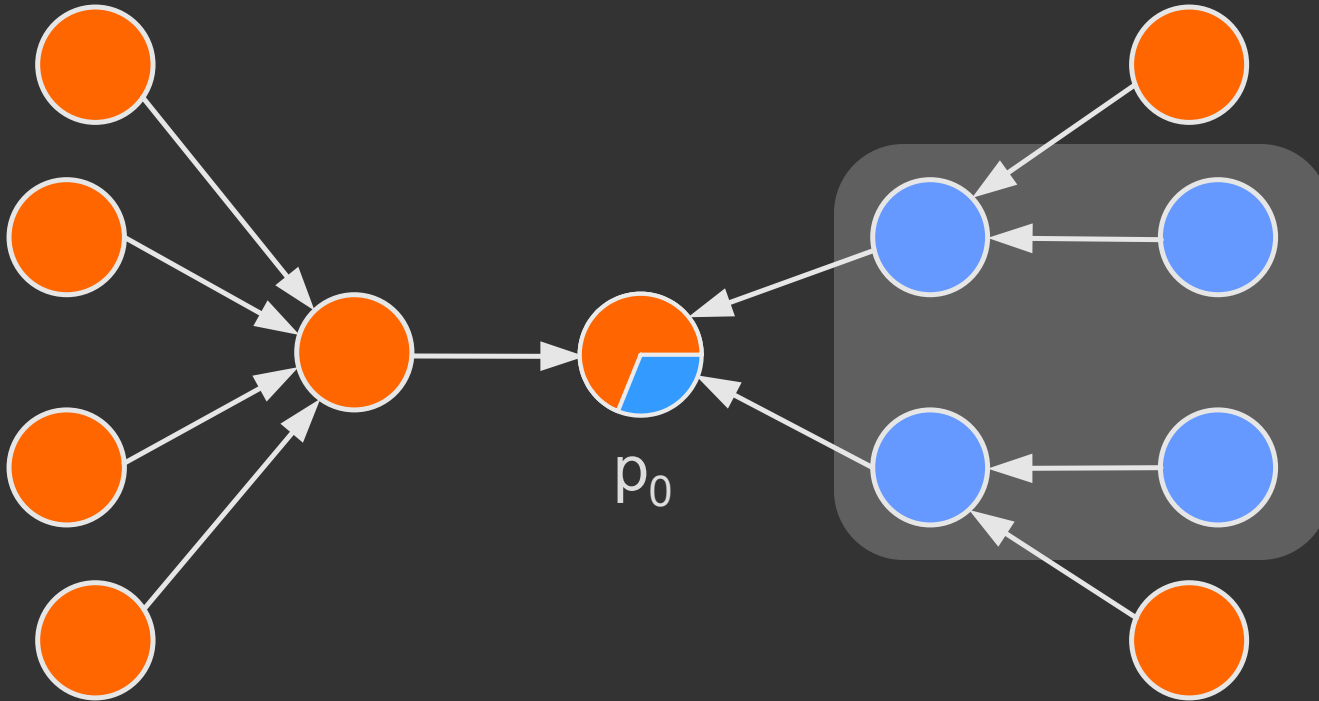


$$\text{r.m.} = \frac{p_0^-}{p_0} = \frac{5}{7}$$



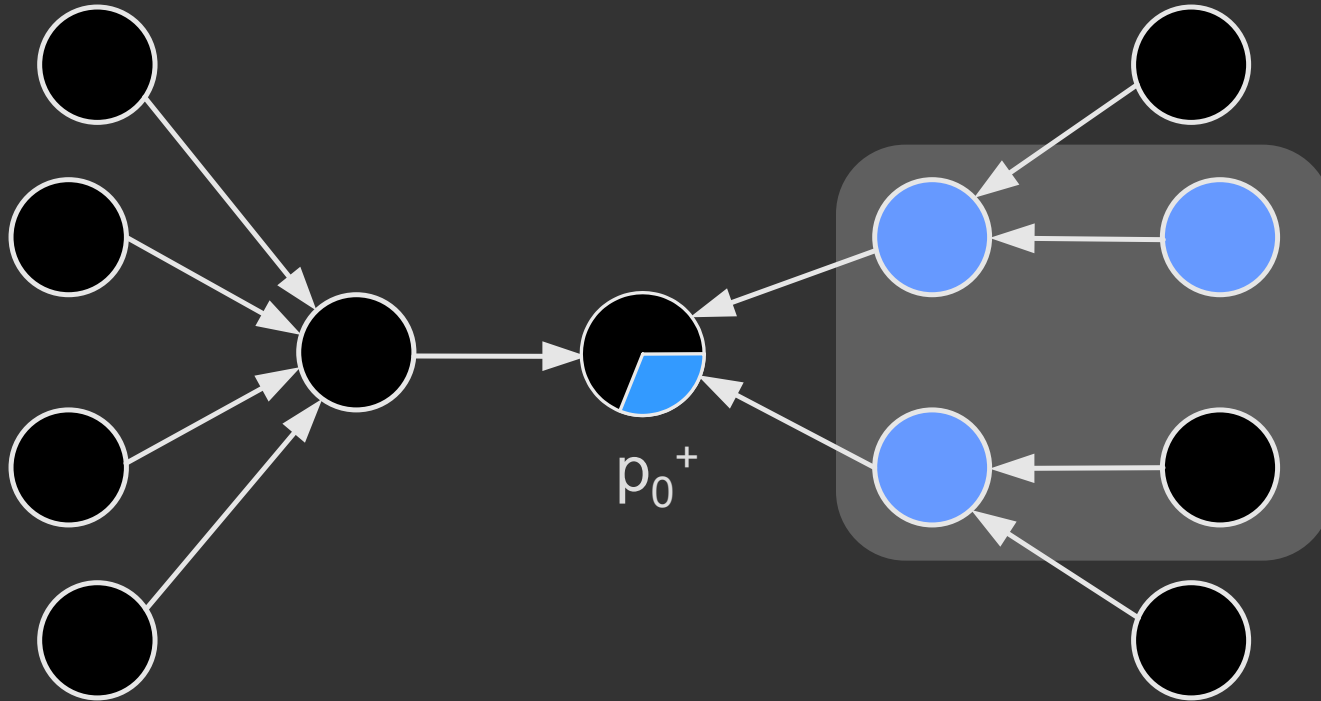
