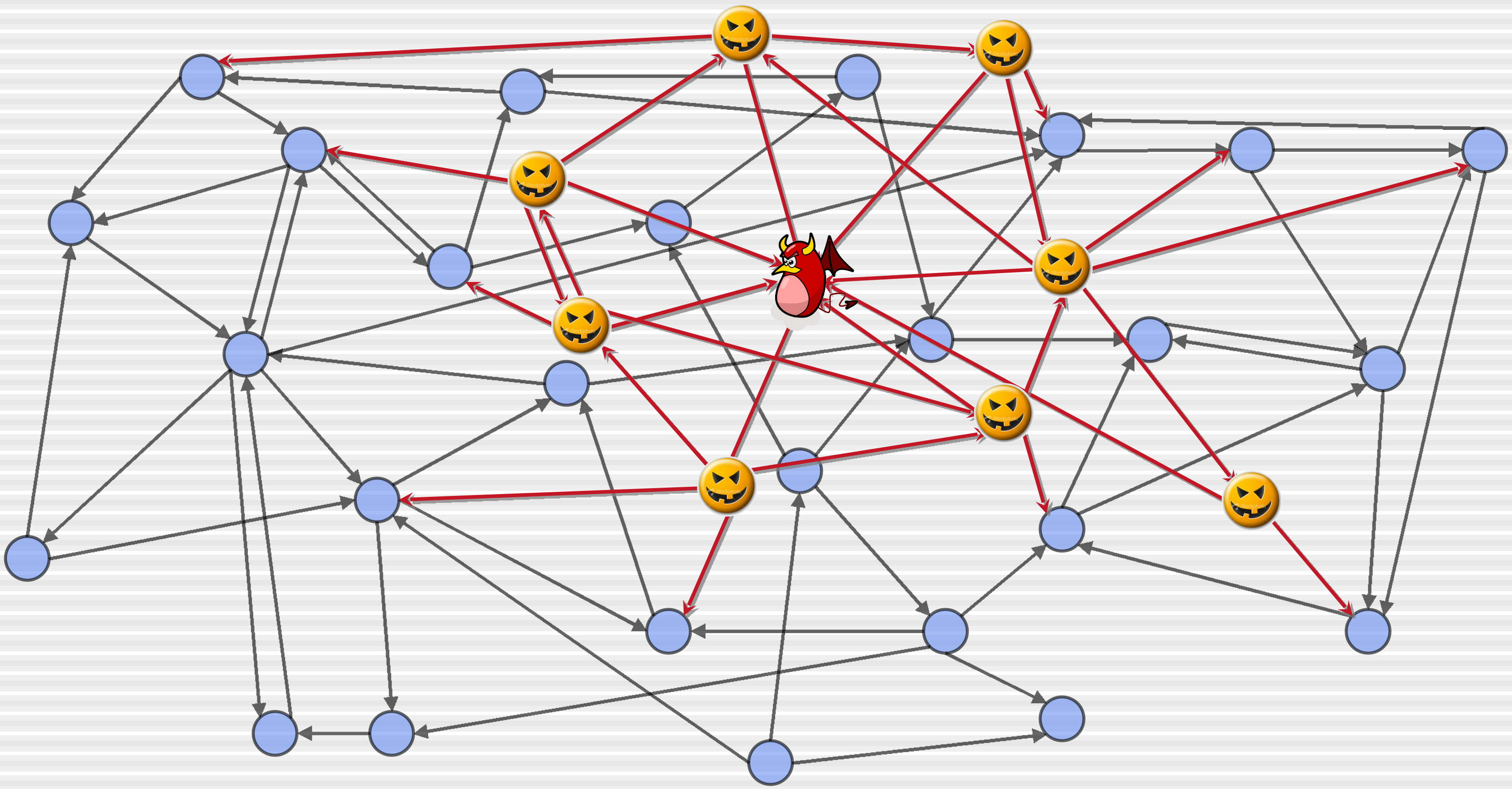
WITCH

A new algorithm for detecting **Web spam** using page features and hyperlinks

Jacob Abernethy, UC Berkeley
(Thanks to Yahoo! Research for two internships and a fellowship!)

