The Game Theoretic Web

Web (2.0) Mining: Analyzing Social Media

Anupam Joshi

Joint work with Tim Finin and several students

Ebiquity Group, UMBC

joshi@cs.umbc.edu

http://ebiquity.umbc.edu/

UMBC
AN HONORS UNIVERSITY IN MARYLAND

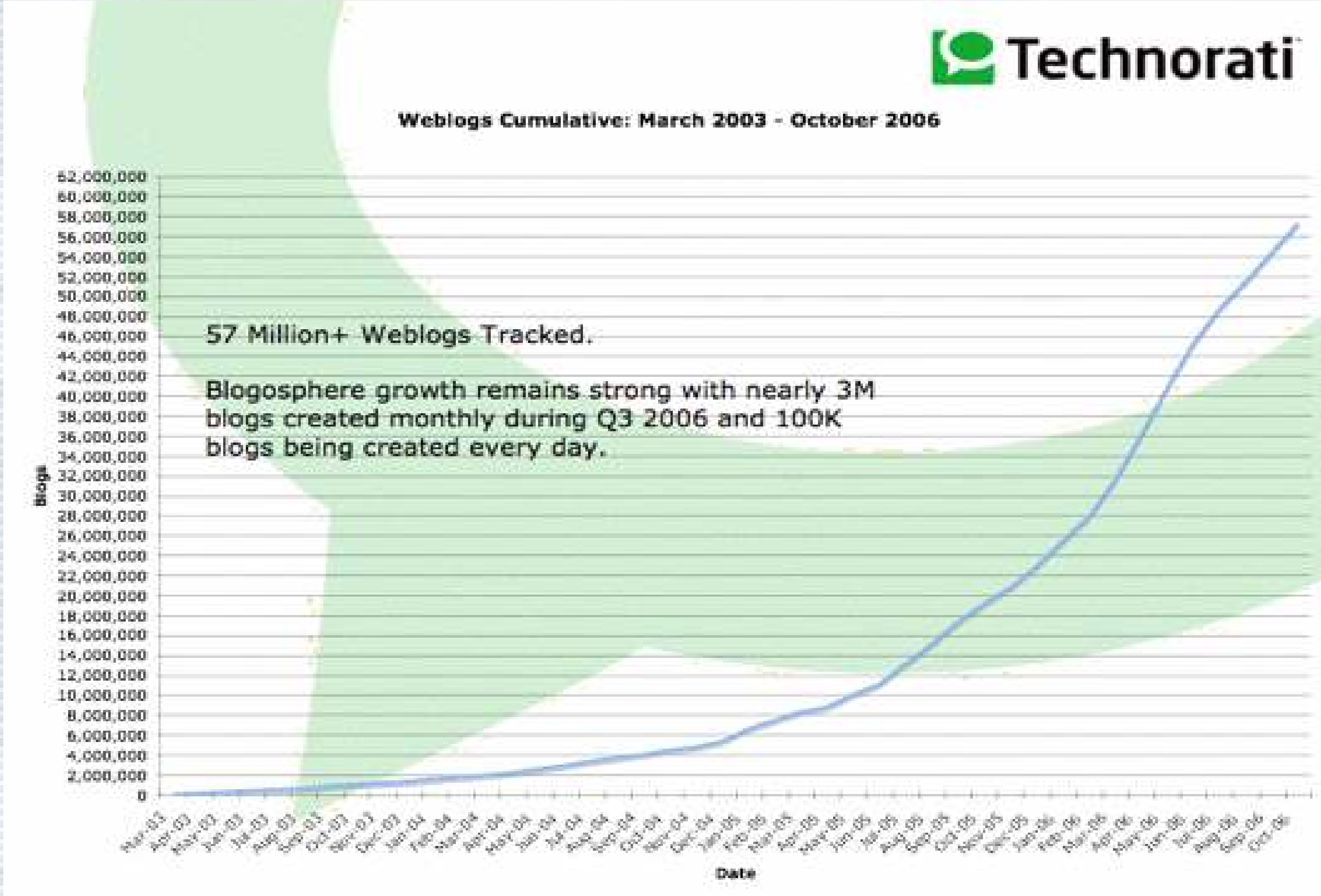
ebiquity

Social Media

- "Social media describes the online tools and platforms that people use to share opinions, insights, experiences, and perspectives" - wikipedia
- Level of user participation and thought sharing across varied topics

Twitterment beta

State of the Blogosphere



"**Blogosphere** is the collective term encompassing all blogs as a community or social network'' Wikipedia Nov 06

Knowing & Influencing your Audience

- Your goal is to campaign for a presidential candidate
- How can you track the buzz about him/her?
- What are the relevant communities and bogs?
- Which communities are supporters, which are skeptical, which are put off by the hype?
- Is your campaign having an effect? The desired effect?
- Which bloggers are influential with political audience? Of these, which are already onboard and which are lost causes?
- To whom should you send details or talk to?

Knowing & Influencing your Market

- Your goal is to market Apple's iPhone
- How can you track the buzz about it?
- What are the relevant communities and blogs?
- Which communities are fans, which are suspicious, which are put off by the hype?
- Is your advertising having an effect? The desired effect?
- Which bloggers are influential in this market? Of these, which are already onboard and which are lost causes?
- To whom should you send details or evaluation samples?

Opinions in Social Media

"Last night in Boston at a mid-dollar fundraiser John Edwards gave a fantastic speech. It was one the loosest most charismatic speeches I have seen him give. Many of the points and line were from his standard stump speech but there was a definite confidence and sense of humor in his delivery.

He also dwelled on the environment more than I have seen him do in other speeches. The environmental section kicked off with with a good and true line that got a big ovation: "On global warming: Al Gore was right."........"[1]

Reader's Perspective
"John Edwards is Good!"

Expressed Opinions

Opinions can influence the votes of others

[1]http://www.dailykos.com/storyonly/2007/10/4/71218/3740

What is Influence?

"the act or power of producing an effect without apparent exertion of force or direct exercise of command"

Measurable Influence

The ability of a blogger to persuade another blogger to

- Take action by means of creating a new post about the topic and commenting on the original ***(text and graph mining)*** .
- Quote the blogger's views in her post ***(text mining)*** .
- Link to the original post via trackbacks, comments ***(graph mining)*** .
- Link to the blogger through other means like del.icio.us, digg, citeULike, Connotea, etc. ***(graph mining)***
- Subscribe to the blog feed ***(graph mining)*** .

Epidemic-based Influence Models

“Find the minimum set of nodes, influencing which would maximize the infection in the network”

- Kemp et al.

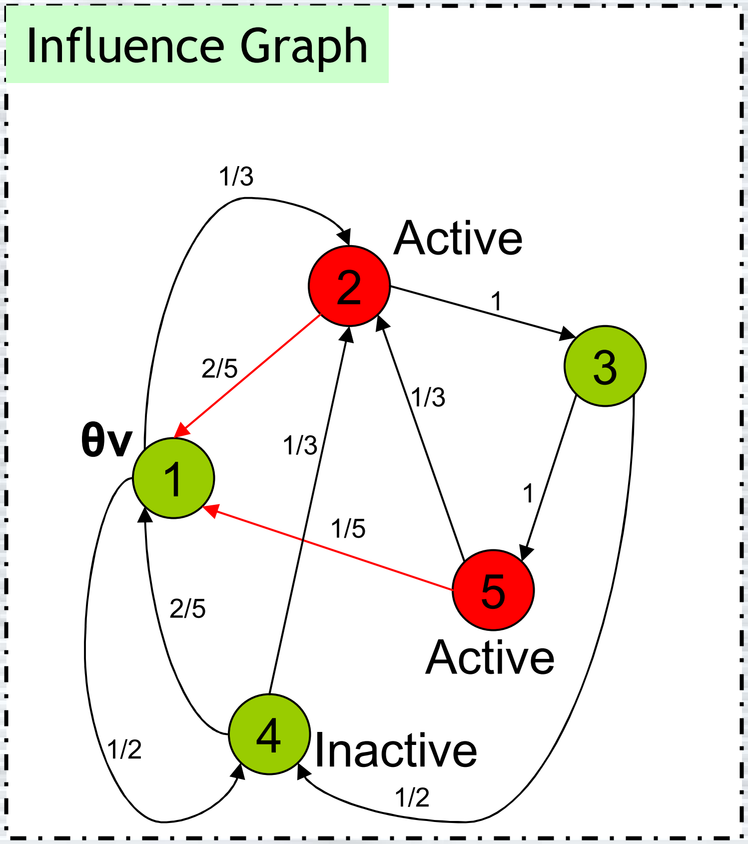
- Linear Threshold Model

  $\Sigma\, b_{wv} \geq \theta_v$

  w is the active neighbor of v,

  θ_v intrinsic threshold for a node

- Greedy Heuristic
  - Assign random θ_v
  - Compute approx influenced set
  - At each step, add the node that increases the marginal gain in the size of the influenced set



Other approaches: Latane’, iRank,

Limitations of Existing Approaches

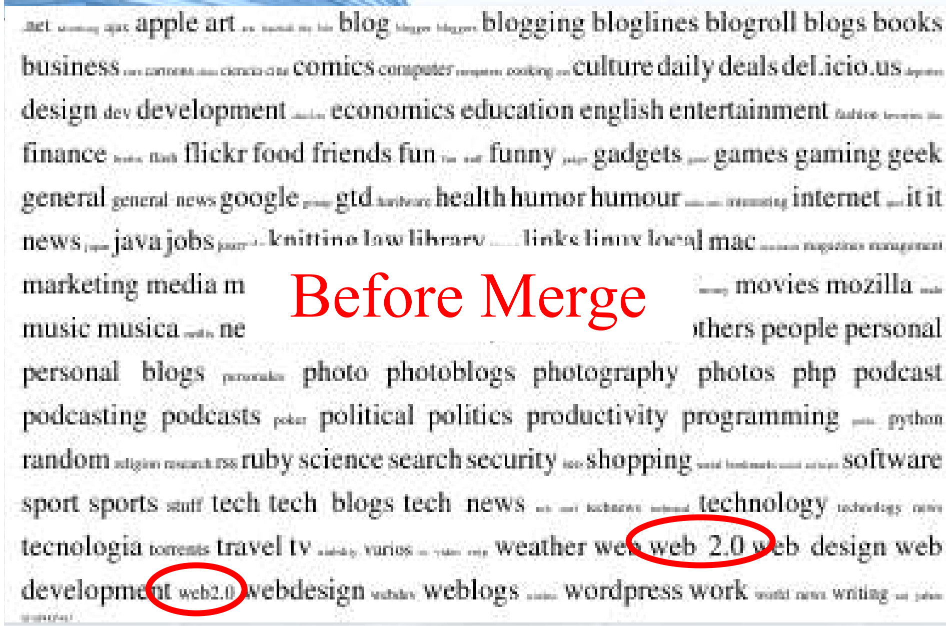
- Selected nodes may belong to different topics
- Opinions or bias not considered
- Information is spread throughout the network without considering social structure
  - Intrinsic threshold θ_v is based on a pseudorandom function
- Static view of the network, no temporal evidence

First 10 nodes selected using Greedy Hill-Climbing Heuristic

http://www.engadget.com
http://www.boingboing.net
http://www.dailykos.com
http://postsecret.blogspot.com
http://slashdot.org
http://www.albinoblacksheep.com
http://www.opinionjournal.com
http://profiles.blogdrive.com
http://godlessmom.blogspot.com
http://thinkprogress.org

TECH, POLITICS, DAILY/NEWS

Finding Communities (and Feeds) That Matter



Analysis of Bloglines Feeds

- 83K publicly listed subscribers
- 2.8M feeds, 500K are unique
- 26K users (35%) use folders to organize subscriptions
- Data collected in May 2006

Top Advertising Feeds

1. Adrants » Marketing and Advertising News With Attitude
2. Adverblog: advertising and new media marketing
3. http://ad-rag.com
4. adfreak
5. AdJab
6. MIT Advertising Lab: future of advertising and advertising technology
7. AdPulp: Daily Juice from the Ad Biz
8. Advertising/Design Goodness

Related Tags: advertising marketing media news design

Feeds That Matter

Top Feeds for "Politics"

Merged folders: "political", "political blogs"

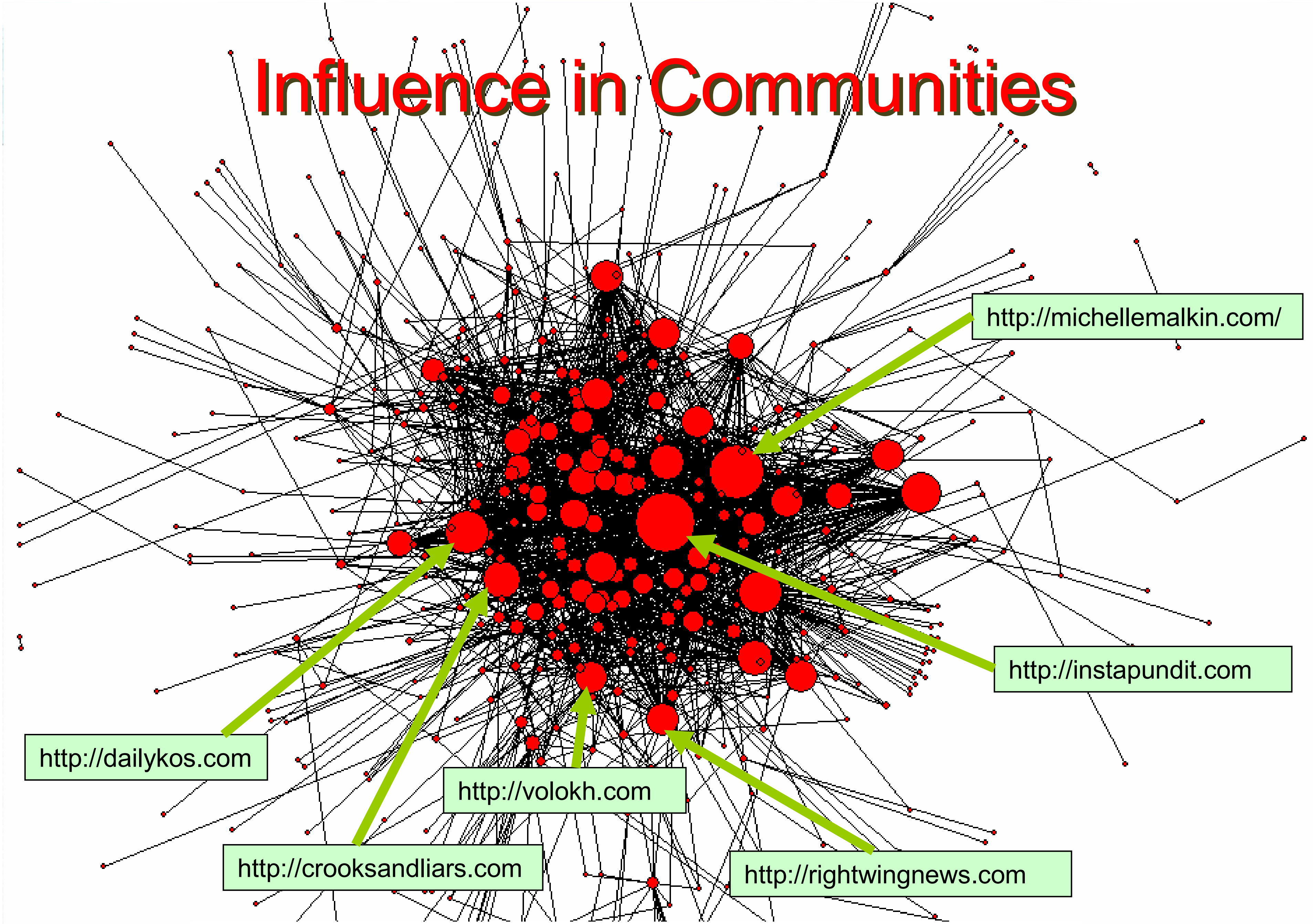
- Talking Points Memo: by Joshua Micah Marshall
- Daily Kos: State of the Nation
- Eschaton
- The Washington Monthly
- Wonkette, Politics for People with Dirty Minds
- http://instapundit.com/
- Informed Comment
- Power Line
- AMERICAblog: Because a great nation deserves the truth
- Crooks and Liars

Top Feeds for "Knitting"

Merged folders "knitting blogs"

- Yarn Harlotknitting
- Wendy Knits!
- See Eunny Knit!
- the blue blog
- Grumperina goes to local yarn shops and Home Depot
- You Knit What??
- Mason-Dixon Knitting
- knit and tonic
- Crazy Aunt Purl
- http://www.lollygirl.com/blog/

Influence in Communities



Communities detected using "Fast algorithm for detecting community structure in networks", M.E. J. Newman

Authority and Popularity

Authority

- contributes to influence
- Influence may be subjective.
- A source, authoritative in one community could influence another community negatively.