# Spam Mass: Estimation

Ideally...



# Spam Mass: Estimation

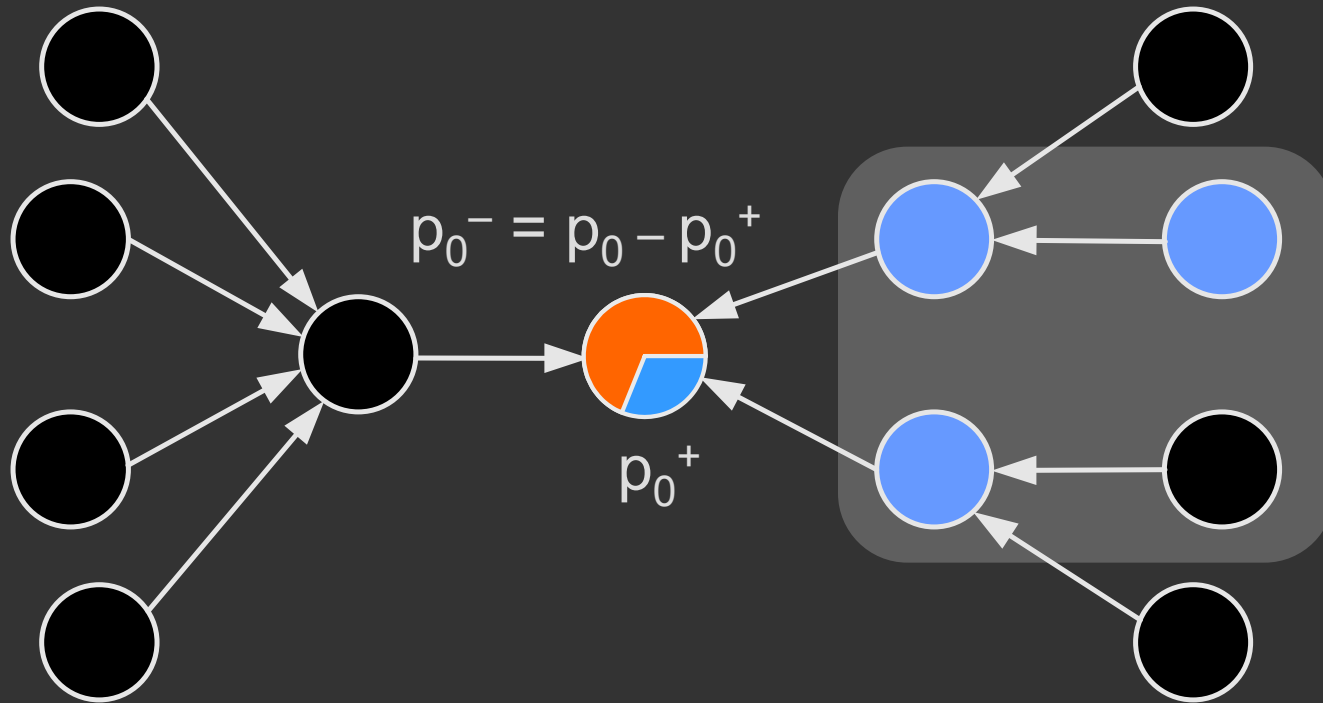
In practice...