Joint work with **Olivier Chapelle** and **Carlos Castillo** (Chato) from Yahoo! Research

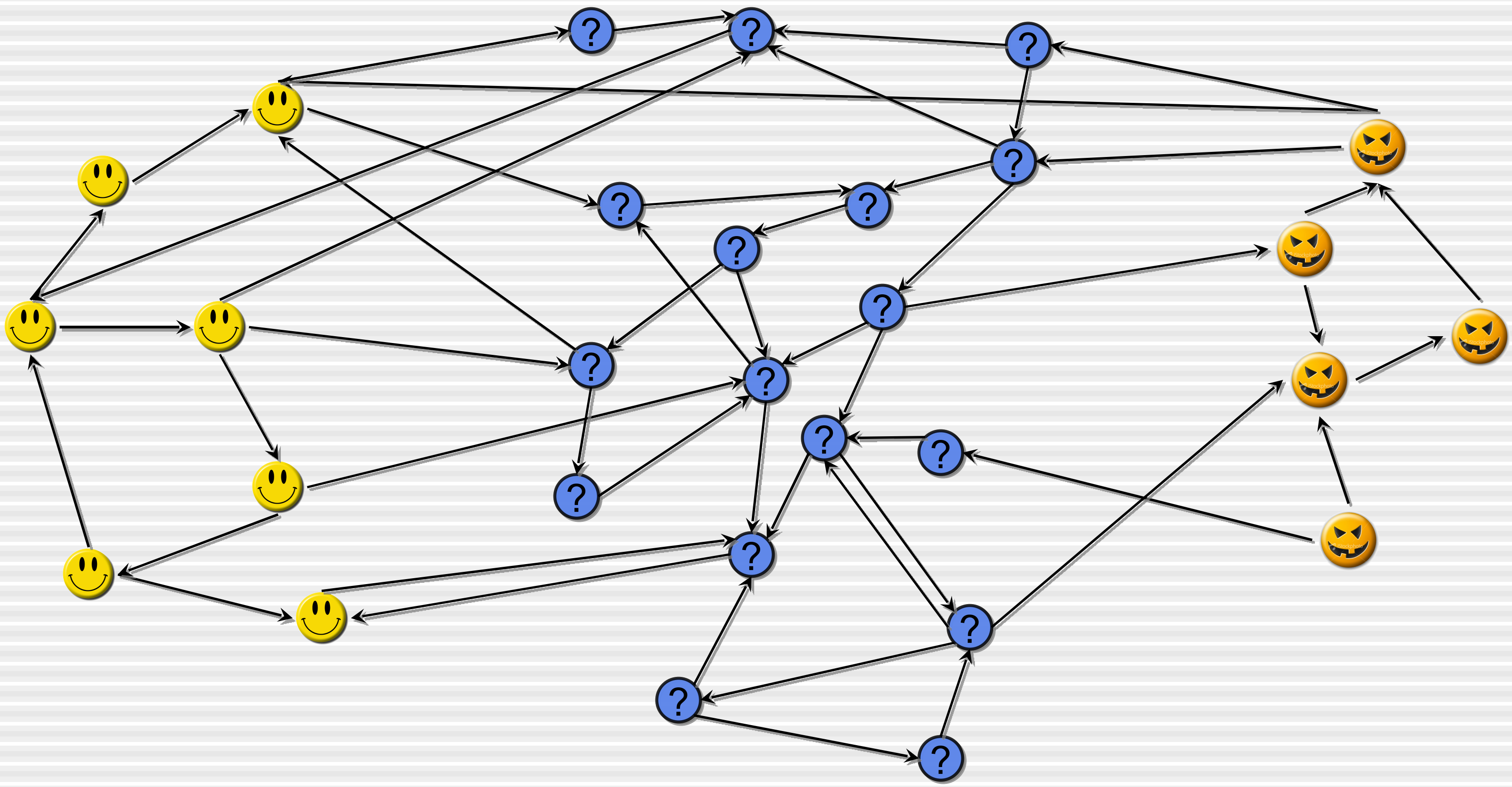
How to Be a Spammer



Learning to Find Spam

- Not a typical learning problem:
  - Web page contents are probably generated adversarially, with the intention of fooling the indexer
  - Given a hyperlink graph, BUT it’s not clear what purpose each link serves: may be natural, may be used for spam, or may simply be there to confuse the indexer

Which of the Blue Hosts are Bad?