Within a community, an authoritative source would be influential.

Popularity

- Authority and popularity often treated equally
- On blog search engines, authority is measured using inlinks, which is at best popularity
- Popularity doesn't mean influence
  - Dilbert is extremely popular but not influential

Link Polarity / Bias

- Linking alone is not indicator of influence
- Polarity can indicate the type of influence
- Consistent negative / positive opinion over a period of time can indicate bias
- Link polarity/citation signal can also be helpful in determining trust

Our Approach to Link Polarity

- Shallow Sentiment Analysis
  - Calculate the number of positively oriented (*Np*) and Negatively oriented words (*Nn*) in the text-window around the link
  - Apply Stemming, basic canonicalization
  - Corpus includes simple bi-grams of the form "*not_good*"
- Polarity = (Np – Nn) / (Np + Nn)
  - Denominator acts as a normalization mechanism
- Natural Language Processing is *shallow*, yet large-scale effects help !

Link Polarity Example

- “Stephen Colbert's performance at the White House Correspondents' Association dinner has garnered him **huge applause** in the blogosphere and also on C-Span where it was shown more than once. Those of us who have been **angry** with Bush for quite some time because of his **arrogant** and feckless **corruption** of our country were even more thrilled to see and know that he had no recourse but to sit there and watch his aspirations for greatness be destroyed by a **master** of irony.

  This will be his **legacy**: I stand by this man. I stand by this man because he stands for things. Not only for things, he stands on things. Things like aircraft carriers and rubble and recently flooded city squares. And that sends a **strong** message, that no matter what happens to America, she will always rebound -- with the most **powerfully** staged photo ops in the world. We who have been watching Stephen Colbert eviscerate politicians that have come on his show knew he was a **gifted** comedian. But it took Saturday's dinner to demonstrate how incredibly **effective** the art form Colbert has chosen is for exposing the Potemkin Regime Bush and his henchmen have created. Rove and the right wing machine have no answer to the performance but to say "it **bombed**", "it wasn't funny", and to hope that by ignoring it, the caustic cleansing agent it has lobbed into their camp can be contained. Yet, the Republican spinmeisters are the masters of spin.”[2]

This - http://dailykos.com/storyonly/2006/4/30/1441/59811

Np = 8, Nn = 4 ; Polarity = Np – Nn / Np + Nn = 0.33

[2]http://www.pacificviews.org/weblog/archives/001989.html

Propagating Influence

- Based on work of Guha et al[1] for modeling propagation of trust and distrust
- Framework
  - M_{ij} represents influence/bias from user i to j.($0 <= M_{ij} <= 1$)
  - M_{ij} is initialized to the polarity from i to j.
  - Belief Matrix *M* represents the initial set of known beliefs, and is sparse
  - Goal is to compute all unknown values in M
  - Belief Matrix after ith atomic propagation
    - $M_{i+1} = M_i * C_i$
  - Combined Operator
    - $C_i = a_1 * M + a_2 * M^T*M + a_3 * M^T + a_4 * M*M^T$
    - a {0.4, 0.4, 0.1, 0.1} represents weighing factor

[1] Guha R, Kumar R, Raghavan P, Tomkins A. Propagation of trust and distrust. In: *Proceedings of the Thirteenth International World Wide Web Conference, New York, NY, USA, May 2004. ACM Press, 2004.*

Experiments

- Domain
  - Political Blogosphere
  - Dataset from Buzzmetrics[2] provides post-post link structure over 14 million posts
  - Few off-the-topic posts help aggregation
  - Potential business value

- Reference Dataset
  - Hand-labeled dataset from Lada Adamic et al[3] classifying political blogs into right and left leaning bloggers
  - Timeframe : 2004 presidential elections, over 1500 blogs analyzed
  - Overlap of 300 blogs between Buzzmetrics and reference dataset

- Goal
  - Classify the blogs in Buzzmetrics dataset as democrat and republican and compare with reference dataset

[2] Lada A. Adamic and Natalie Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop

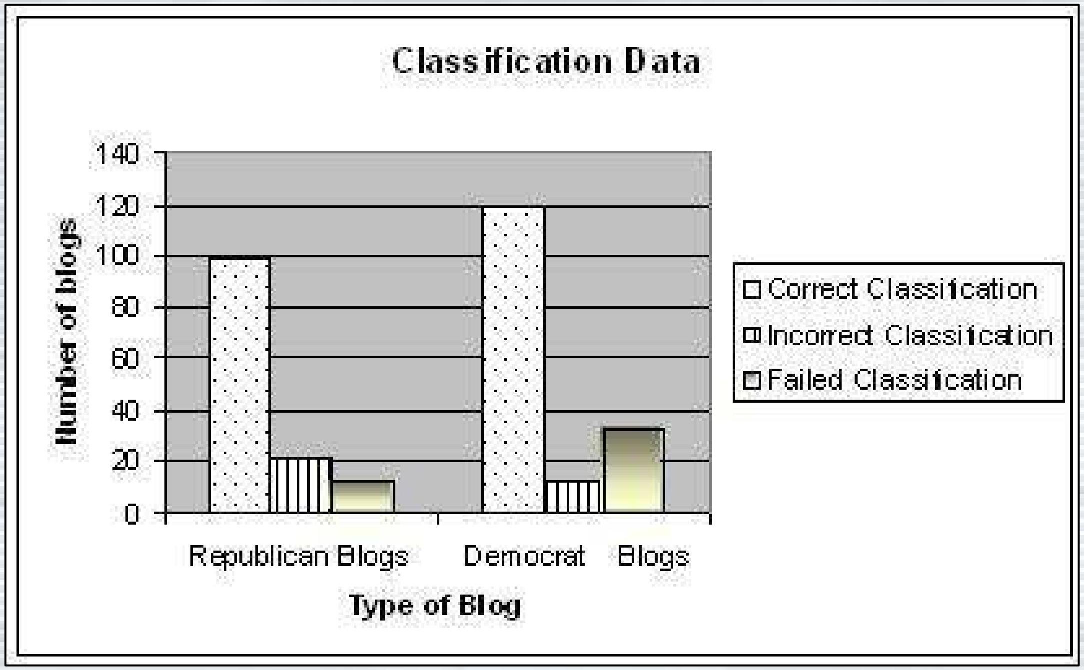
Buzzmetrics – www.buzzmetrics.com

Evaluation Metrics

Polarity Improves Classification by almost 26%

Confusion Matrix

		Predicted	
		Democrat	Republican
Actual	Democrat	99	45
	Republican	33	120



- Accuracy = 73%
- True Positive Rate (Recall) = 78%
- False Positive Rate (FP) = 31%
- True Negative Rate (Recall) = 69%
- False Negative Rate (FN) = 21%
- Precision (R) = 75%
- Precision (D) = 72%
- (

Sample Data

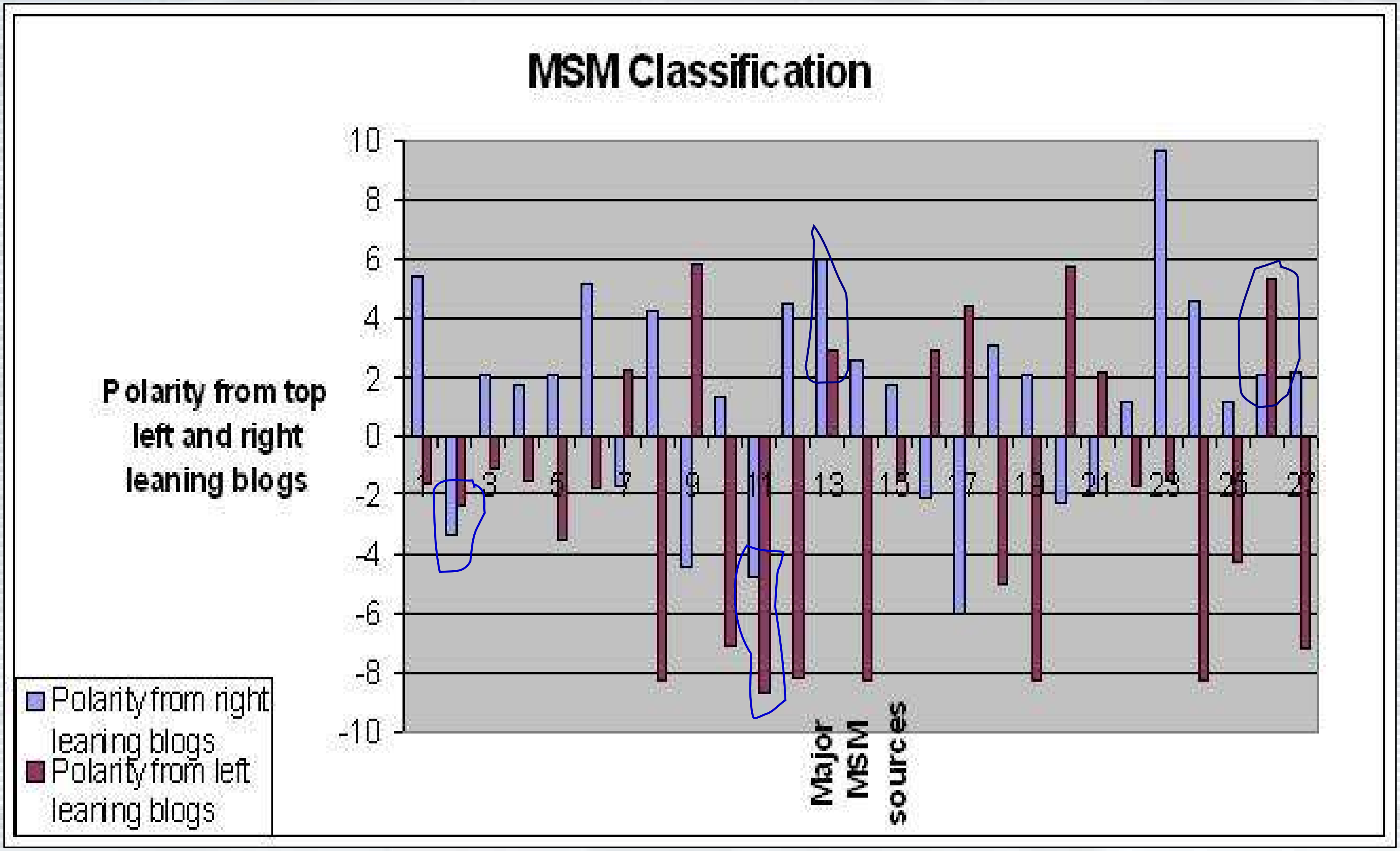
- Trust propagation compensates for initial incorrect polarity (DK – AT)
- Trust propagation does not change correct polarity (AT-DK)
- Trust propagation assigns correct polarity for non-existent direct links (AT-IP)
- Numbers in *italics* problematic (MM-AT)
  - Improve sentiment detection ?

Table 4.1. Polarity Values for Sample Influential Blogs

From-To	Number of links	Polarity before trust propagation	Polarity after trust propagation
MM–MM	0	N/A	3.53
MM–DK	0	N/A	-2.9
MM-IP	0	N/A	2.2
MM-AT	0	N/A	*1.09*
DK-MM	0	N/A	-2.9
DK-DK	0	N/A	2.02
DK-IP	0	N/A	*1.71*
DK-AT	20	0	8.51
IP-MM	8	1	2.2
IP-DK	6	0	*1.71*
IP-IP	0	N/A	1.06
IP-AT	0	N/A	-7.19
AT-MM	0	N/A	*1.09*
AT-DK	5	0.342	8.51
AT-IP	0	N/A	7.19
AT-AT	0	N/A	3.57

MM–http://michellemalkin.com, DK–http://dailykos.com
IP–http://instapundit.com, AT–http://atrios.blogspot.com

MSM Classification Results



Interesting Observations

- 24 out of 27 sources classified “correctly”
  - guardian, foxnews, humaneventsonline, mediamatters
- Main Outliers -- “thenation” and “boston globe”
- Both left and right leaning blogs talk negatively about “nytimes” and “abcnews” and positively about “rawstory” and “examiner”

1	http://www.washingtonpost.com	15	http://www.truthout.org
2	http://www.nytimes.com	16	http://today.reuters.com
3	http://news.yahoo.com	17	http://mediamatters.org
4	http://news.bbc.co.uk	18	http://www.townhall.com
5	http://www.msnbc.msn.com	19	http://www.timesonline.co.uk
6	http://www.cnn.com	20	http://www.guardian.co.uk
7	http://news.google.com	21	http://www.salon.com
8	http://www.usatoday.com	22	http://www.thenation.com
9	http://www.latimes.com	23	http://apnews.myway.com
10	http://www.boston.com	24	http://www.xaminr.com
11	http://www.abcnews.go.com	25	http://www.humaneventsonline.com
12	http://www.foxnews.com	26	http://www.dailybulltin.com
13	http://www.rawstory.com	27	http://www.spectator.org
14	http://www.cbsnews.com		

Identifying Bias using KL Divergence

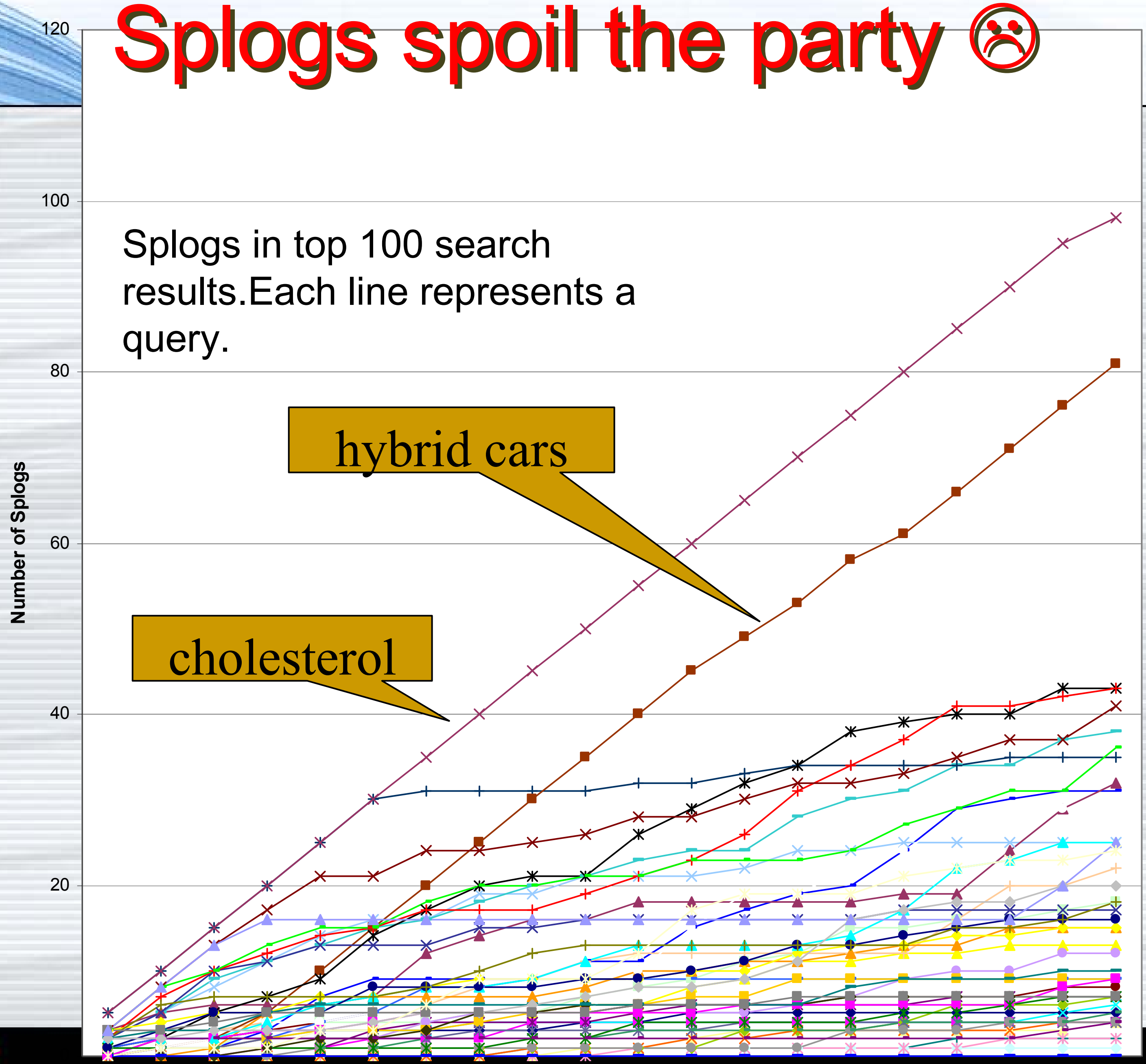
MSM sources for Democrats

Rank	MSM	Links from Dems	Links from Reps
1	http://mediamatters.org	76 from 28 blogs	5 from 4 blogs
2	http://www.rawstory.com	108 from 38 blog	14 from 11 blogs
3	http://www.nytimes.com	503 from 83 blogs	199 from 50 blogs
4	http://www.alternet.org	38 from 19 blogs	2 from 2 blogs
5	http://www.washingtonpost.com	750 from 91 blogs	355 from 61 blogs
6	http://news.independent.co.uk	59 from 20 blogs	5 from 5 blogs
7	http://www.salon.com	48 from 25 blogs	8 from 2 blogs
8	http://www.truthout.org	85 from 35 blogs	24 from 10 blogs
9	http://www.usatoday.com	168 from 55 blogs	71 from 36 blogs
10	http://www.thenation.com	29 from 17 blogs	4 from 3 blogs