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

# 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  $m_i > \text{threshold}$

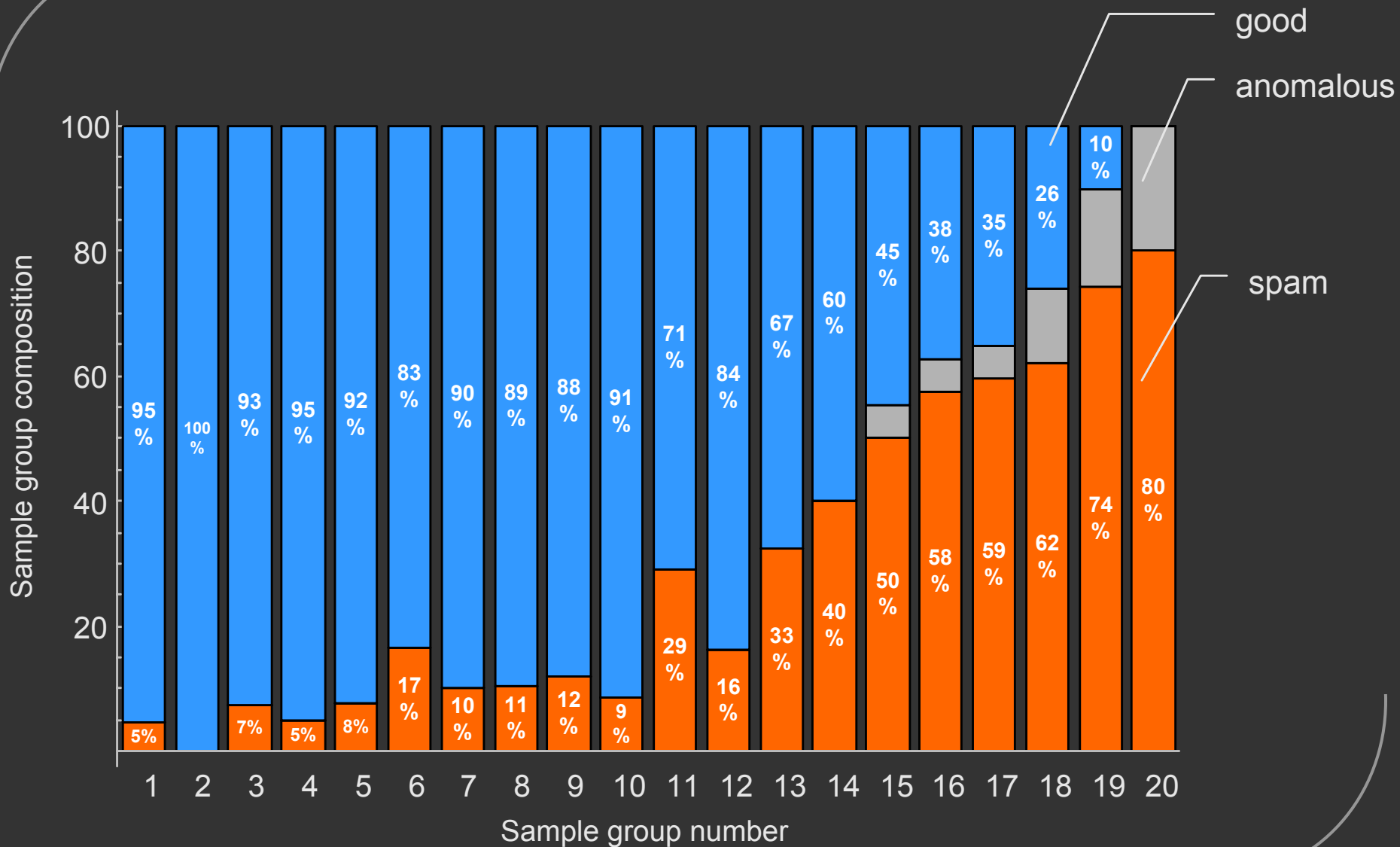