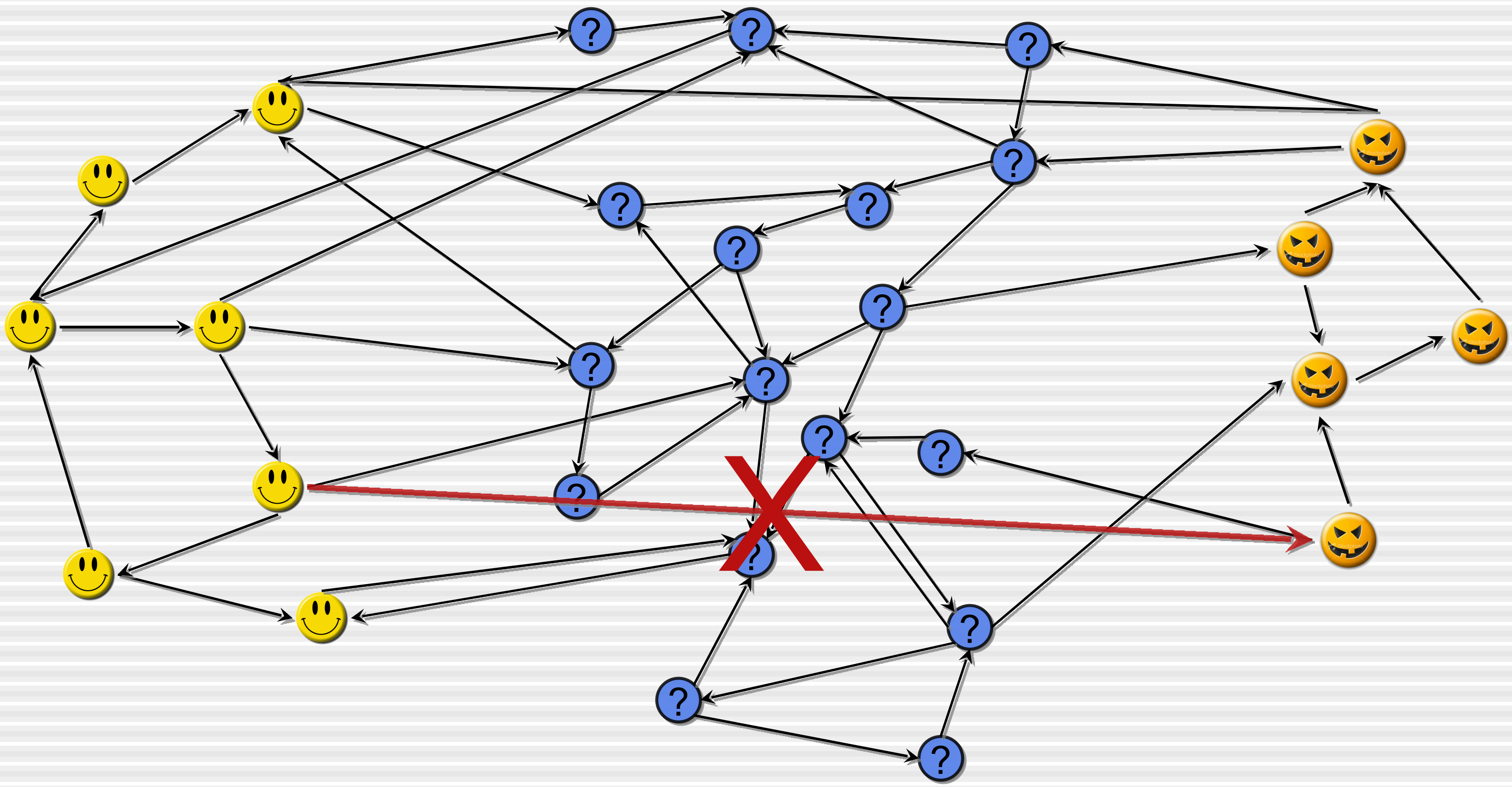


One Key Fact

- An extremely useful observation for spam detection:

Good hosts almost NEVER link to spam hosts!!

Good does NOT link to Bad!



Methods For Web Spam Detection

Graph Based Detection Methods

- Graph-based methods try to compute the “spamicity” of a given page using only the hyperlink graph.
- Perhaps most well-known is **TrustRank,** based on the PageRank algorithm.