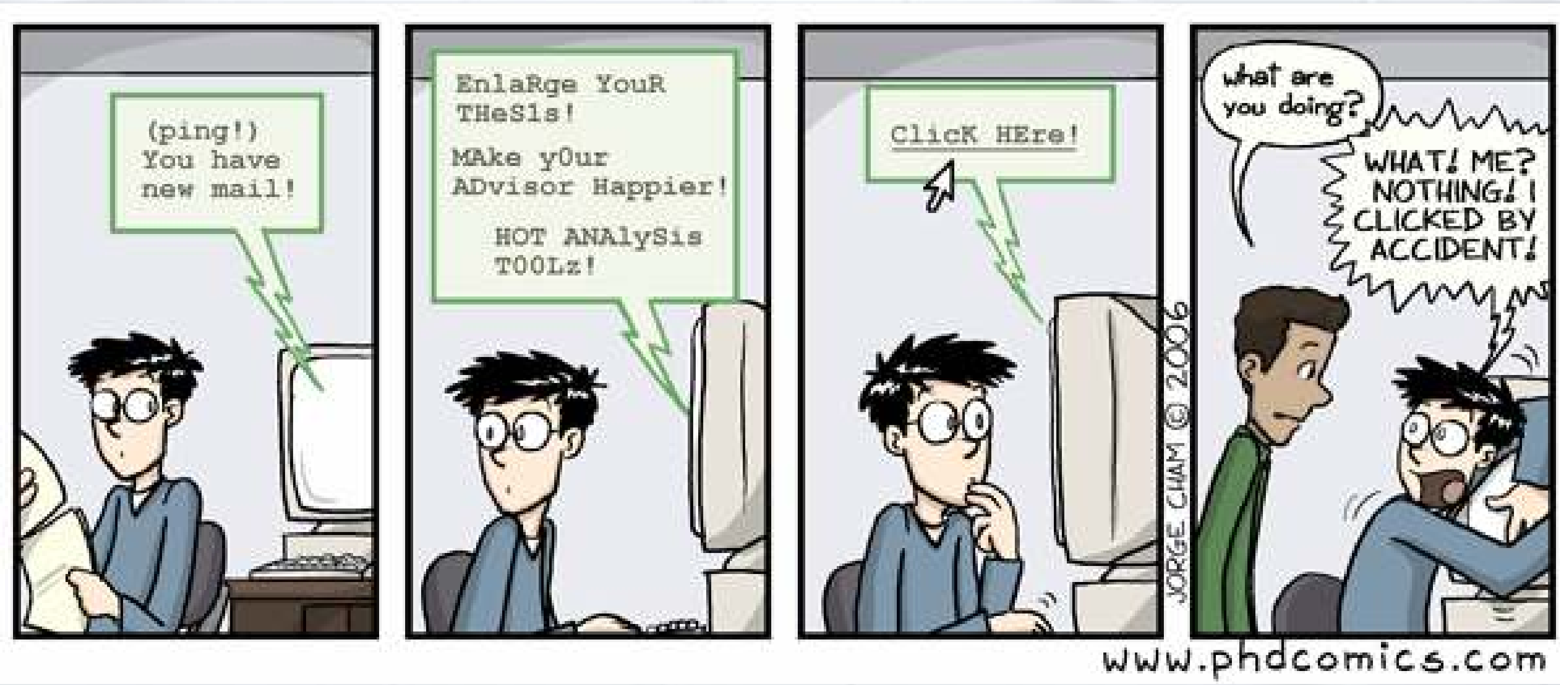
MSM sources for Republicans

Rank	MSM	Links from Dems	Links from Reps
1	http://www.washingtontimes.com	17 from 11 blogs	65 from 33 blogs
2	http://www.foxnews.com	64 from 23 blogs	165 from 44 blogs
3	http://apnews.myway.com	4 from 3 blogs	33 from 17 blogs
4	http://www.examiner.com	4 from 4 blogs	23 from 17 blogs
5	http://www.frontpagemag.com	3 from 3 blogs	23 from 13 blogs
6	http://www.humaneventsonline.cor	6 from 5 blogs	22 from 16 blogs
7	http://www.townhall.com	31 from 8 blogs	72 from 24 blogs
8	http://www.dailybulletin.com	5 from 3 blogs	19 from 14 blogs
9	http://www.sacbee.com	0 from 0 blogs	6 from 6 blogs
10	http://www.spectator.org	5 from 3 blogs	17 from 11 blogs

Splogs spoil the party ☹



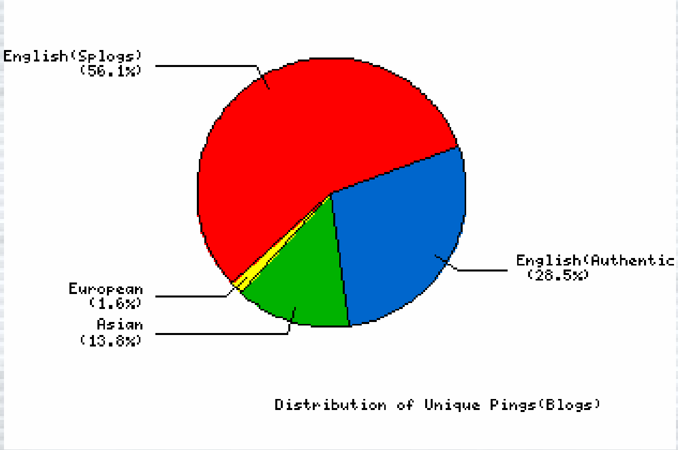
SPLOGS!


(ping!) You have new mail!
EnlaRge YouR THeS1s!
MAke y0ur ADvisor Happier!
HOT ANAlySis TOOLz!
ClicK HEre!
JORGE CHAM © 2006
what are you doing?
WHAT! ME? NOTHING! I CLICKED BY ACCIDENT!
www.phdcomics.com

SPLOGS BY NUMBERS

- 75% of update pings (eBiquity 2006)
- 20% of indexed Blogosphere (Umbria 2006)
- 56% of update pings (eBiquity 2007)

56% of all active blogs are splogs! (2007)



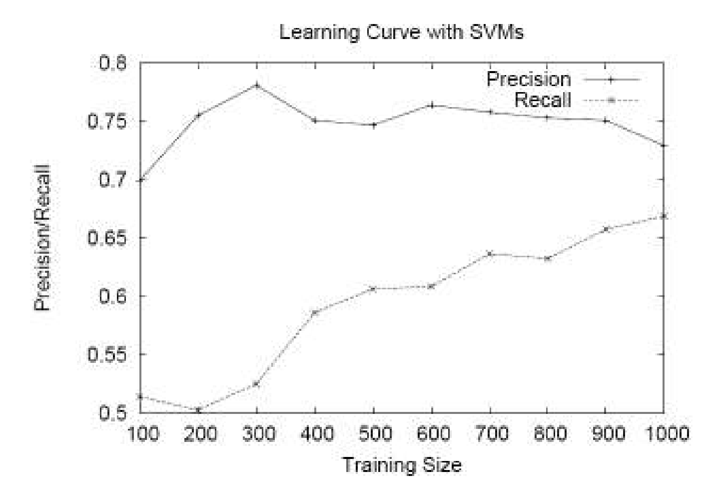
DATASETS

- SPLOG-2005
  - Sampled Summer 2005 at Technorati
    - A search engine, so many splogs already removed
  - Labeled samples of 700 blogs and 700 splogs
  - Only Blog-homepages
- SPLOG-2006
  - Sampled Oct 2006 at Weblogs.com
  - Labeled samples of 750 blogs and 750 splogs
  - Blog-homepages + feeds

EXPERIMENTAL SETUP

- Binary feature encoding
- Top 50K selected using frequency count
- SVMs
  - Default parameters
  - Linear Kernel
- No stemming or stop word elimination
- Naïve Bayes
- Ten fold cross-validation

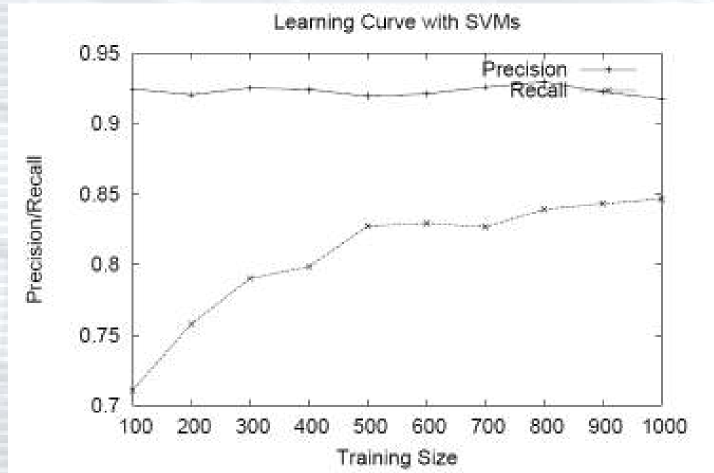
URL



	P	R	F1
SVM	0.70	0.70	0.70
NB	0.70	0.67	0.69

Authentic
fif, sig, yww, lee
enk, mod, hop, dae
ose, edu, mode, bab
baby, aby, ile, blu
evie, file, evi, hat
Spam
nhg, vkq, hot, mat
chao, ree, urs, herb
cha, she, shev, hev
ool, karl, rlz, des
info, ate, inf, ies

2005 | 2006



	P	R	F1
SVM	0.92	0.88	0.90
NB	0.82	0.90	0.85

Authentic
law, xre, org, cha
hds, ibn, bnl, ibnl
bnliv, bnli, ibnli, clau
poo, rea, log, lau
aus, rma, webl, weblo
Spam
htm, imv, nfo, info
inf, car, eac, each
abe, ach, blogs, job
sta, grac, grace, ogs
logs, star, ace, loan

URL

- 3,4,5 charactergrams from URL
- Captures profitable contexts
- Highly effective at ping streams
- Supports an extremely low cost classifier

2005

	P	R	F1
SVM	0.70	0.70	0.70
NB	0.70	0.67	0.69

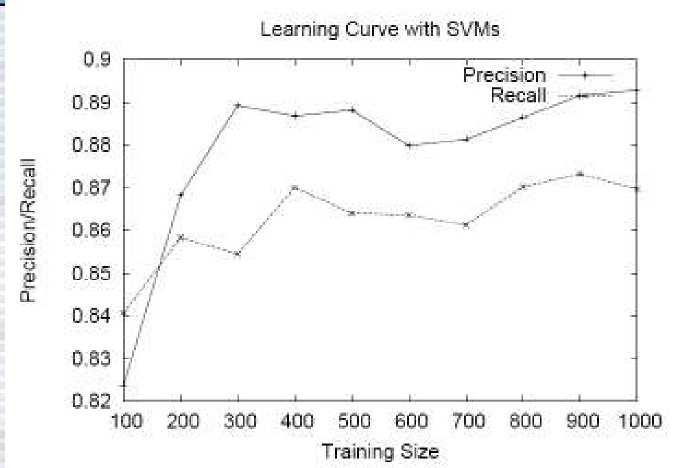
Authentic
fif, sig, yww, lee
enk, mod, hop, dae
ose, edu, mode, bab
baby, aby, ile, blu
evie, file, evi, hat
Spam
nhg, vkq, hot, mat
chao, ree, urs, herb
cha, she, shev, hev
ool, karl, rlz, des
info, ate, inf, ies

2006

	P	R	F1
SVM	0.92	0.88	0.90
NB	0.82	0.90	0.85

Authentic
law, xre, org, cha
hds, ibn, bnl, ibnl
bnliv, bnli, ibnli, clau
poo, rea, log, lau
aus, rma, webl, weblo
Spam
htm, imv, nfo, info
inf, car, eac, each
abe, ach, blogs, job
sta, grac, grace, ogs
logs, star, ace, loan

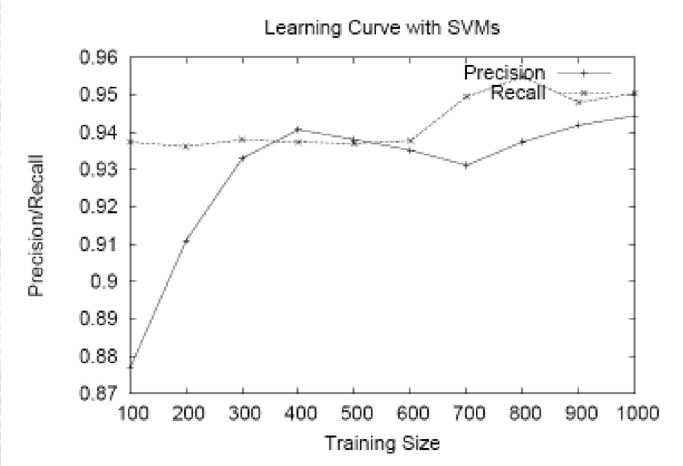
WORDS



	P	R	F1
SVM	0.90	0.88	0.89
NB	0.82	0.85	0.83

Authentic
my, we, s, new, this,
please, what, org, log, linux
gallery, words, paper, jane,
political, web, september,
reseller, flickr, open
Spam
news, your, on, find, previous,
info, uncategorized, best, information,
top, posted, com, laquo, related,
business, articles, august, may,
action, looking

2005 | 2006



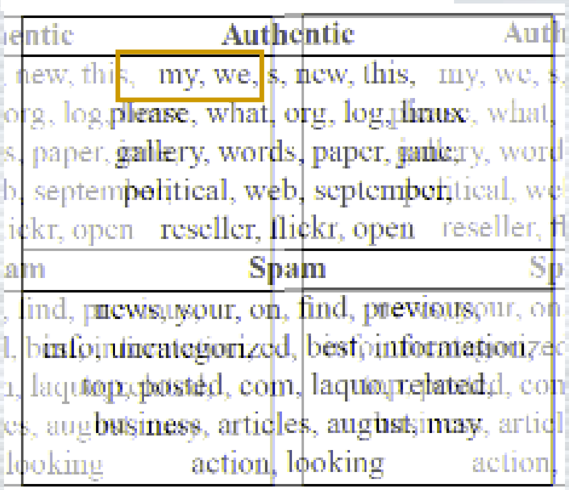
	P	R	F1
SVM	0.95	0.95	0.95
NB	0.92	0.92	0.92

Authentic
location, april, february, june
march, pm, november, creative
friday, july, may, thursday
link, article, fun, wednesday
privacy, down, today, read
Spam
by, net, posts, technorati
tag, edit, at, sitemap
as, to, start, info
friendly, and, free, news
buy, sponsors, copyright, ads

WORDS

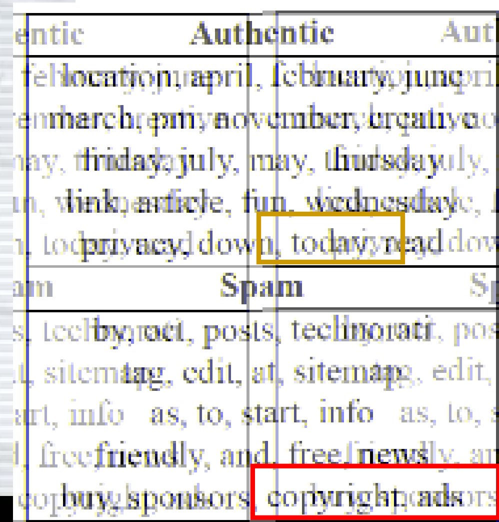
- Words (Text) on a Blog
- Previously effective in topic classification
- Captures profitable advertising contexts
- Interesting Authentic Genre Observed