# Experiments: Data

- 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

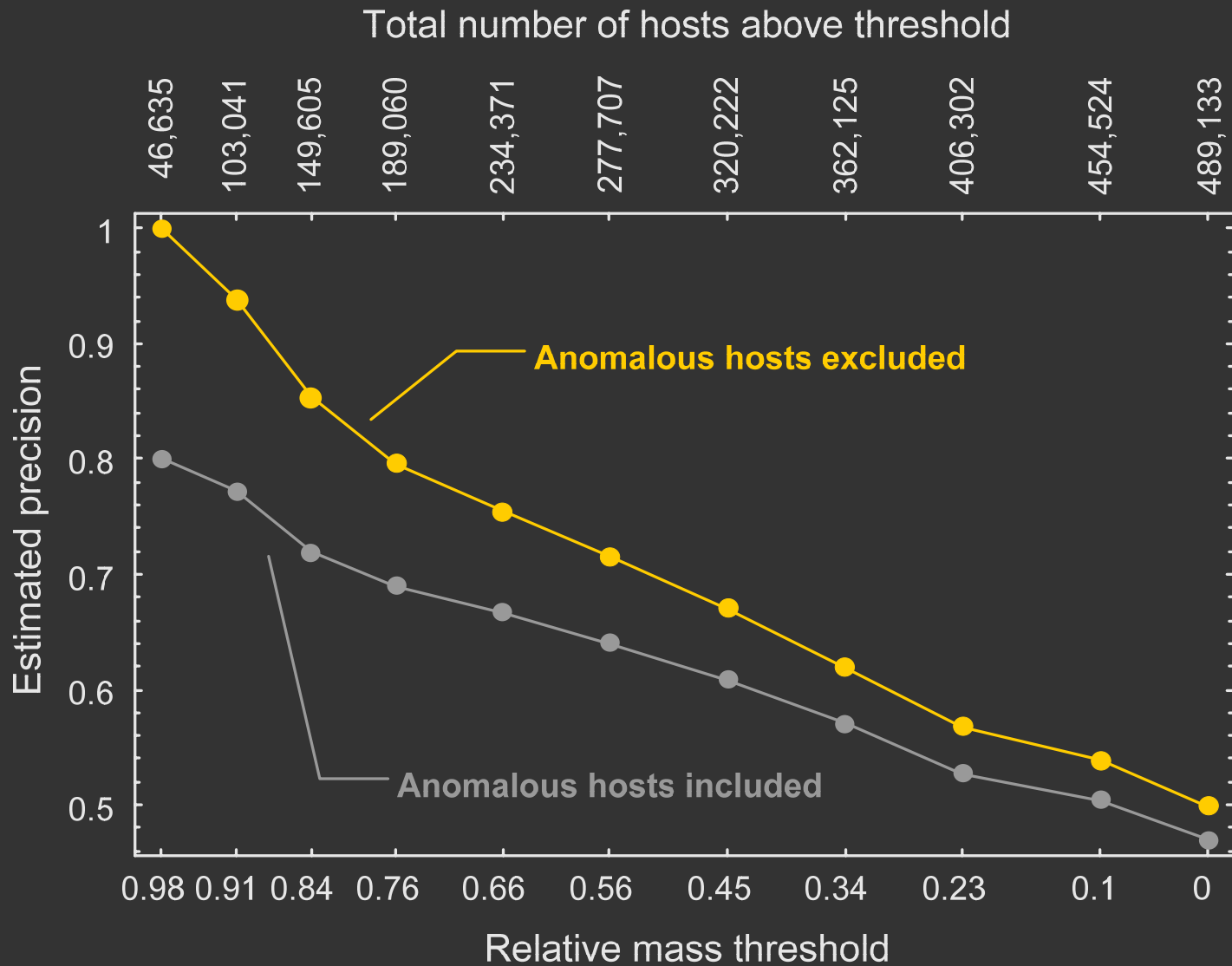