Content-Based Methods

- Train a classifier based on page features:
  1. # words in page
  2. Fraction of visible words
  3. Fraction of anchor text
  4. Average word length
  5. Compression rate

WITCH

Web spam Identification Through Content and Hyperlinks

Key Ingredients

- Support Vector Machine (SVM) type framework
- Additional slack variable per node
- "Semi-directed" graph regularization
- Efficient Newton-like optimization

WITCH Framework 1

- Standard **SVM**: fit your data, but make sure your classifier isn't too complicated (aka has a large margin)

$$\Omega(\mathbf{w}) = \frac{1}{l}\sum_{i=1}^{l}[1 - y_i\mathbf{w}\cdot\mathbf{x}_i]_+^2 + \lambda\mathbf{w}\cdot\mathbf{w}$$

WITCH Framework 2

- **Graph Regularized** SVM: fit your data, control complexity, AND make sure your classifier “predicts smoothly along the graph”

$$\Omega(\mathbf{w}) = \frac{1}{l}\sum_{i=1}^{l}[1 - y_i \mathbf{w}\cdot\mathbf{x}_i]_+ + \lambda \mathbf{w}\cdot\mathbf{w} + \gamma \sum_{(i,j)\in E} a_{ij}(\mathbf{w}\cdot\mathbf{x}_i - \mathbf{w}\cdot\mathbf{x}_j)^2$$

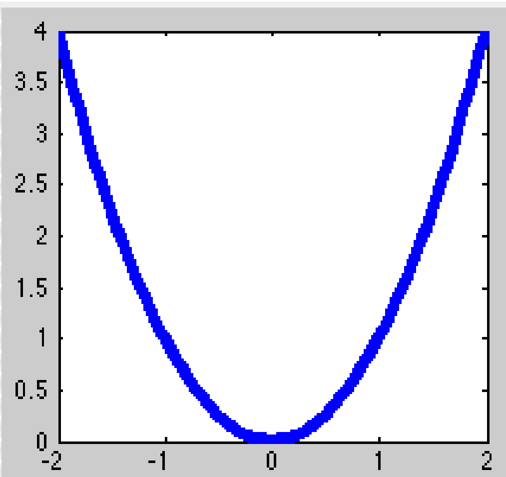
WITCH Framework 3

- Graph Regularized SVM with **Slack**: Same as before, but also learn a spam weight for each node.

$$\Omega(\mathbf{w},\mathbf{z}) = \frac{1}{l}\sum_{i=1}^{l}[1-y_i(\mathbf{w}\cdot\mathbf{x}_i+z_i)]_+ + \lambda_1\mathbf{w}\cdot\mathbf{w} + \lambda_2\mathbf{z}\cdot\mathbf{z} + \gamma\sum_{(i,j)\in E} a_{ij}((\mathbf{w}\cdot\mathbf{x}_i+z_i)-(\mathbf{w}\cdot\mathbf{x}_j+z_j))^2$$

Better Graph Regularization:

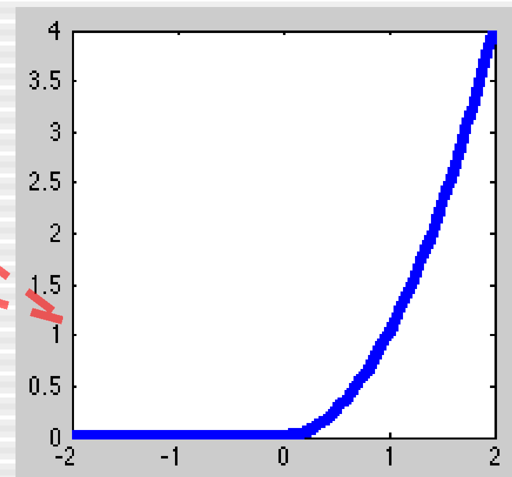
- When A links to B, penalizing the spam score as $(S_A - S_B)^2$ isn't quite right. This hurts sites that *receive* links from spam sites.



