Modeling Community & Sentiment

using latent variable models

Ramnath Balasubramanyan rbalasub@cs.cmu.edu
(with William Cohen, Alek Kolcz and other collaborators)

Modeling Polarizing Topics

When Do Different Political Communities Respond Differently to the Same News?

"essentially all models are wrong, but some are useful"

Peter Norvig

MCR-LDA

Modeling Polarizing topics in Politics

Political decision making is based on an immediate emotional response *[Lodge & Taber, 2000]*

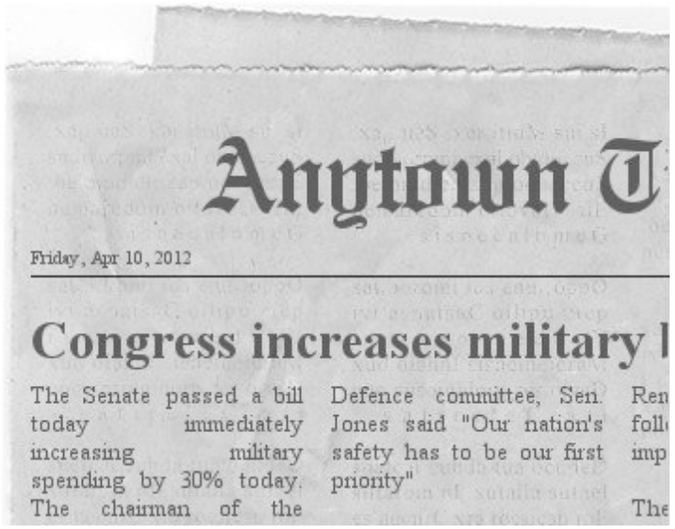
It is important to understand how different communities react to political stimuli.

My City Tr

Thursday, May 17, 2012

Assault weapons banned

The federal government announced today that assault weapons will be immediate today, sale of prohibited


Anytown T

Friday, Apr 10, 2012

Congress increases military

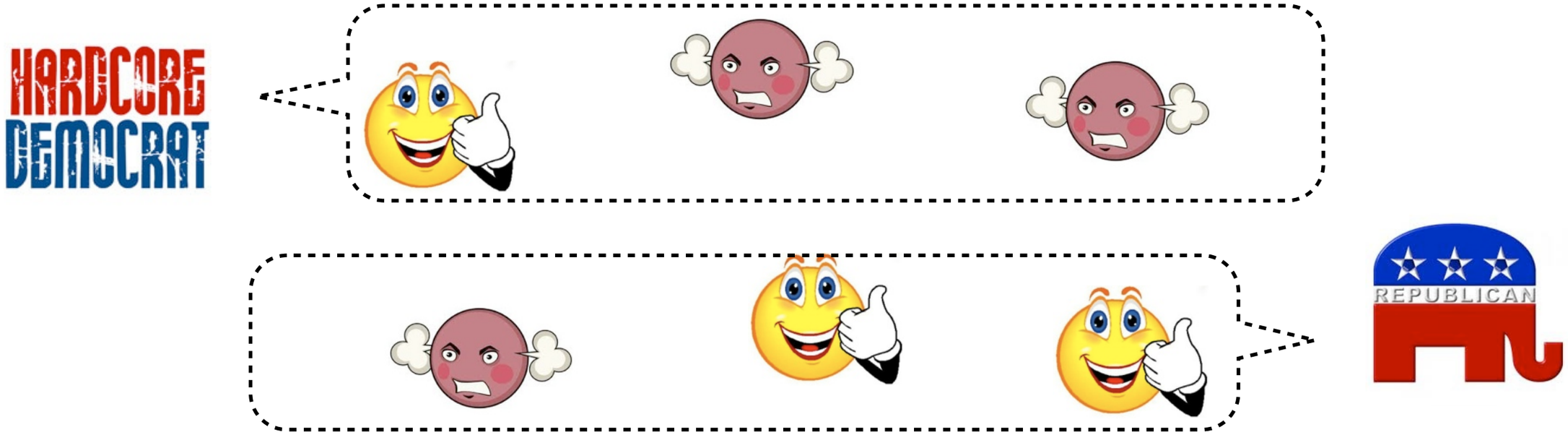
The Senate passed a bill today immediately increasing military spending by 30% today. The chairman of the Defence committee, Sen. Jones said "Our nation's safety has to be our first priority"

Washington

Friday, Apr 10, 2012

Capital gains tax rate slashe

The President signed into law a bill slashing the capital gains tax rate to 1%.