# Experiments: Relative Mass

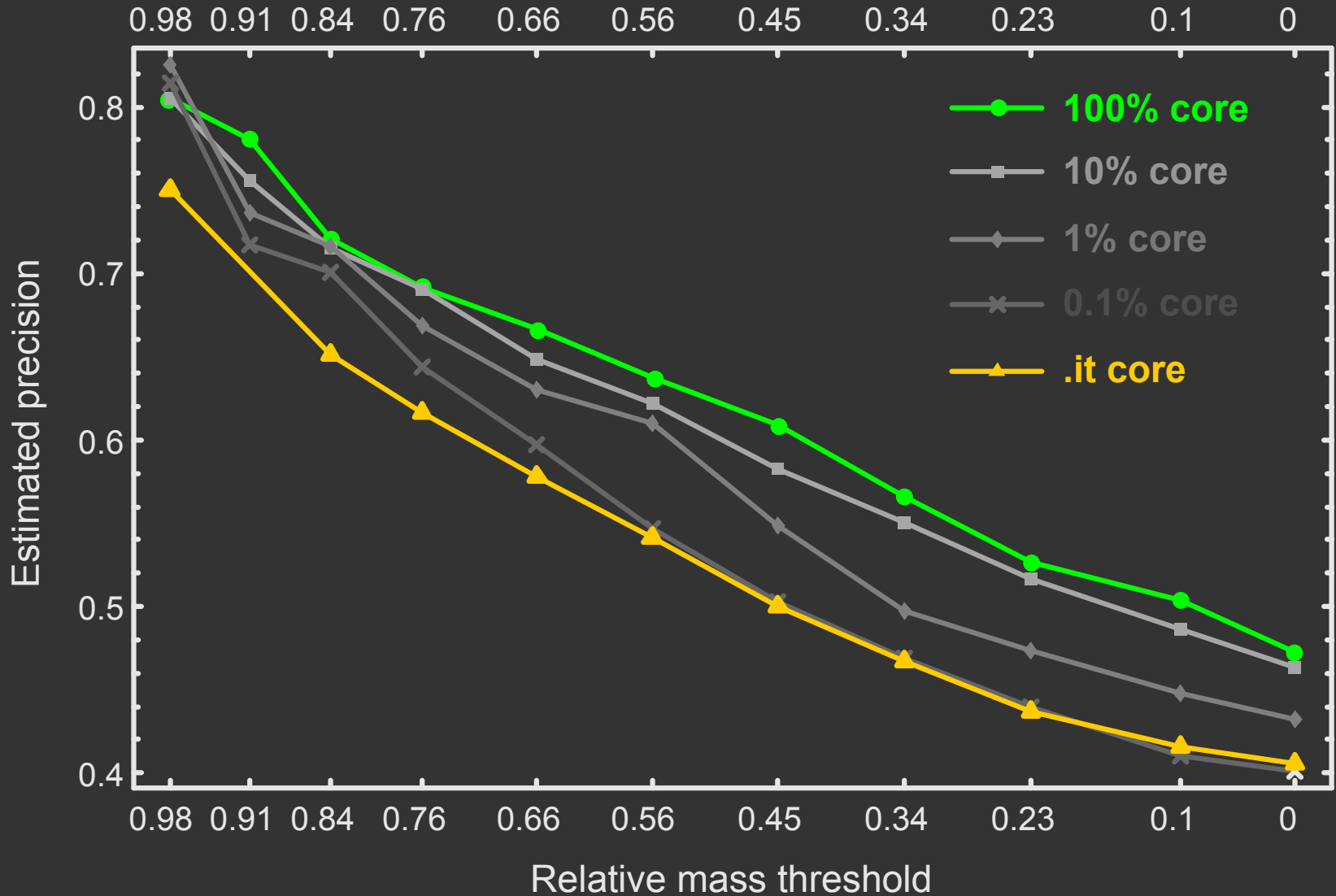
- Anomalies
  - \*.alibaba.com
  - \*.blogger.com.br
  - Polish hosts → only 12 .pl in good core