	P	R	F1
SVM	0.90	0.88	0.89
NB	0.82	0.85	0.83

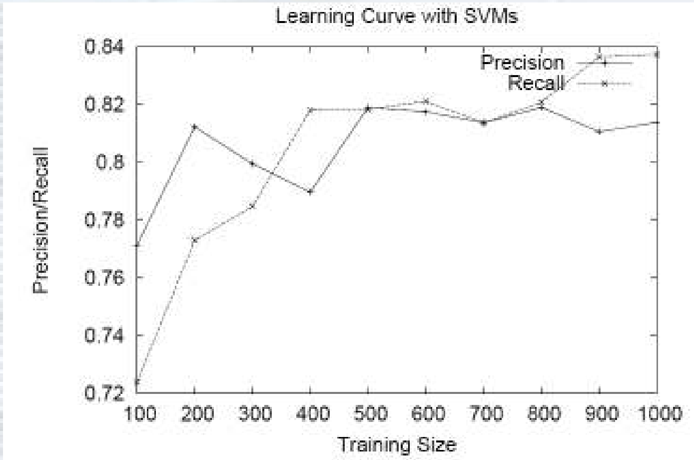


2005 2006

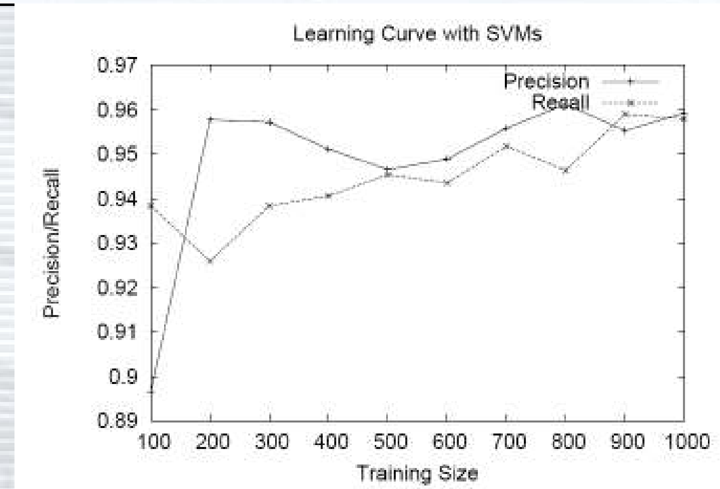
	P	R	F1
SVM	0.95	0.95	0.95
NB	0.92	0.92	0.92



OUTLINKS



	P	R	F1
SVM	0.81	0.83	0.82
NB	0.79	0.81	0.80



	P	R	F1
SVM	0.95	0.96	0.96
NB	0.95	0.57	0.71

Authentic
rudayday, weblog, archives
October, august, id
sundaymornings, email, jpg
mailto, september, photos
brin, org, of
Spam
info, com, prop
cessna, technorati, solution
page, proactiv, mybeautyadviceblog
tag, www, profile
comment, google, post

2005 2006

OUTLINKS

- Out-links tokenized by non-alphabets
- Similar to URL n-grams, likely more robust

- **Novel feature space**

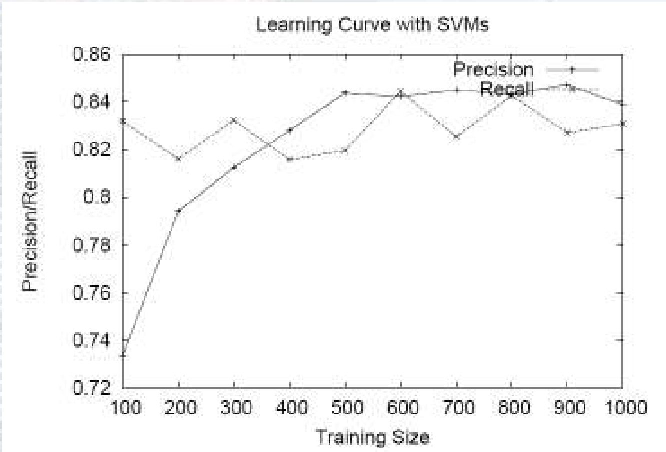
	P	R	F1
SVM	0.81	0.83	0.82
NB	0.79	0.81	0.80

	P	R	F1
SVM	0.95	0.96	0.96
NB	0.95	0.57	0.71

Authentic
rudayday, weblog, archives
October, august, id
sundaymornings, email, jpg
mailto, september, photos
brn, org, of
Spam
info, com, prop
cessna, technorati, solution
page, proactiv, mybeautyadviceblog
tag, www, profile
comment, google, post

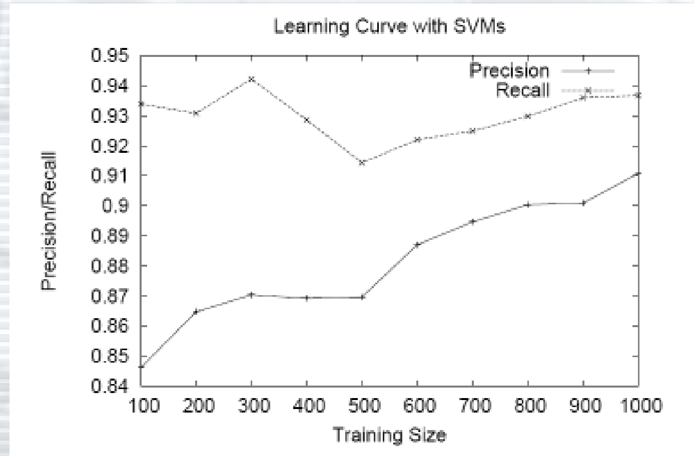
2005 2006

ANCHORS



	P	R	F1
SVM	0.84	0.85	0.85
NB	0.83	0.82	0.82

Authentic
greymatter, rant, monk, chapitre
terrorism, comment, jane, the
postcount, permalink, archives, disclaimer
flickr, trackback, journals, about
s, space, report, random
Spam
read, chapter, revisionaryjen, generation
ii, laquo, lost, more
biz, to, top, jaguar
soulessencehealing, now, used, directory
august, free, town, an



	P	R	F1
SVM	0.92	0.94	0.93
NB	0.88	0.56	0.68

Authentic
december, site, about, flickr
july, links, august, this
september, november, memories, link
here, february, march, projects
archives, photos, email, article
Spam
prop, start, comments, nbsp
by, edit, google, for
sitemap, and, oceanriver, freedom
search, hawaii, university, xhtml
news, to, mmorpgsource, superforum

2005 | 2006

ANCHORS

- Anchor text tokenized into words
- Subsumed by words, but obfuscation difficult
- Capture personalization of publishing template
- **Novel feature space**

	P	R	F1
SVM	0.84	0.85	0.85
NB	0.83	0.82	0.82

Authentic
greymatter, rant, monk, chapitre
terrorism, comment, jane, the
postcount, permalink, archives, disclaimer
flickr, trackback, journals, about
s, space, report, random
Spam
read, chapter, revisionaryjen, generation
ii, laquo, lost, more
biz, to, top, jaguar
soulessencehealing, now, used, directory
august, free, town, an

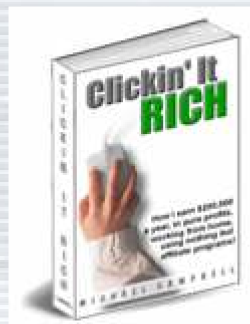
	P	R	F1
SVM	0.92	0.94	0.93
NB	0.88	0.56	0.68

Authentic
december, site, about, flickr
july, links, august, this
september, november, memories, link
here, february, march, projects
archives, photos, email, article
Spam
prop, start, comments, nbsp
by, edit, google, for
sitemap, and, oceanriver, freedom
search, hawaii, university, xhtml
news, to, mmorpgsource, superforum

2005 | 2006

Splog software ?!

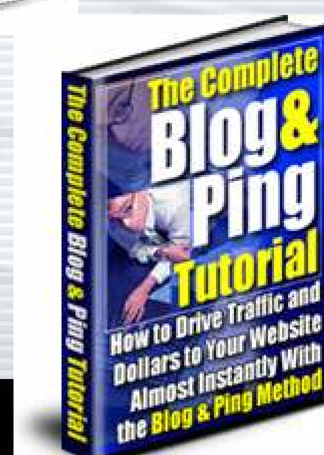
"Honestly, Do you think people who make $10k/month from adsense make blogs manually? Come on, they need to make them as fast as possible. Save Time = More Money! It's Common SENSE! How much money do you think you will save if you can increase your work pace by a hundred times? Think about it…"



"Discover The Amazing Stealth Traffic Secrets Insiders Use To Drive Thousands Of Targeted Visitors To Any Site They Desire!"

"Holy Grail Of Advertising… "

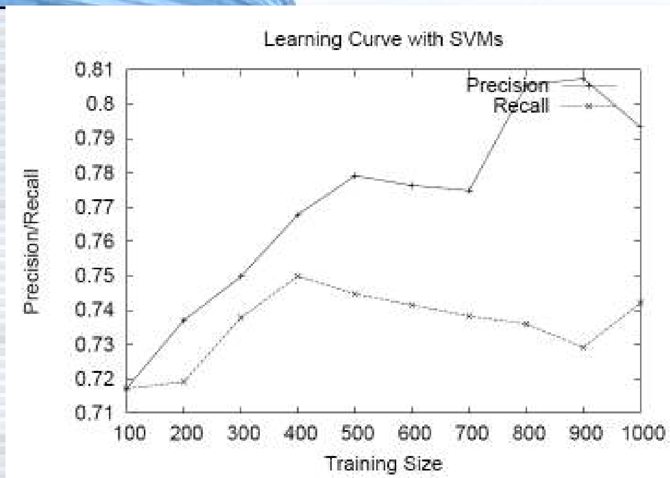
$ 197



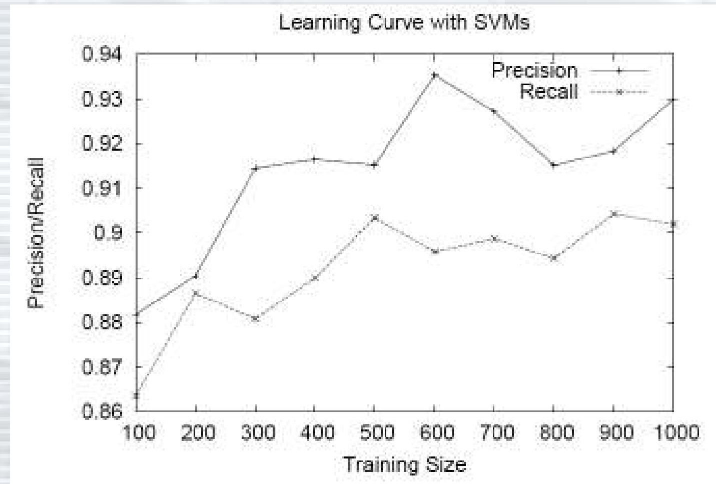
"Easily Dominate Any Market, Any Search Engine, Any Keyword."

Capture HTML Stylistic Patterns in Authentic Blogs

HTMLTAGS



Authentic
dt, marquee, table, pre
wbr, embed, img, s
noembed, warning, ahem, link
background, no, del, blockquote
basefont, description, i, ins
Spam
entrytitle, script, tt, bgsound
dl, nyt, byline, li
tr, td, nobr, content
hr, mainorarchivepage, state, meta
tblog, font, activated, status



	P	R	F1
SVM	0.94	0.91	0.92
NB	0.94	0.85	0.90

Authentic
blockquote, sup, html, mainorarchivepage
dt, del, span, img
tag, th, option, select
noscript, em, strike, ol
big, o, noembed, embed
Spam
link, h, acronym, d
marquee, thead, tfoot, fieldset
dl, b, doctype, street
center, abbr, title, a
head, meta, description, nobr

2005 2006

HTMLTAGS

- Use HTML Tags – stylistic information
- Capture signatures of splog software
- Fully language independent
- **Novel feature space**

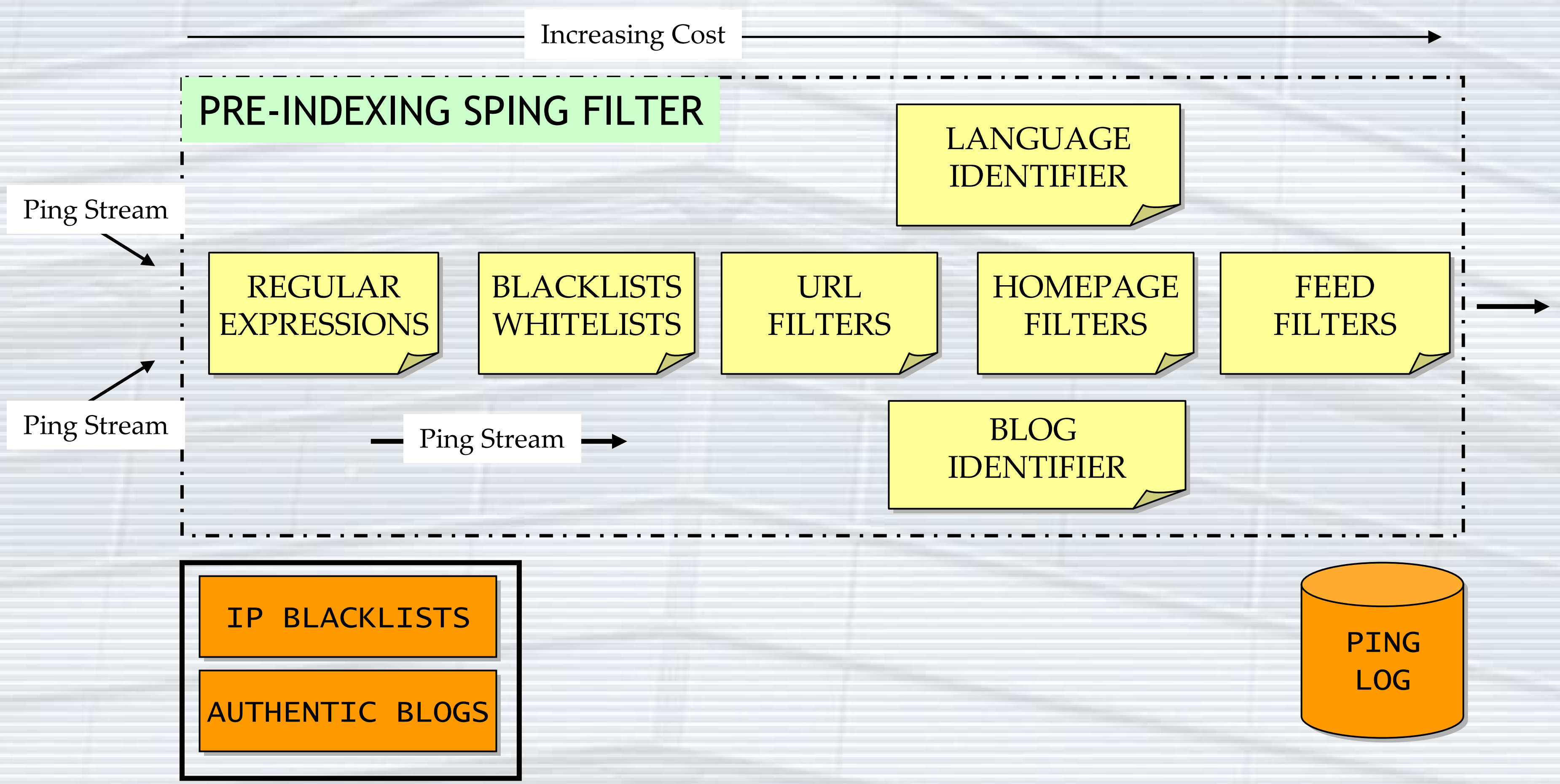
Authentic
dt, marquee, table, pre
wbr, embed, img, s
noembed, warning, ahem, link
background, no, del, blockquote
basefont, description, i, ins
Spam
entrytitle, script, tt, bgsound
dl, nyt, byline, li
tr, td, nobr, content
hr, mainorarchivepage, state, meta
tblog, font, activated, status

	P	R	F1
SVM	0.94	0.91	0.92
NB	0.94	0.85	0.90

Authentic
blockquote, sup, html, mainorarchivepage
dt, del, span, img
tag, th, option, select
noscript, em, strike, ol
big, o, noembed, embed
Spam
link, h, acronym, d
marquee, thead, tfoot, fieldset
dl, b, doctype, street
center, abbr, title, a
head, meta, description, nobr

2005 2006

META-PING SYSTEM



THE GAME THEORETIC WEB

Qouth Peter Norvig

- "The other thing that I hadn't really thought about when we started this all is how kind of game theoretic the whole thing is. At first we thought of ourselves as this observer of the Web. That the Web was out there and we made a copy of it and indexed it and if people wanted they could come and access that index. But it was just a reflection of the Web out there. And now we understand that we're co-evolving with the Web and that when we make a move it changes the Web and when the Web changes we change and going back and forth. And so all the search engine optimizers and so on are watching and what we do and we watch what they do and the Web is the interaction between us. And that is something I hadn't even considered before we saw it happening."
- *In Singularity 2007*

This is true of Social Media as well

- If I know that you are out there, trying to infer my opinions (or prevent me from spamming) then I will actively work to defeat that. Since the content is user generated, I can do that fairly quickly.
- Spam adaptation is a classic example.