Undirected Regularization

$(S_A - S_B)^2$

Intuitively, this should be better

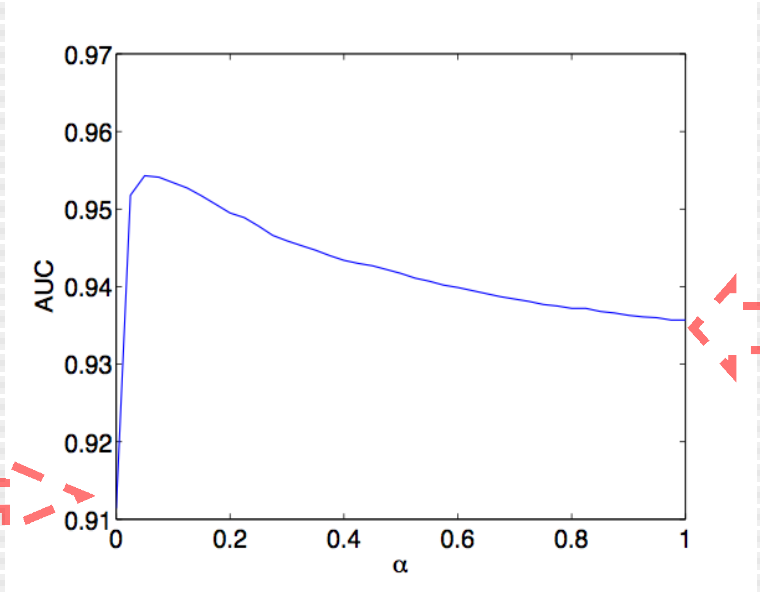


Directed Regularization

$\max(0, S_A - S_B)^2$

NOT TRUE!!

- Interestingly, the issue is more complex

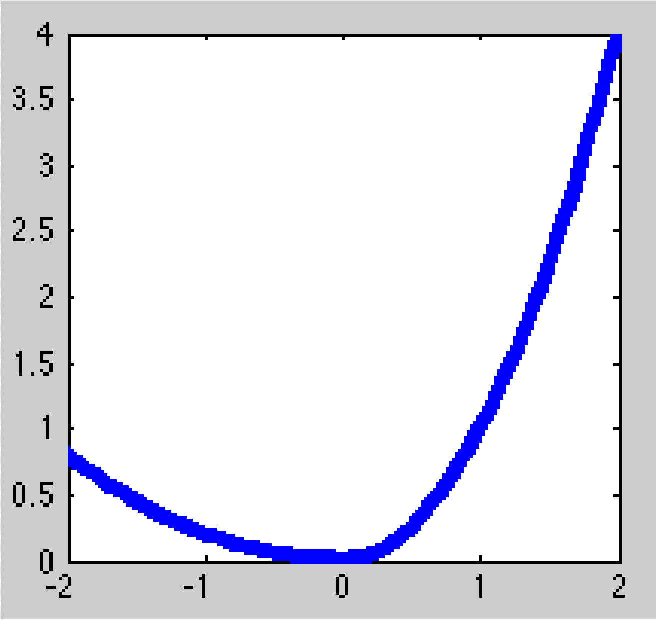


Undirected Regularization

Directed Regularization

A *mixture* of the two types of regularization is better!

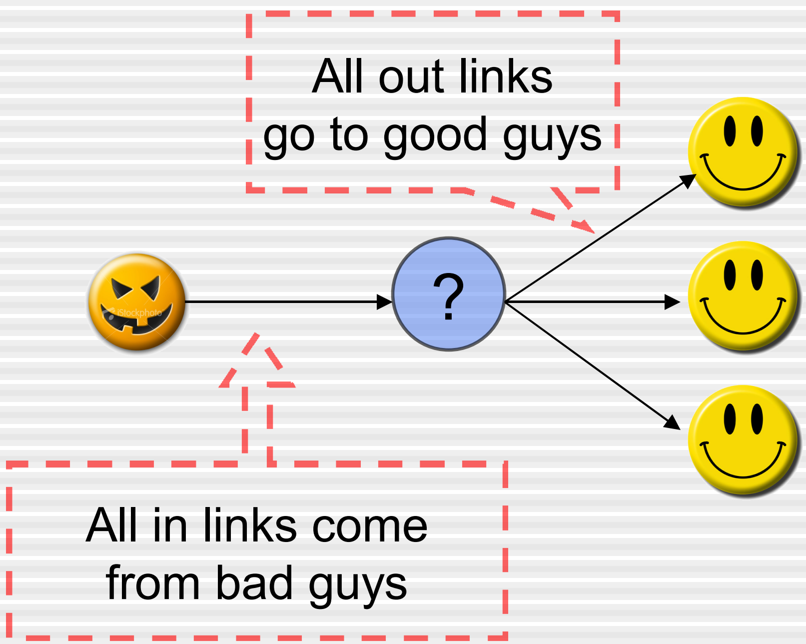
Optimal Regularizer



Semi-Directed Regularization

Seems Strange, BUT…

- Why didn't simple directed regularization work?
- It will **fail** on certain cases:

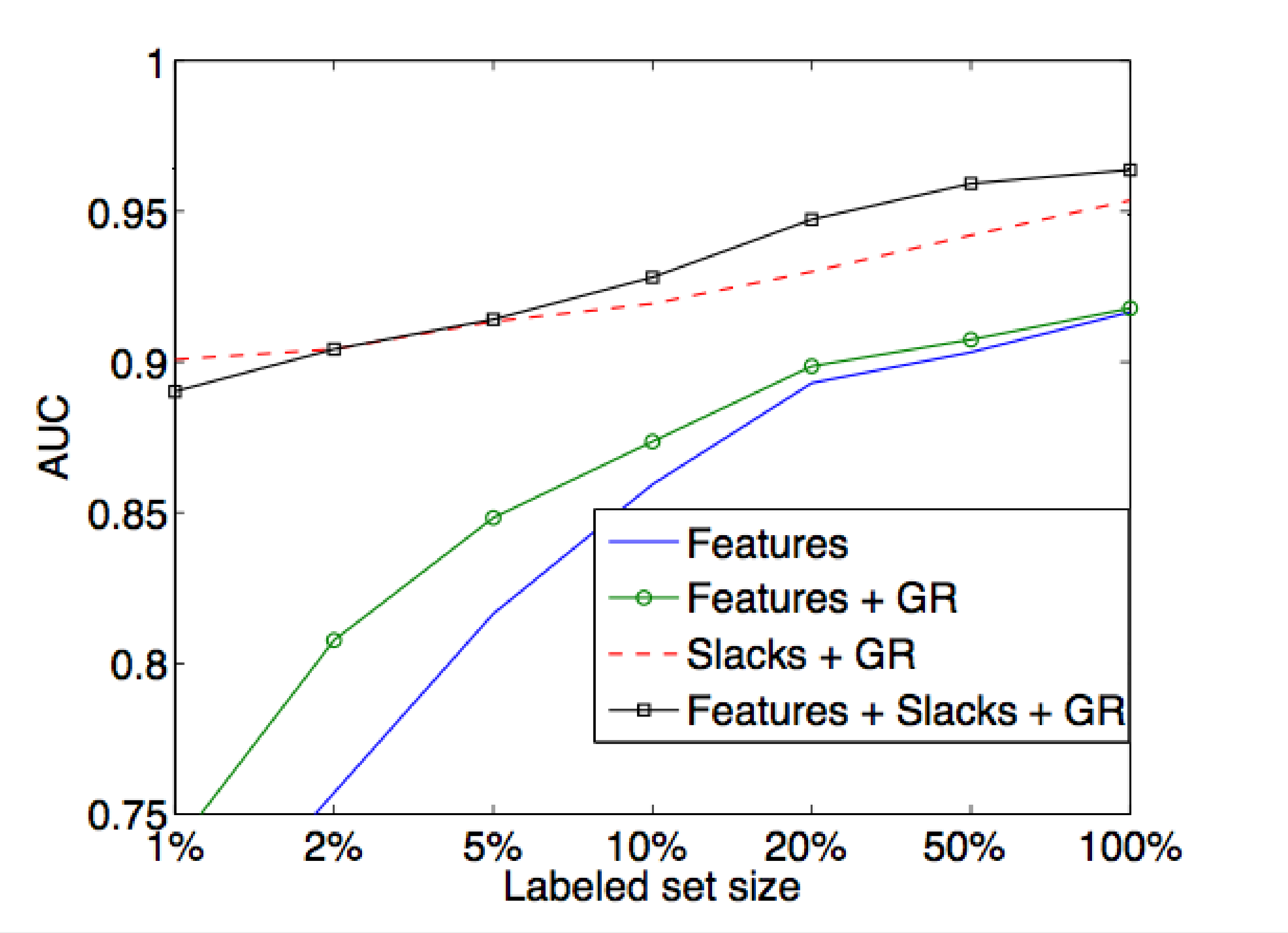


Optimization

- Roughly a Newton-method type optimization.
- Hard part is computing the Newton Step
- Can be accomplished using linear conjugate gradient, ~50 passes over data to get one approximate Hessian.
- Requires roughly 10 Newton steps

WITCH Performance Results

Performance Comparison



Web Spam Challenge

- Organized By Researchers at Yahoo! Research Barcelona and University Paris 6
- Used a web spam dataset consisting of 10,000 hosts including:
  - 1,000 labelled hosts, roughly 10% spam
  - A Hyperlink graph
  - Content-based features

Web Spam Challenge

- We won the 2nd Track of the Web spam Challenge 2007 (measured by AUC, host-level only)
- Our algorithm outperforms the winner of the Track I competition (we were too late to compete).