# Experiments: Relative Mass



# Experiments: 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

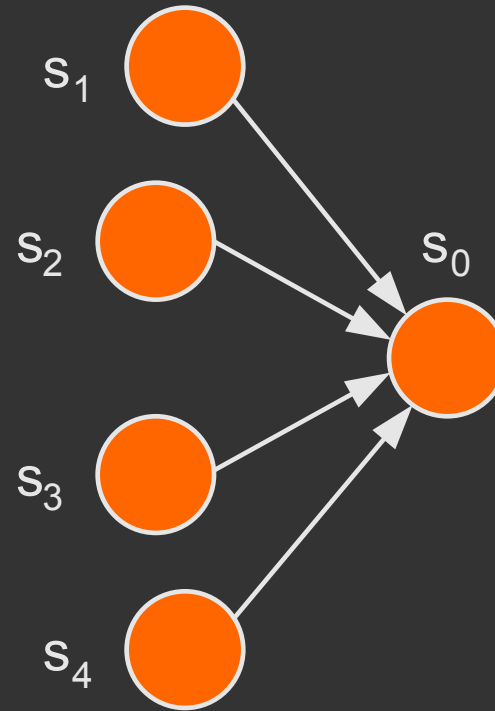


# Link Spamming: Model

- Spam farm
  1. Target node
  2. Boosting nodes

Ski Austria travel...

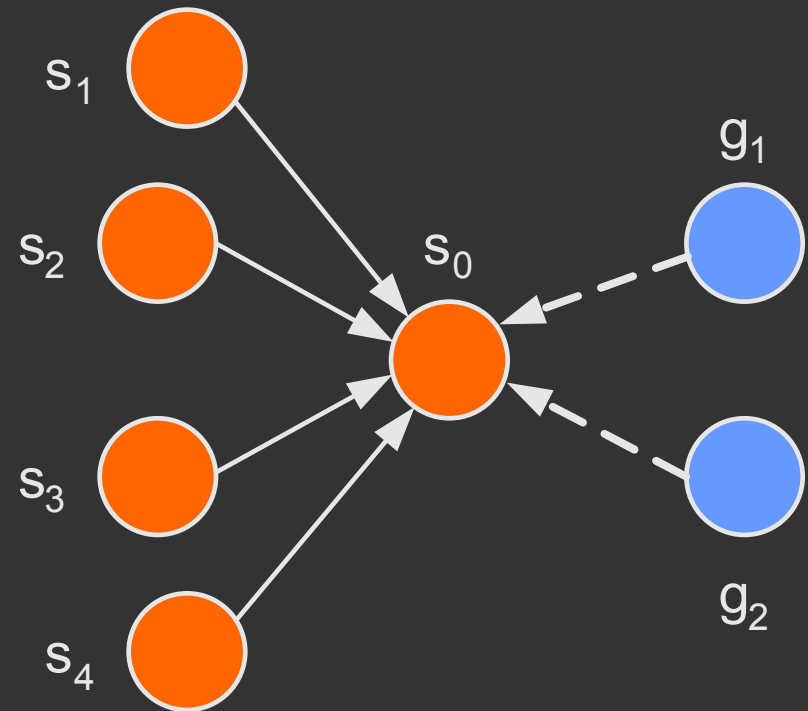
[Great cheap ski](#)  
[Switzerland Italy travel](#)  
[best rates winter sports](#)  
[hotels](#)