ADAPTIVE CONTEXT

- Change in distribution in feature space
- Concept Drift – Seasonal, seen in both splogs and blogs

$f_1, f_2, f_3 .. f_m$

- Adversarial Scenario – seen in splogs

$P(splog(x)/O(x))$

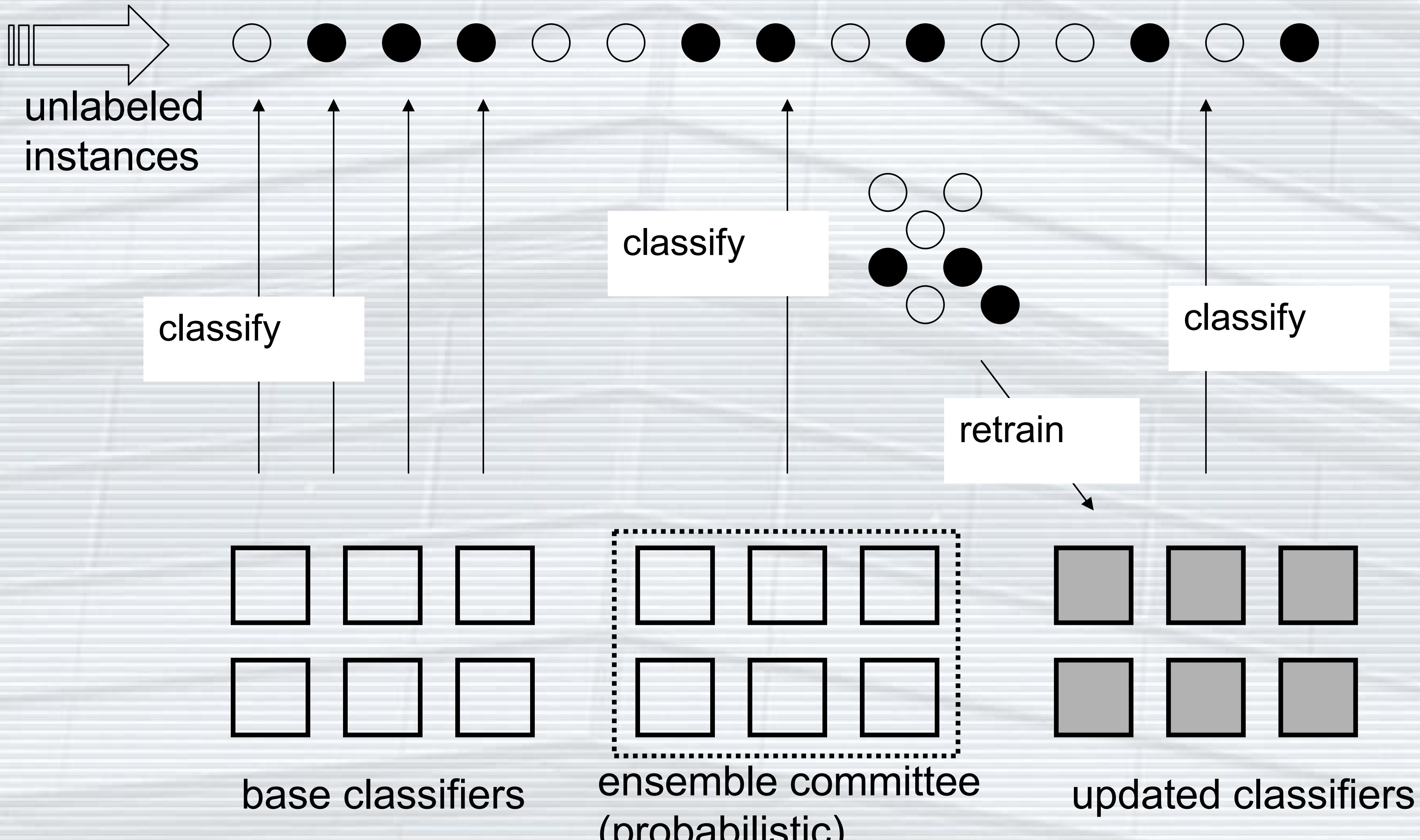
- Concept Description needs to be updated

$P(O(x)/splog(x))$

ENSEMBLE INTUITION

- Stream of unlabeled instances (drifting)
- Base classifiers with potentially independent feature spaces
- Is an ensemble (probabilistic committee) of the catalogue more robust to drift?
- Are instances classified by the ensemble effective to retrain base classifiers (semi-supervised learning)?
- Motivated by co-training

ENSEMBLE INTUITION



POTENTIAL TO ADAPT

Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.71	0.75	[illegible]
SPLOG-2006,SPLOG-2006	0.69	0.85	0.87
SPLOG-2005,SPLOG-2006	0.61	0.72	0.66

URL

Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.80	0.82	[illegible]
SPLOG-2006,SPLOG-2006	0.94	0.96	[illegible]
SPLOG-2005,SPLOG-2006	0.86	0.88	0.87

Outlink

Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.82	0.85	0.83
SPLOG-2006,SPLOG-2006	0.92	0.94	0.93
SPLOG-2005,SPLOG-2006	0.83	0.81	0.82

Anchor

Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.77	0.76	0.77
SPLOG-2006,SPLOG-2006	0.93	0.90	0.92
SPLOG-2005,SPLOG-2006	0.77	0.80	0.78

Tag

Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.88	0.86	0.87
SPLOG-2006,SPLOG-2006	0.93	0.94	0.94
SPLOG-2005,SPLOG-2006	0.84	0.71	0.77

Chargram

Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.86	0.85	0.86
SPLOG-2006,SPLOG-2006	0.89	0.96	[illegible]
SPLOG-2005,SPLOG-2006	0.87	0.85	0.86

Wordgrams

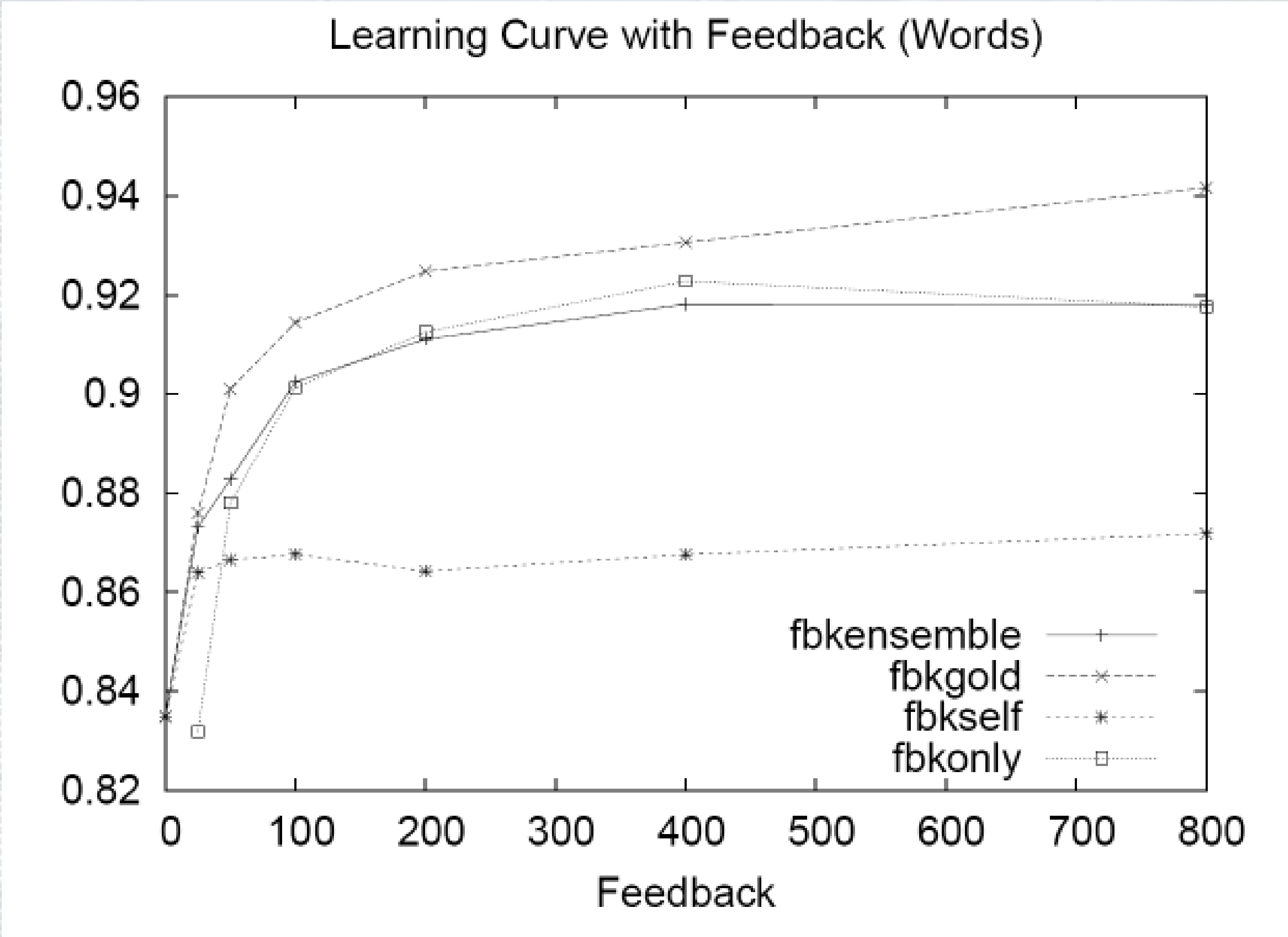
Train,Test	P	R	F1
SPLOG-2005,SPLOG-2005	0.89	0.87	0.88
SPLOG-2006,SPLOG-2006	0.96	0.96	0.96
SPLOG-2005,SPLOG-2006	0.88	0.83	0.85

Words

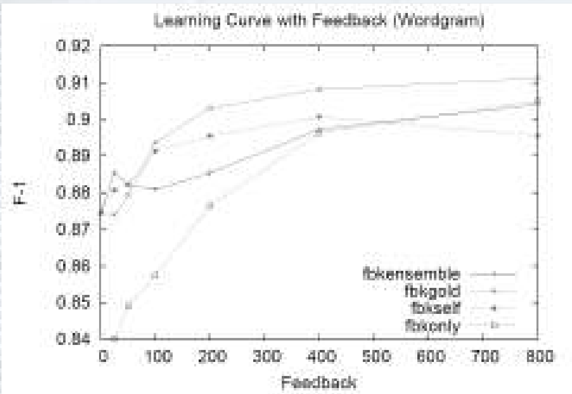
EXPERIMENTAL SETUP

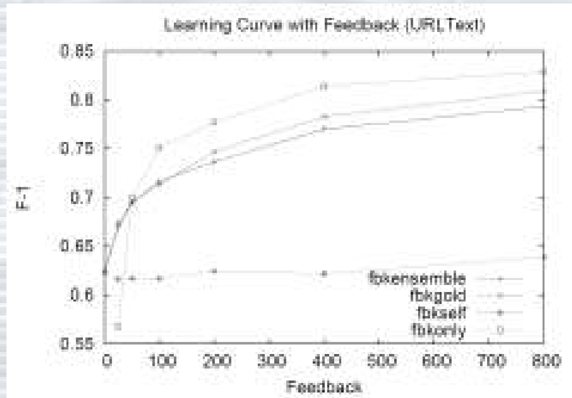
- A catalog of seven classifiers
- SPLOG-2005 as base labeled dataset
- SPLOG-2006 as evaluation stream
- 10K Top Features
- SVM based learning
- SPLOG-2006 separated out into unlabeled stream and test set (3-fold)
- F-1 performance metric evaluation

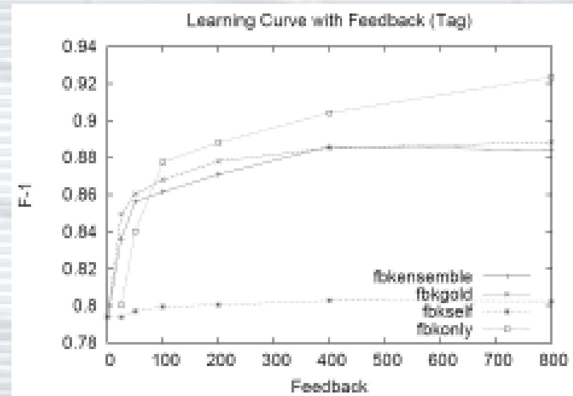
RESULTS – WORD DRIFT

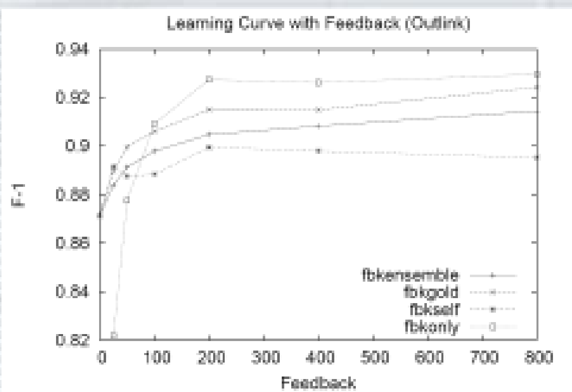


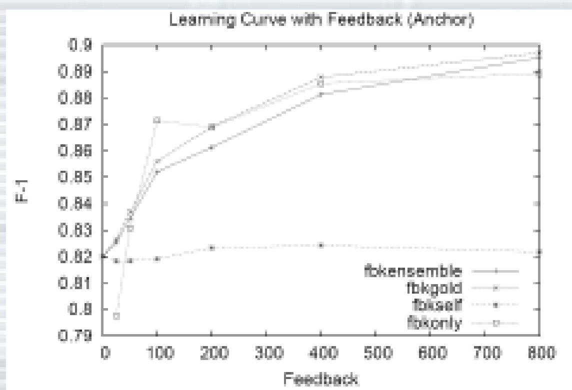
RESULTS – ALL CLASSIFIERS

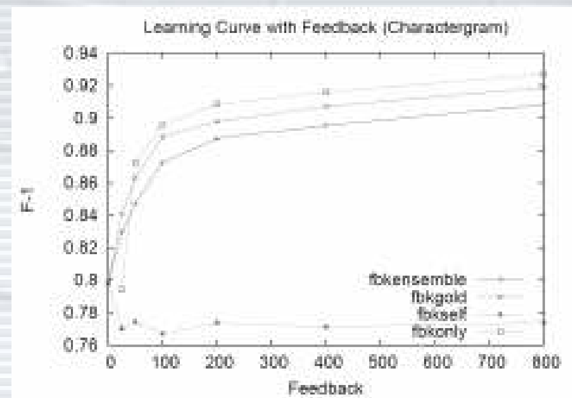












Conclusion

- Using *topic, social structure, opinions and temporal information* we can develop an accurate model for influence, bias and trust on the Blogosphere.
- We apply this framework on real-world data and describe techniques for identifying influence on the Blogosphere.
- Splogs are a big issue – we have developed efficient techniques to detect them in near real time.
- Does the Game Theoretic Nature of this system raise fundamental new challenges for Data Mining.

Backup Slides

Generative Models for Blogosphere

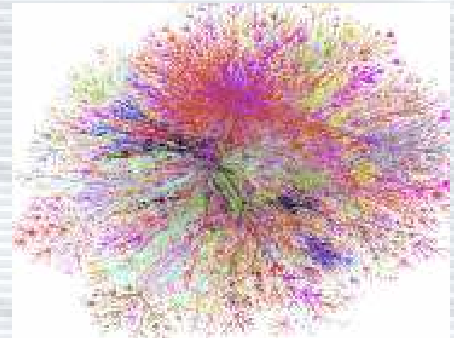
Graphs are everywhere .. and so are Power laws!!