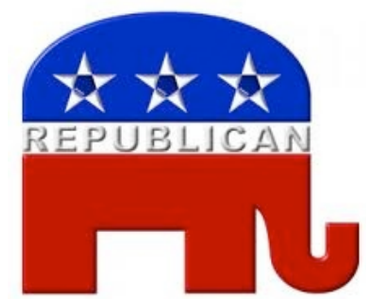


MCR-LDA

Problem statement

Predict response

reaction?



reaction?

+

THE HUFFINGTON POST

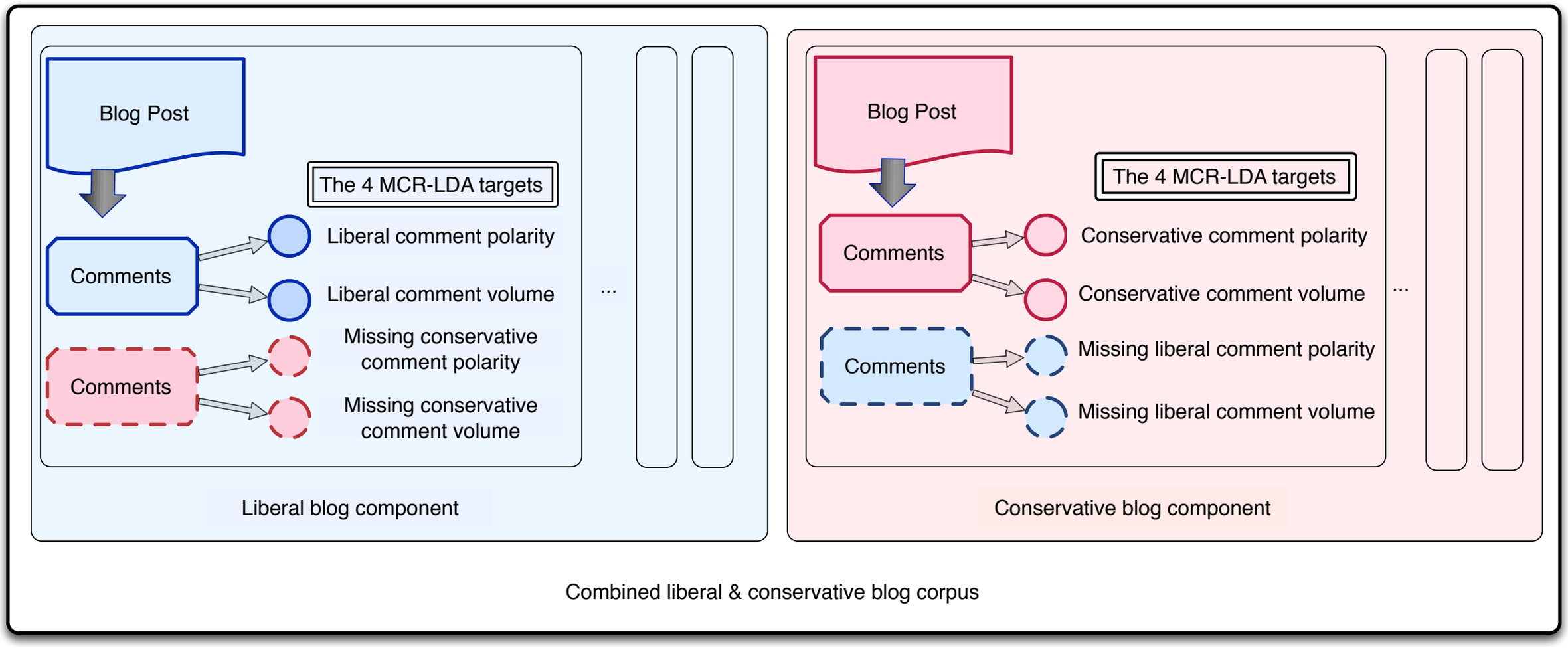
THE CARPETBAGGER REPORT Commentary and Analysis on Politics in America by Steve Benen

DAILY KOS

What issues are they talking about?