# Link Spamming: Model

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

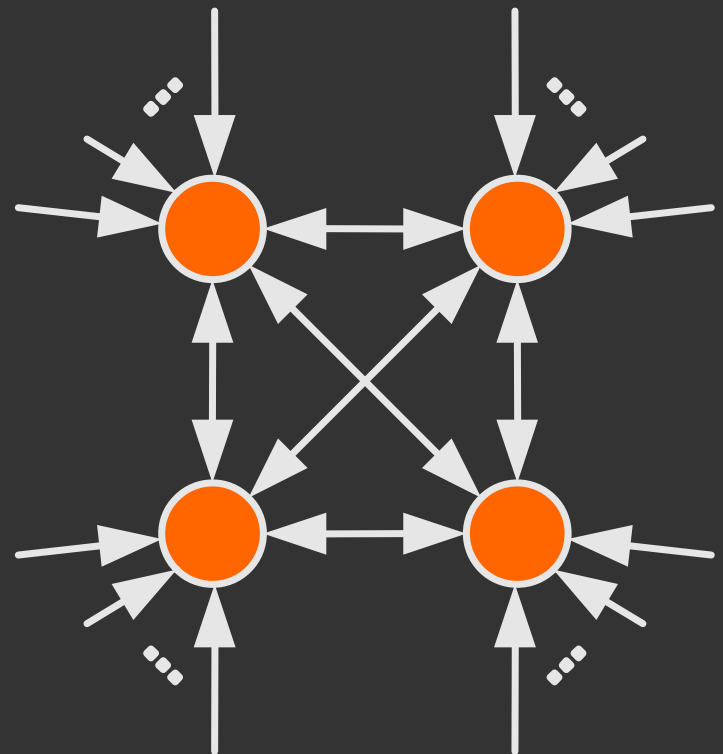
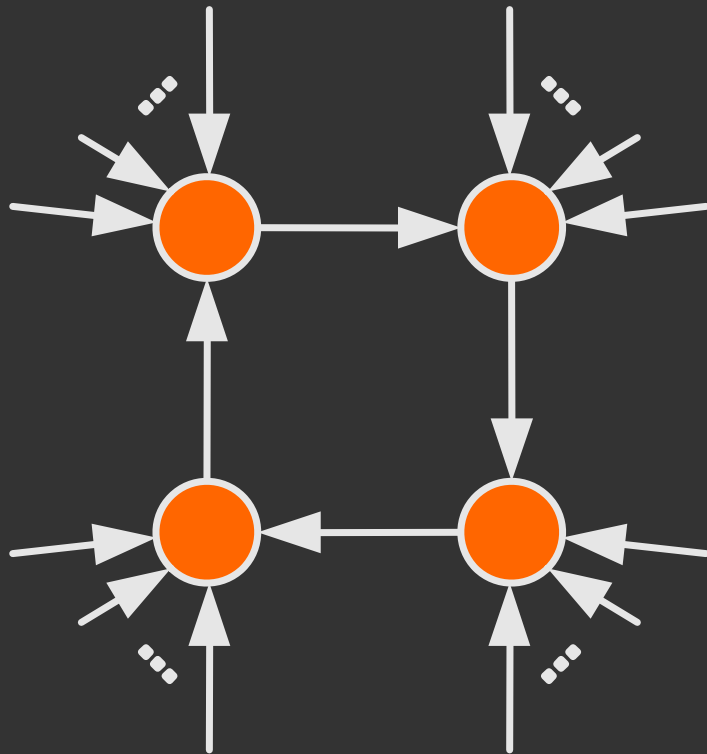
Joe's Blog
<b>Comments</b> Great pictures! See my <a href="#">Austria ski</a> vacation. (by <a href="#">as7869</a> )





# 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  $p = \text{PR}(v)$  and  $v = v_1 + v_2$ ,  $p = p_1 + p_2$  where  $p_1 = \text{PR}(v_1)$  and  $p_2 = \text{PR}(v_2)$
    - Faster to compute

# PageRank Contribution

- Walk  $W$  from  $x$  to  $y$ :  $x = x_0, x_1, \dots, x_k = y$ 
  - Weight  $\pi(W) = \text{out}(x_0)^{-1} \cdot \dots \cdot \text{out}(x_{k-1})^{-1}$
- 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
  - $p_y^x = \text{PR}(\text{random jump to } x \text{ only})$
- See also [Jeh+2003]