In simple words, power law can be explained by "**rich get richer phenomenon**" OR "**20% of the population holds 80% of the wealth**"

Considering web as a graph:

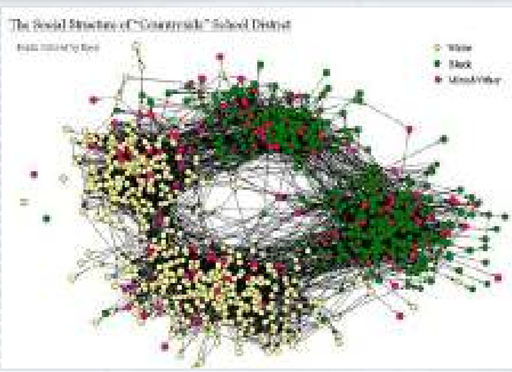
$$P(k) = k^{-\gamma}$$

k is degree of the node



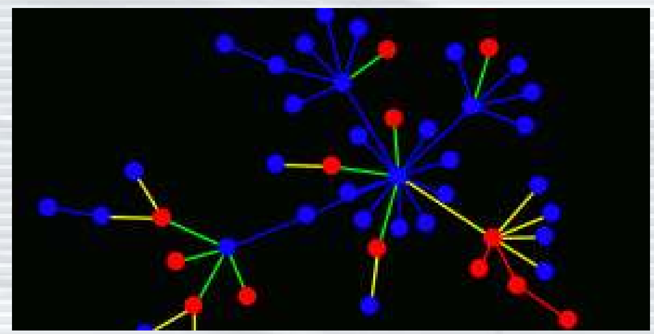
Internet Mapping Project [lumeta.com]

Friendship Network [Moody '01]



Scale-free network: Structure and properties independent of network size

Few high connectivity node (hubs)



http://www.prefuse.org/gallery/

Properties of interest (graph theory)

Average degree of node, degree distribution, degree correlation, distribution of strongly/weakly connected components, clustering coefficient and reciprocity

Generative Models for Blogosphere

- **Reduce time to generate data**
  - crawling the blogosphere over a few weeks
  - sampling the right blogs to get a representative sample
- **Reduce time in preprocessing and data cleaning**
  - removing links pointing outside the dataset, outside the time frame
  - splog removal [1]
- **Generate graphs of different properties\sizes**
  - average degree of node, degree distributions
- **Testing of new algorithms for blog graphs**
  - e.g. spread of influence in blogosphere [2], community detection [3]
- **Extrapolation**
  - how will fast growth affect the blogosphere properties?
  - how does this affect the connected components?

[1] Kolari et al "Svms for the blogosphere: Blog identification and splog detection," in AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006.
[2] Java et al "Modeling the spread of influence on the blogosphere," tech. rep., University of Maryland, Baltimore County, March 2006.
[3] Lin et al "Discovery of Blog Communities based on Mutual Awareness

Existing Approaches

Erdos-Renyi random model

Barabasi Albert preferential attachment

Property	ER model	BA model	Simulation	Blogopshere
Type	undirected	undirected	directed	directed
Degree distribution	Poisson refer [1]	Power Law refer [3]	Power Law	Power Law refer [7, 32]
Slope [inlinks,outlinks]	-	[2.08,-]	[1.7-2.1,1.5-1.6]	[1.66-1.8,1.6-1.75]
Avg. degree	constant (for given p)	constant (adds m edges)	increases	increases
Component distribution	-	-	Power Law	Power Law [7]
Correlation coeffi cient	-	1 (high - fully preferential)	0.1 (low)	0.024 (low-WWE)
Avg clustering coeff.	0.00017 (low)	0.00018	0.0242 (high)	0.0235 (WWE)
Reciprocity	N/A (undirected)	N/A (undirected)	0.6	0.6 (WWE)

Preferential Attachment: The likelihood of linking to a popular website is higher

- **Two level network: blog and post level**
- **Inlinks and outlinks to and from posts**
- **NEED to model blogger interactions**

[1] M. Newman, "The structure and function of complex networks," 2003

[3] R. Albert, *Statistical mechanics of complex networks*. PhD thesis, 2001.

[7] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs", *ICWSM*, 2007

Model Parameters

1. Probability of random reads (**rR**)
2. Probability of randomly selecting writer (**rW**)
3. Probability that new node does not link to the existing network (**pD**)
4. Growth exponent (**g**)
   – how many links should be added every step?

Proposed Model

1. Add new blog node
2. Select writer
3. Writers read blog posts, write posts

I will not link to anyone!

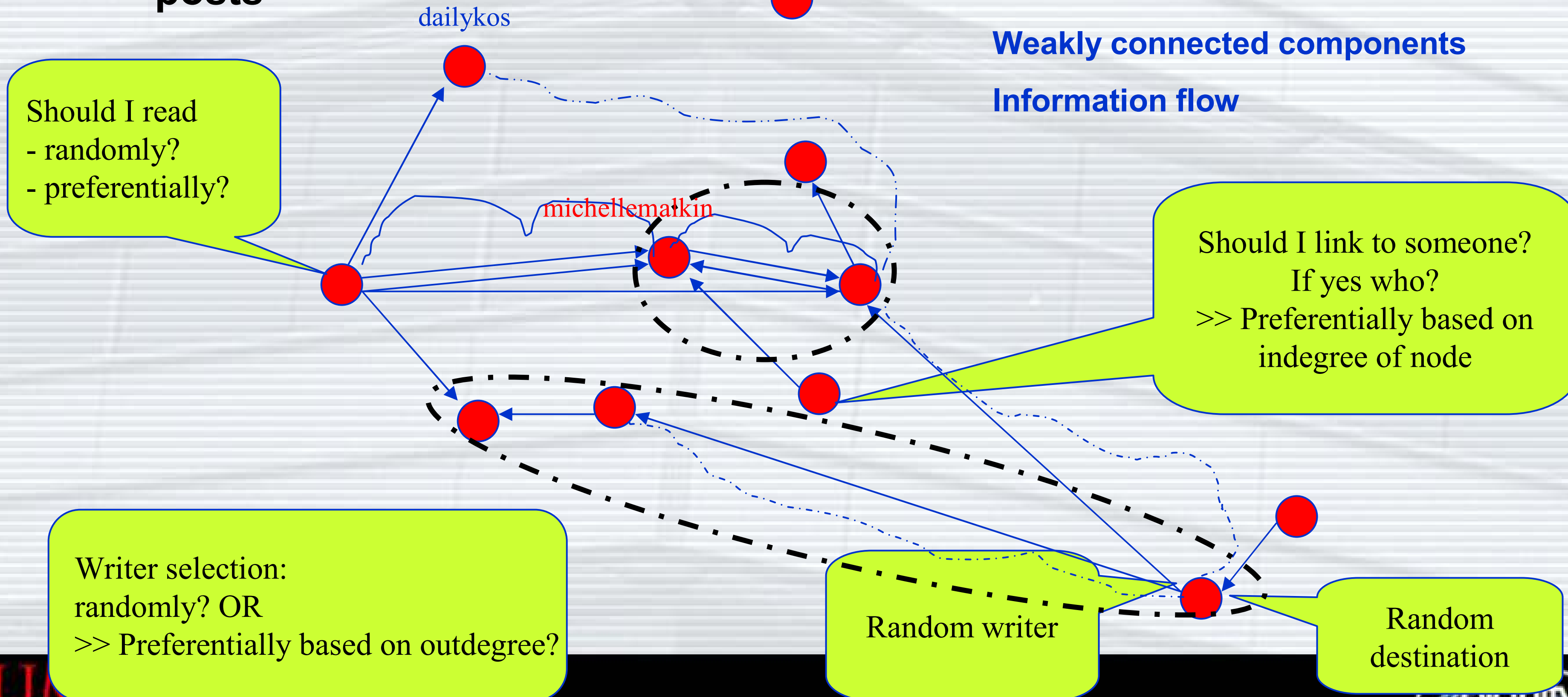
Step=1
Step=2

Reciprocal links

Strongly connected components
Subset of nodes having directed path from every node to every other node

Weakly connected components

Information flow



Properties of Simulated Blog Graphs

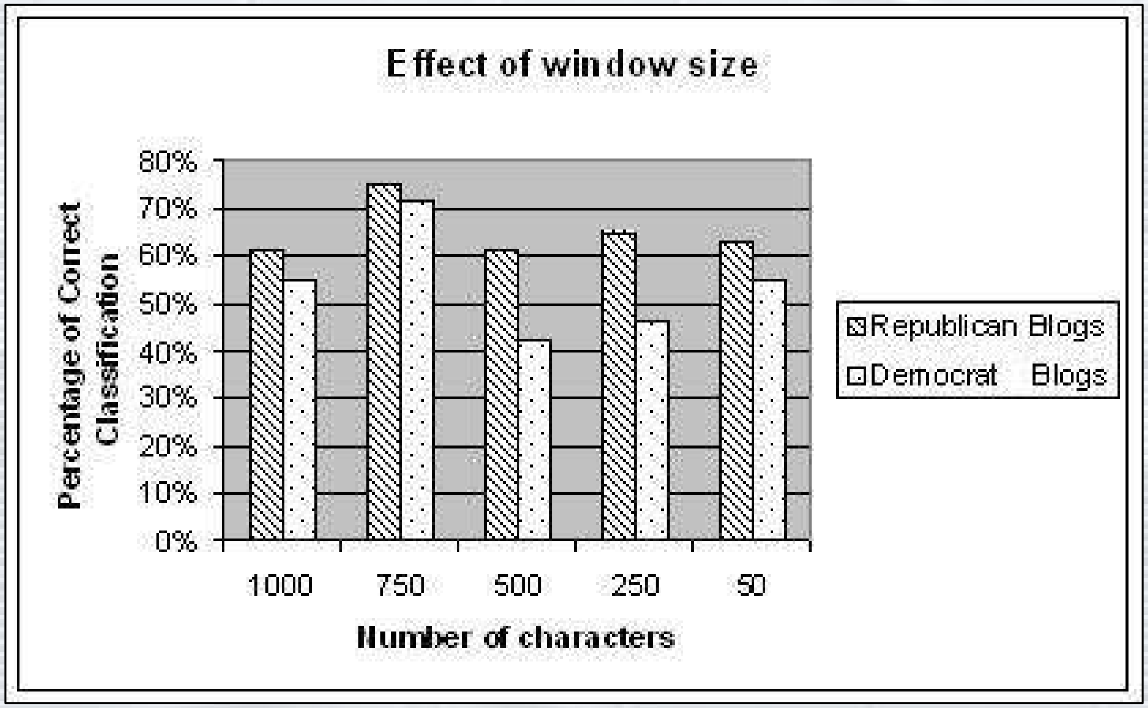
Table 5.4. Comparison of blog network properties of datasets and simulation

Blog network properties	ICWSM 2007	WWE 2006	Simulation
Total blogs	159,036	650,660	650,000
Total blog-blog links	435,675	1,893,187	1,451,069
Unique blog-blog links	245,840	648,566	1,158,803
Average degree	5.47	5.73	4.47
Indegree distribution	-2.07	-2.0	-1.71
Outdegree distribution	-1.51	-1.6	-1.76
Degree correlation coefficient	0.056	0.002	0.10
Diameter	14	12	6
Largest WCC size	96,806	263,515	617,044
Largest SCC size	4,787	4,614	72,303
Clustering coefficients	0.04429	0.0235	0.0242
Percent Reciprocity	3.03	0.6838	0.6902

Rank	MSM sources for Democrats MSM	Links from Dems	Links from Reps	Polarity Dem	Polarity Rep
1	http://mediamatters.org	76 from 28 blogs	5 from 4 blogs	4.368336871	-5.9562827
2	http://www.rawstory.com	108 from 38 blog	14 from 11 blogs	2.873328203	6.013103206
3	http://www.nytimes.com	503 from 83 blogs	199 from 50 blogs	-2.31435096	-3.35244371
4	http://www.alternet.org	38 from 19 blogs	2 from 2 blogs	?	?
5	http://www.washingtonpost.com	750 from 91 blogs	355 from 61 blogs	-1.647123666	5.449887525
6	http://news.independent.co.uk	59 from 20 blogs	5 from 5 blogs	?	?
7	http://www.salon.com	48 from 25 blogs	8 from 2 blogs	2.163055083	-1.88452348
8	http://www.truthout.org	85 from 35 blogs	24 from 10 blogs	-1.484073313	1.772874119
9	http://www.usatoday.com	168 from 55 blogs	71 from 36 blogs	-8.239055964	4.202658984
10	http://www.thenation.com	29 from 17 blogs	4 from 3 blogs	-1.663142934	1.106710739

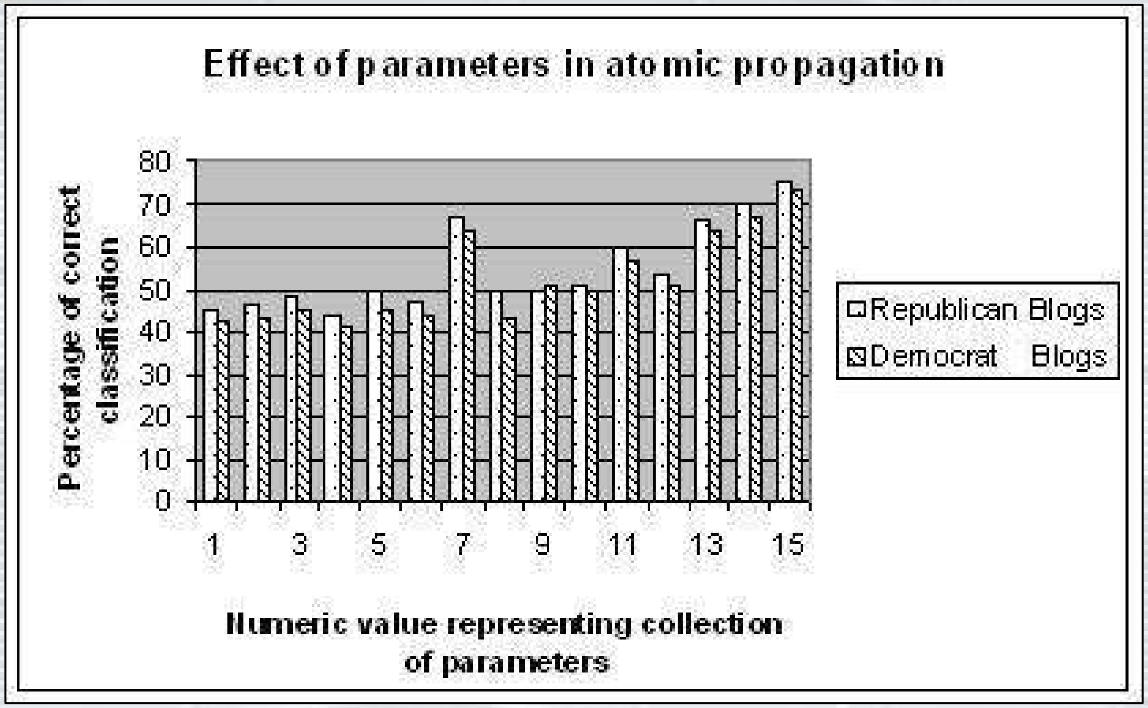
Rank	MSM sources for Republicans MSM	Links from Dems	Links from Reps	Polarity Dem	Polarity Rep
1	http://www.washingtontimes.com	17 from 11 blogs	65 from 33 blogs	?	?
2	http://www.foxnews.com	64 from 23 blogs	165 from 44 blogs	-8.197277972	4.502696152
3	http://apnews.myway.com	4 from 3 blogs	33 from 17 blogs	-1.477490333	9.633693436
4	http://www.examiner.com	4 from 4 blogs	23 from 17 blogs	?	?
5	http://www.frontpagemag.com	3 from 3 blogs	23 from 13 blogs	?	?
6	http://www.humaneventsonline.com	6 from 5 blogs	22 from 16 blogs	-4.314417358	1.140630351
7	http://www.townhall.com	31 from 8 blogs	72 from 24 blogs	-4.980464907	3.116320103
8	http://www.dailybulletin.com	5 from 3 blogs	19 from 14 blogs	5.272860746	2.064693675
9	http://www.sacbee.com	0 from 0 blogs	6 from 6 blogs	?	?
10	http://www.spectator.org	5 from 3 blogs	17 from 11 blogs	-7.205228528	2.09956978

Effect of text window size



- Optimal window size is 750 characters for our experiments
- Small window size – Non-opinionated phrases
- Large Window size – Analysis of non-related text
- Specific to our experiments, numbers may not be generalized

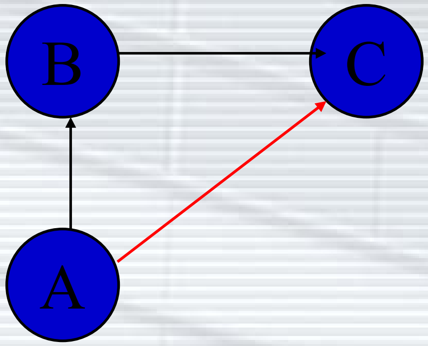
Effect of atomic propagation parameters



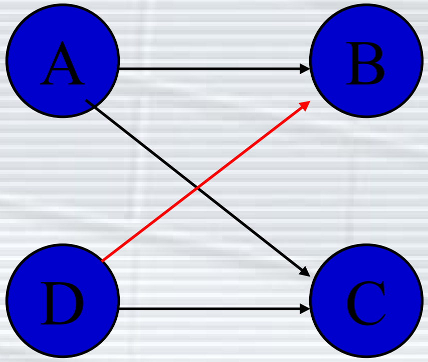
- X-axis Bitset = {direct trust, co–citation, transpose trust and trust coupling} = {0001 - 1111}
- Each parameter set to either 0 or its optimal value
- Collective influence of all parameters helps !