Performance Results

Training Algorithm	AUC 10%	AUC 100%
SVM + stacked g.l.	0.919	0.953
Link based (no features)	0.906	0.948
Challenge winner	–	0.956
Only Features	0.859	0.917
Features + GR	0.874	0.917
Slack + GR	0.919	0.954
WITCH (Feat. + Slack + GR)	**0.928**	**0.963**

Final Thoughts

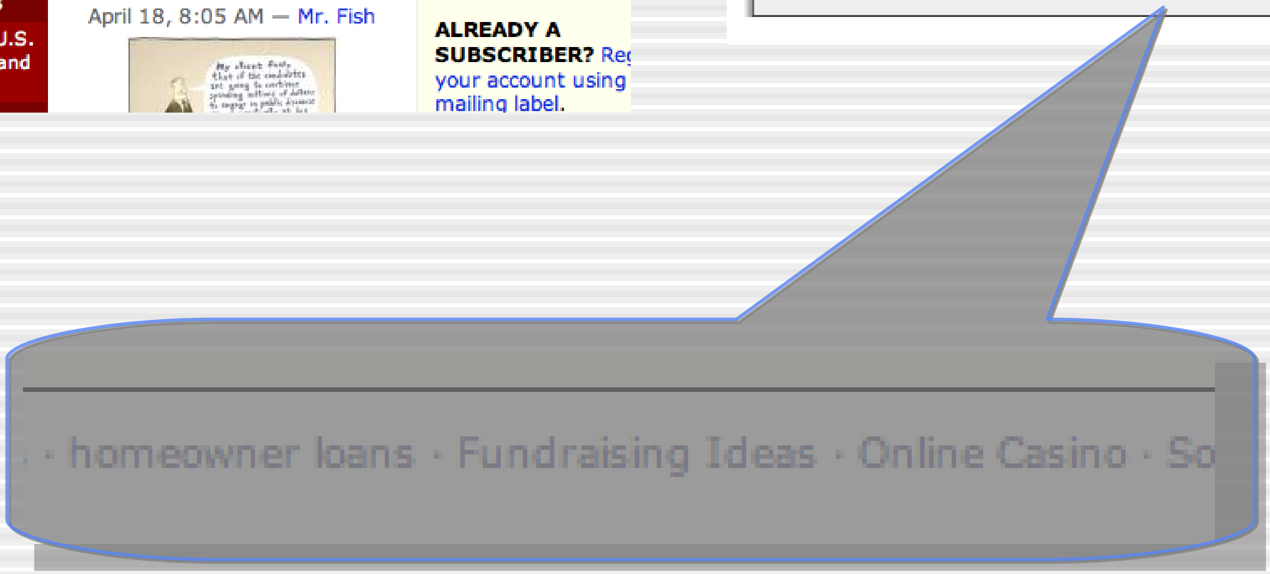
“No Good → Bad Links” Assumption?

- Perhaps good sites will link to bad sites occasionally:
  - Blog spam
  - “link swapping”
  - Harpers (thanks to reviewer for pointing this out!)
- How can we deal with this?

Harpers:


HOME
SUBSCRIBE
ARCHIVE
HARPER'S
MAGAZINE
1850 1860 1870 1880 1890 1900 1910 1920 1930
2000 2001 2002 2003 2004 2005
Jan Feb Mar Apr May Jun Jul
FIND HARPER'S
Find a U.S. newsstand
April 18, 8:05 AM — Mr. Fish
ALREADY A SUBSCRIBER?
your account using
mailing label.


[MORE FROM APRIL]
Harper's Magazine is an American journal of literature, politics, culture, and the
© The Harper's Magazine Foundation. All rights reserved. · About Harper's ·
Partners: Kids bedding · Prom dresses · Personalized Gift Ideas · Poker California · Casinos
Cross Tattoos · BCS Championship Game Tickets · Place Your Link Here!


· homeowner loans · Fundraising Ideas · Online Casino ·

Thank You!!

Questions?

(and thanks to Alexandra Meliou for the PowerPoint Animations)