Atomic Propagation

- Direct Propagation
  - Given: A trusts B and B trusts C
  - Implies: A trusts C
  - Operator : M

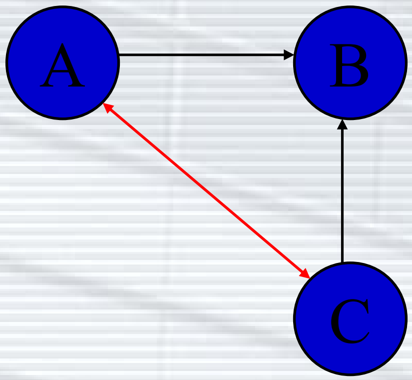


- Co-citation
  - Given: A trusts B and C, D trust C
  - Implies: D trusts B
  - Operator : $M^T * M$

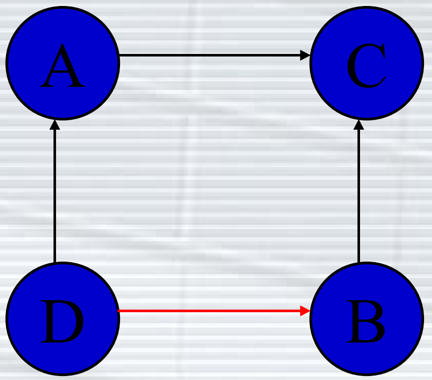


Atomic Propagation Contd…

- Transpose Trust
  - Given: A trusts B and C trusts B
  - Implies: C trusts A, A trusts C
  - Operator : M^T



- Trust Coupling
  - Given: D trusts A, A trusts C
    and B trusts C
  - Implies: D trusts B
  - Operator : $M * M^T$



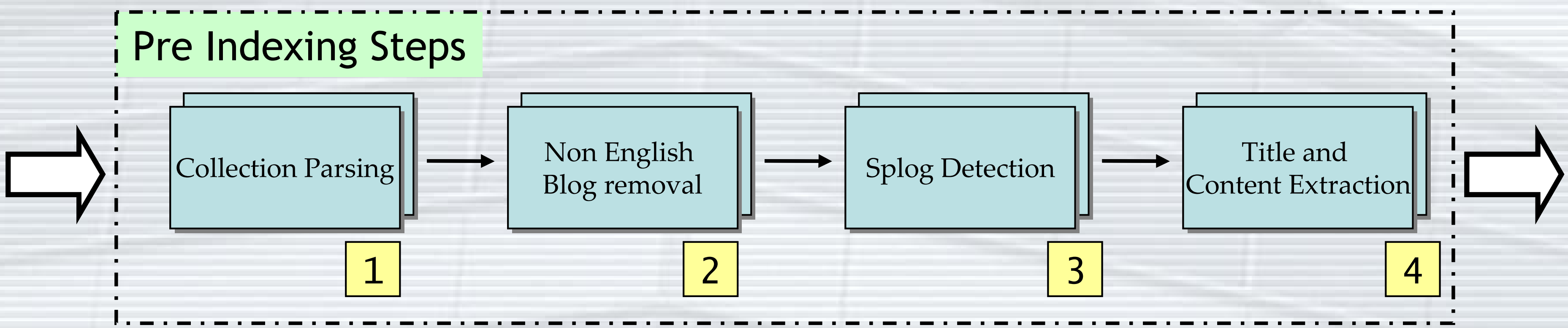
Atomic Propagation contd…

- Combined Operator
  - $Ci = a_1 * M + a_2 * M^T*M + a_3 * M^T + a_4 * M*M^T$
  - a_i {0.4, 0.4, 0.1, 0.1} represents weighing factor
- Belief Matrix after i^{th} atomic propagation
  - $M_{i+1} = M_i * C_i$
- We perform propagations till "convergence" (till the new iteration does not change values in M above "threshold")

Separating Blog Wheat from Blog Chaff

Data cleaning for

- Splog removal
- Post content identification



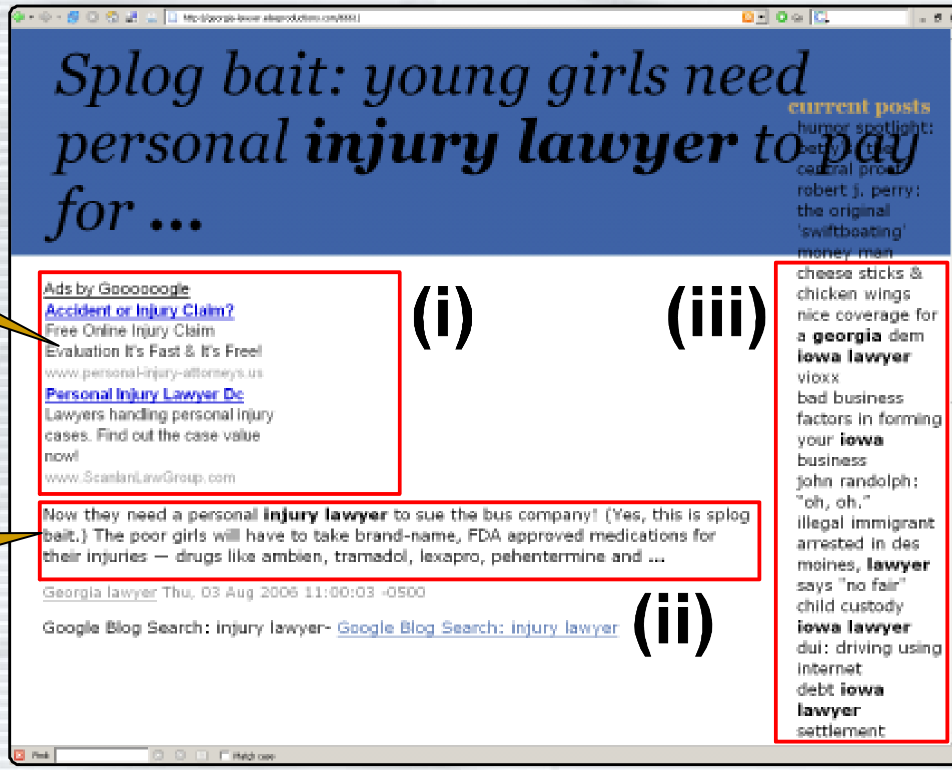
BlogVox: Separating Blog Wheat from Blog Chaff", IJCAI 2007 Analytics of Noisy and Unstructured Text

Data Cleaning: Splogs

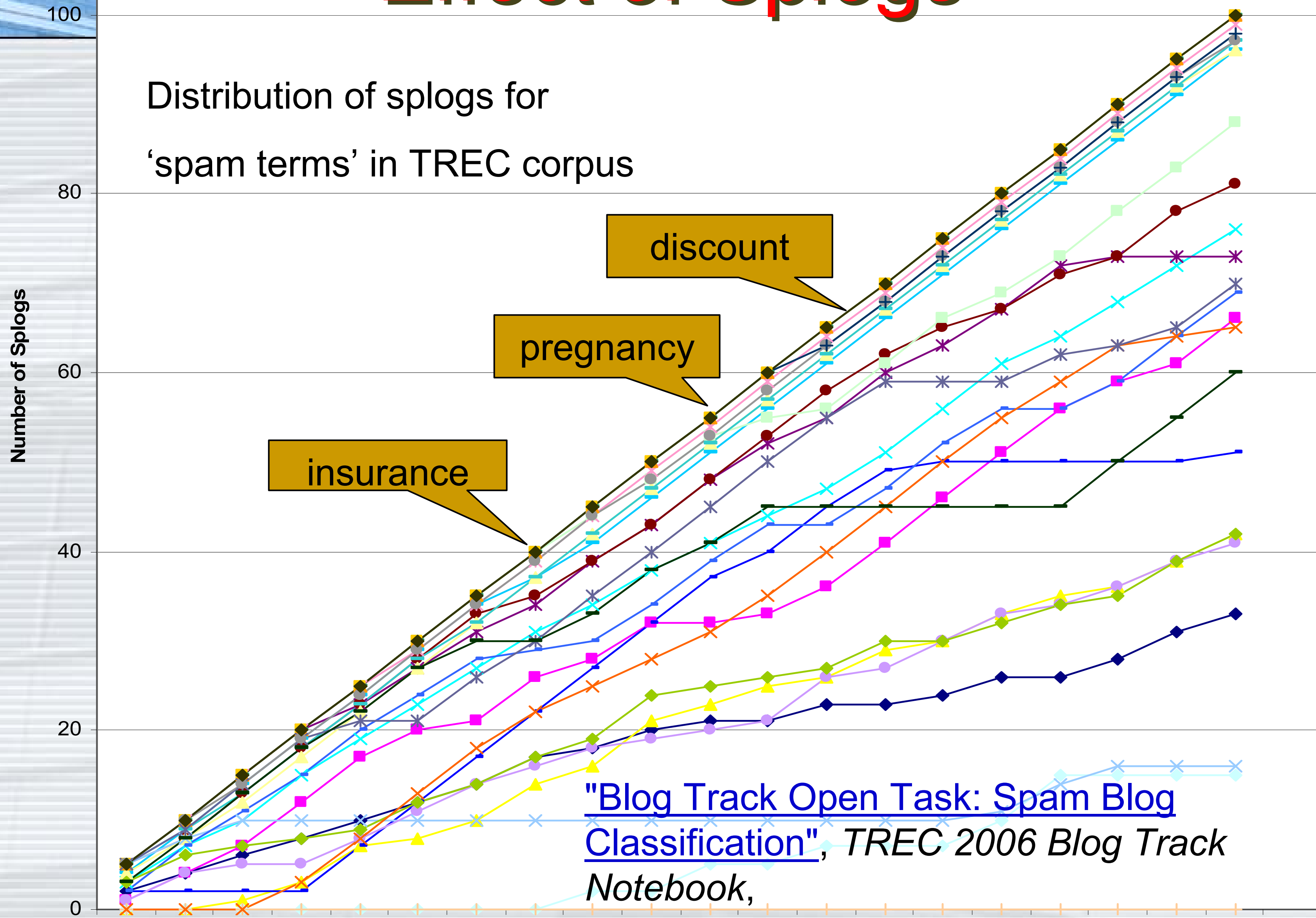
Host Ads

Plagiarized content

Index affiliates, Promote pageRank



Effect of Splogs



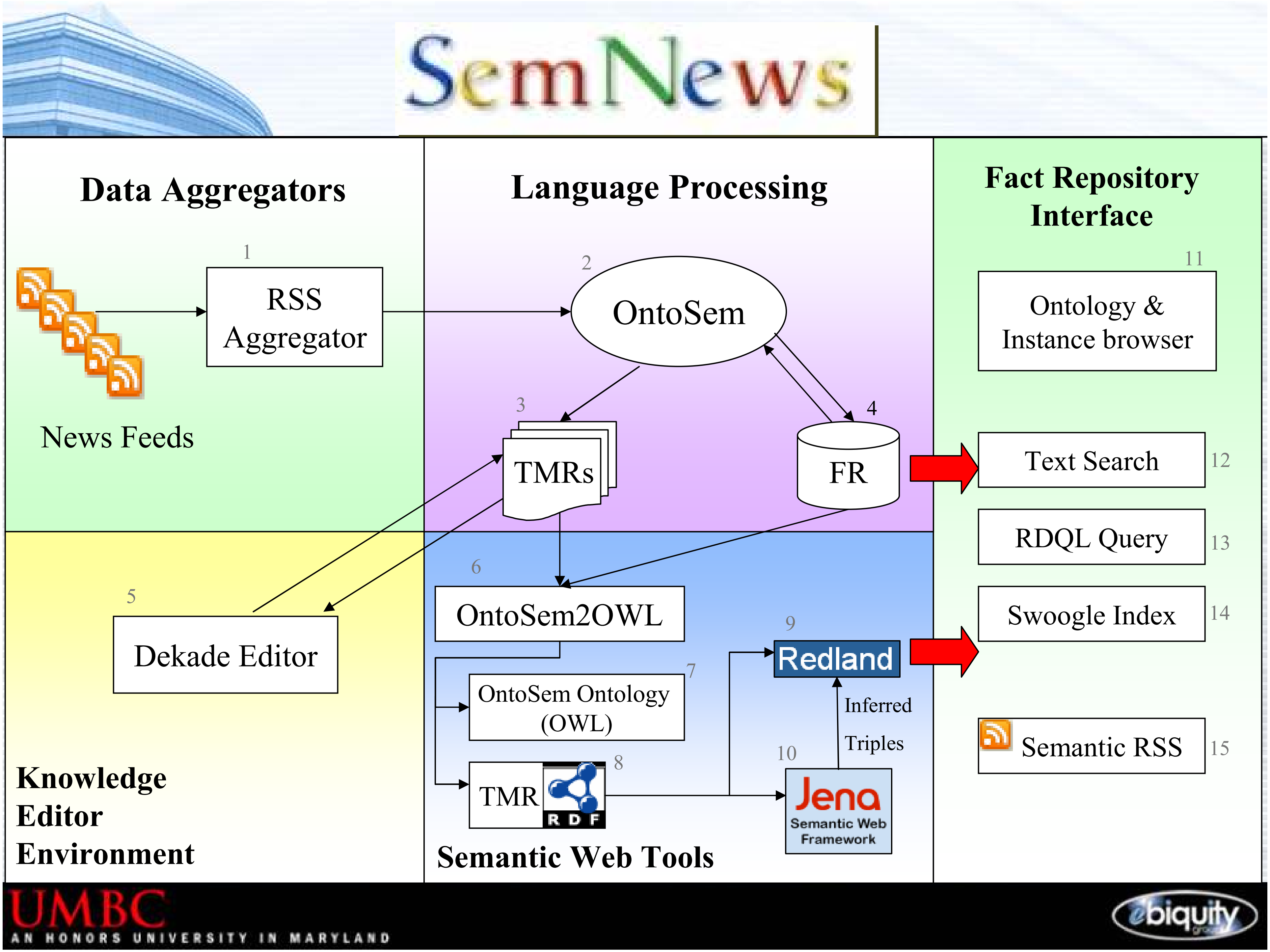
"Blog Track Open Task: Spam Blog Classification", *TREC 2006 Blog Track Notebook*,

Data Cleaning: Content Identification

- Baseline Heuristic
- SVM Method

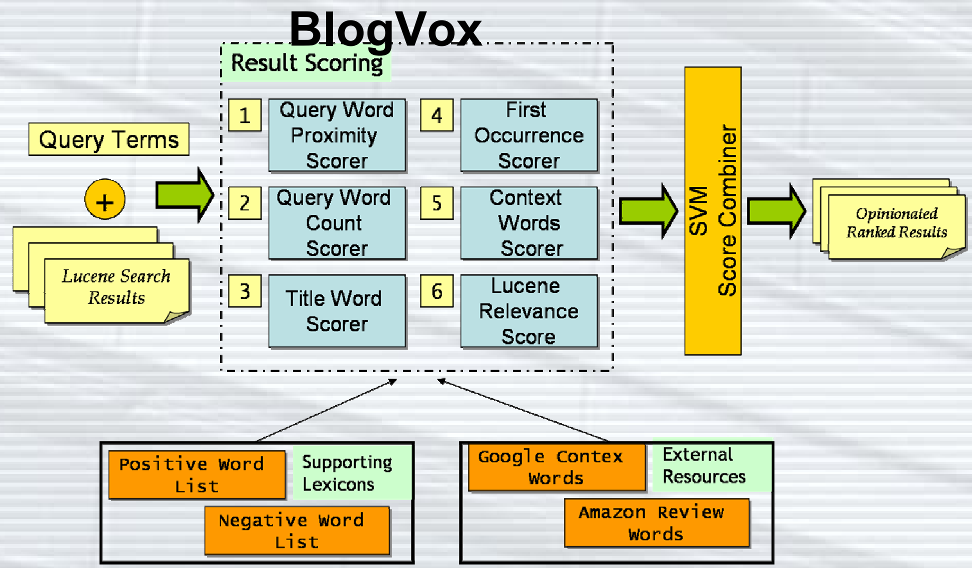
ID	Features
1	Previous Node
2	Next Node
3	Parent Node
4	Previous N Tags
5	Next N Tags
6	Sibling Nodes
7	Child Nodes
8	Depth in DOM Tree
9	Char offset from page start
10	links outside the blog?
11	Anchor text words
12	Previous N words
13	Next N words

Method	*Precision*	*Recall*	*F1*
baseline heuristic	0.83	0.87	0.849
svm cleaner (tag features)	0.79	0.78	0.784
svm cleaner (all features)	0.86	0.94	0.898


SemNews
Data Aggregators
1
RSS Aggregator
News Feeds
Language Processing
2
OntoSem
3
TMRs
4
FR
Fact Repository Interface
11
Ontology & Instance browser
Text Search
12
RDQL Query
13
Swoogle Index
14
Semantic RSS
15
5
Dekade Editor
Knowledge Editor Environment
6
OntoSem2OWL
7
OntoSem Ontology (OWL)
8
TMR
RDF
9
Redland
Inferred Triples
10
Jena
Semantic Web Framework
Semantic Web Tools
UMBC
AN HONORS UNIVERSITY IN MARYLAND
ebiquity

BlogVox Opinion Extraction System

- **TREC 06**: Finding *opinionated* posts, either positive or negative, about a query
- 2006 TREC Blog corpus:
  - 80K blogs
  - 300K posts
  - 50 test queries
- **BlogVox** opinion extraction system
  - Document and sentence level scorers
  - Combined scores using an SVM meta-learner
  - Data cleaning: splogs and post identification



Run	Opinion		Topic Relevance	
	MAP	R-prec	MAP	R-prec
Unclean Index	0.1275	0.202	0.1928	0.2858
Cleaned Index	0.1548	0.2388	0.2268	0.3272

Brand Monitoring / Business Analytics

Blog Analytics/ Market Intelligence

cymfony harnessing influence 2.0™
umbria
BIZ360
Nielsen BuzzMetrics
analytica
MONITOR110
opinmind discovering bloggers
MOTIVEQUEST
collective intellect
factiva Dow Jones & Reuters
brandimensions
BUZZLOGIC

Buzz
Opinions
Influence
Reputation
Competition
Financial Analyst

Limitations

- Proprietary
- Some companies conduct extensive manual research

Top Cited Media Sources

Top MSM Sources on the Blogosphere

Rank	MSM
1	http://www.nytimes.com
2	http://www.washingtonpost.com
3	http://news.yahoo.com
4	http://news.bbc.co.uk
5	http://www.msnbc.msn.com
6	http://www.cnn.com
7	http://news.google.com
8	http://www.bbc.co.uk
9	http://www.usatoday.com
10	http://sports.espn.go.com

Top MSM from Democrats	Top MSM from Republicans
http://www.washingtonpost.co	http://www.washingtonpost.com
http://www.nytimes.com	http://news.yahoo.com
http://news.yahoo.com	http://www.nytimes.com
http://www.msnbc.msn.com	http://www.foxnews.com
http://www.cnn.com	http://www.cnn.com
http://www.usatoday.com	http://www.msnbc.msn.com
http://www.abcnews.go.com	http://www.usatoday.com
http://www.latimes.com	http://www.washingtontimes.com
http://www.boston.com	http://www.abcnews.go.com
http://www.rawstory.com	http://www.timesonline.co.uk
http://www.truthout.org	http://today.reuters.com
http://news.bbc.co.uk	http://www.sfgate.com
http://www.cbsnews.com	http://news.bbc.co.uk
http://today.reuters.com	http://www.townhall.com
http://mediamatters.org	http://www.canada.com

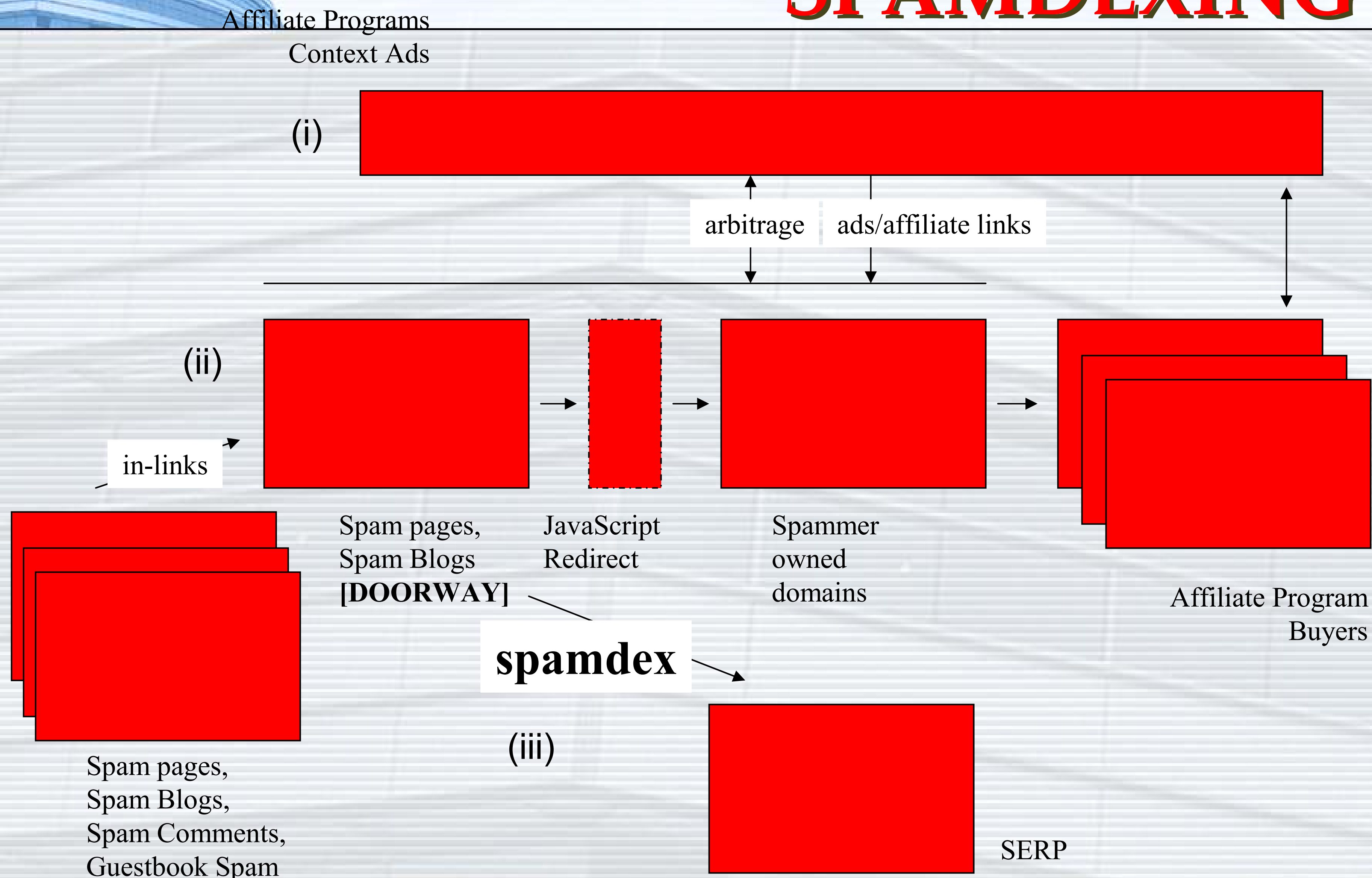
Propagating Influence

- Trust-only
  - Ignore distrust (negative polarities) completely
  - Final Belief Matrix = M_k , $M_0 = T$
    - (K : Number of atomic propagations till convergence)

- One-step Distrust
  - Distrust propagates single step while trust propagates repeatedly
  - Final Belief Matrix = $M_k * (T-D)$, $M_0 = T$
    - (K : Number of atomic propagations till convergence)

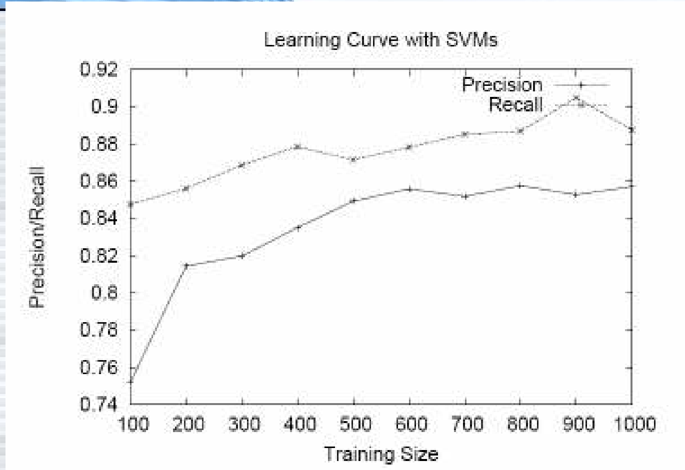
- Propagated Distrust
  - Treat distrust and trust equivalent
  - Final Belief Matrix = M_k , $M_0 = T - D$
    - (K : Number of atomic propagations till convergence)

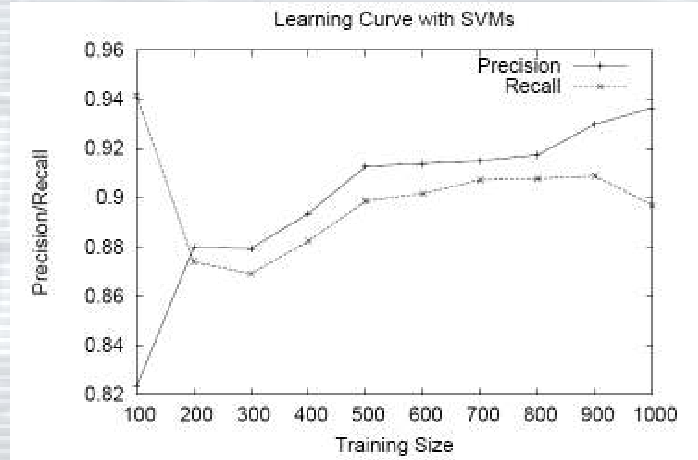
SPAMDEXING



Affiliate Programs
Context Ads
(i)
arbitrage
ads/affiliate links
(ii)
in-links
Spam pages,
Spam Blogs
[DOORWAY]
JavaScript
Redirect
Spammer
owned
domains
Affiliate Program
Buyers
spamdex
(iii)
SERP
Spam pages,
Spam Blogs,
Spam Comments,
Guestbook Spam

WORDGRAMS





	P	R	F1
SVM	0.86	0.86	0.86
NB	0.80	0.83	0.82

Authentic
nbsp-nbsp, personal-web, s-personal
please-read, read-my, are-my
a-new, nbsp-blog, blog-nbsp,
about-me, here-are, search-this
september-august, i-have, s-blog
Spam
to-us, at-am, uncategorized-no
comments-off, linking-to, com-archives
site-index, self-publishing, writer-s
archives-august, the-internet, in-den
new-york, the-best, many-people

	P	R	F1
SVM	0.93	0.92	0.92
NB	0.90	0.92	0.91

Authentic
pm-nbsp, me-do, profile-links
this-post, comments-links, am-nbsp
to-this, previous-posts, nbsp-about
nbsp-friday, the-new, nbsp-thursday
links-to, post-nbsp, march-april
Spam
technorati-tag, recent-posts, comments-nbsp
tuesday-october, am-comments, friendly-blogs
tue-oct, my-favorites, mon-oct
original-post, blog-tag, sun-oct
sponsors-ads, thu-oct, ads-recent

2005 2006

WORDGRAMS

- Word-2-grams, 2 adjacent words
- Shallow NLP technique to tackle word salad
- Word salad less common in web spam (TFIDF)
- Word-x-gram features, exponential with x

	P	R	F1
SVM	0.86	0.86	0.86
NB	0.80	0.83	0.82

Authentic
nbsp-nbsp, personal-web, s-personal
please-read, read-my, are-my
a-new, nbsp-blog, blog-nbsp,
about-me, here-are, search-this
september-august, i-have, s-blog
Spam
to-us, at-am, uncategorized-no
comments-off, linking-to, com-archives
site-index, self-publishing, writer-s
archives-august, the-internet, in-den
new-york, the-best, many-people

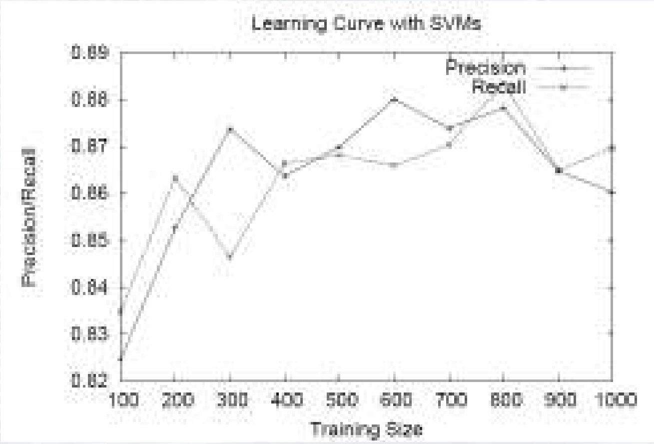
2005

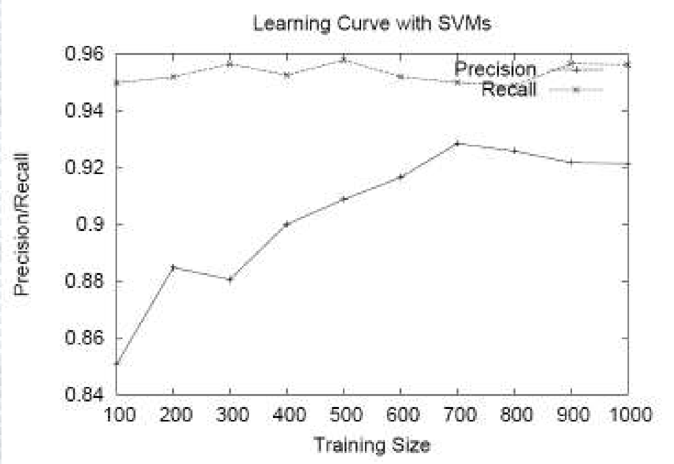
	P	R	F1
SVM	0.93	0.92	0.92
NB	0.90	0.92	0.91

Authentic
pm-nbsp, me-do, profile-links
this-post, comments-links, am-nbsp
to-this, previous-posts, nbsp-about
nbsp-friday, the-new, nbsp-thursday
links-to, post-nbsp, march-april
Spam
technorati-tag, recent-posts, comments-nbsp
tuesday-october, am-comments, friendly-blogs
tue-oct, my-favorites, mon-oct
original-post, blog-tag, sun-oct
sponsors-ads, thu-oct, ads-recent

2006

CHARACTERGRAMS





	P	R	F1
SVM	0.86	0.87	0.87
NB	0.78	0.83	0.80

Authentic
lle, blo, gal, see
ami, thin, add, pleas
plea, woul, son, lou
inu, gall, flic, gue
jan, galle, wha, erenc
Spam
new, ver, rti, bes
oste, poste, aqu, ail
prev, inf, ran, hei
icl, man, pro, fin
tra, itie, rov, che

	P	R	F1
SVM	0.93	0.93	0.93
NB	0.86	0.87	0.86

Authentic
oca, ocati, ocat, loca
apr, locat, apri, loc
catio, marc, jun, cati
bru, vem, riv, jul
feb, lic, ebr, vemb
Spam
pos, ost, post, blo
lin, new, tio, pro
rec, edi, com, ssn
essn, rat, ess, ehnor
hnor, hnora, norat, ent

2005 2006

CHARACTERGRAMS

- 3,4,5 charactergrams from blog content
- Can capture character salad (e.g. p1lls)
- Feature selection important

	P	R	F1
SVM	0.86	0.87	0.87
NB	0.78	0.83	0.80

Authentic
lle, blo, gal, see
ami, thin, add, pleas
plea, woul, son, lou
inu, gall, flic, gue
jan, galle, wha, erenc
Spam
new, ver, rti, bes
oste, poste, aqu, ail
prev, inf, ran, hei
icl, man, pro, fin
tra, itie, rov, che

2005 | 2006

	P	R	F1
SVM	0.93	0.93	0.93
NB	0.86	0.87	0.86

Authentic
oca, ocati, ocat, loca
apr, locat, apri, loc
catio, marc, jun, cati
bru, vem, riv, jul
feb, lic, ebr, vemb
Spam
pos, ost, post, blo
lin, new, tio, pro
rec, edi, com, ssn
essn, rat, ess, chnor
hnor, hnora, norat, ent