Multi Community Response LDA (MCR-LDA)

Multi target Semi-supervised LDA



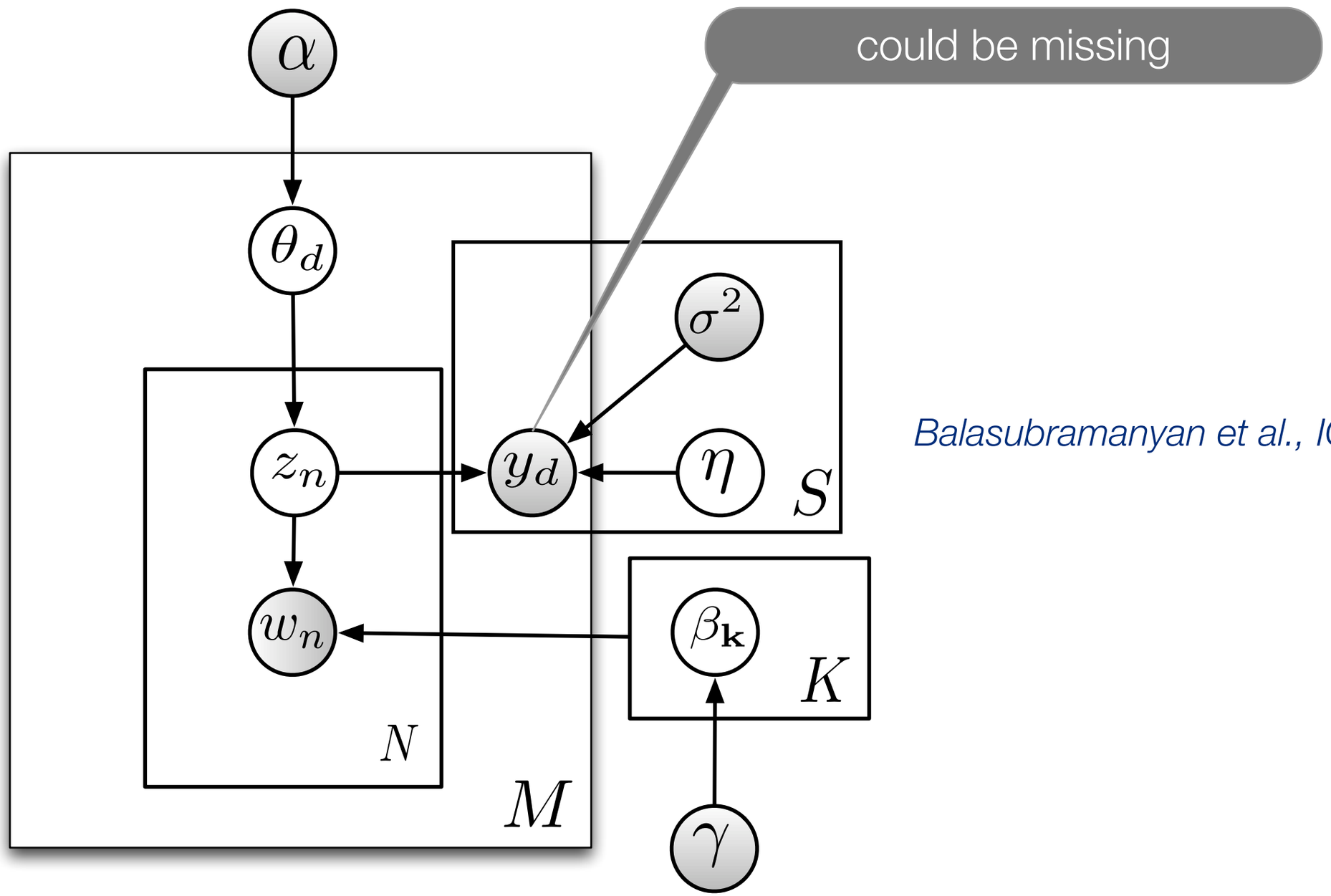
Obtaining sentiment polarity from comments

Data: blog vocabulary V, standard sentiment word lists P and N
Result: Blog specific sentiment word lists
for w in V **do**
$avg_pos_PMI \leftarrow \frac{\sum_{s \in P} PMI(w,s)}{|P|}$
$avg_neg_PMI \leftarrow \frac{\sum_{s \in N} PMI(w,s)}{|N|}$
$polarity \leftarrow avg_pos_PMI - avg_neg_PMI$
end for
$sorted_V \leftarrow V$ sorted by $polarity$
$positive_words \leftarrow$ top N of $sorted_V$
$negative_words \leftarrow$ bottom N of $sorted_V$
return $positive_words, negative_words$

Algorithm 1: Using PMI to construct blog specific sentiment word lists

Multi Community Response LDA (MCR-LDA)

Multi target Semi-supervised LDA

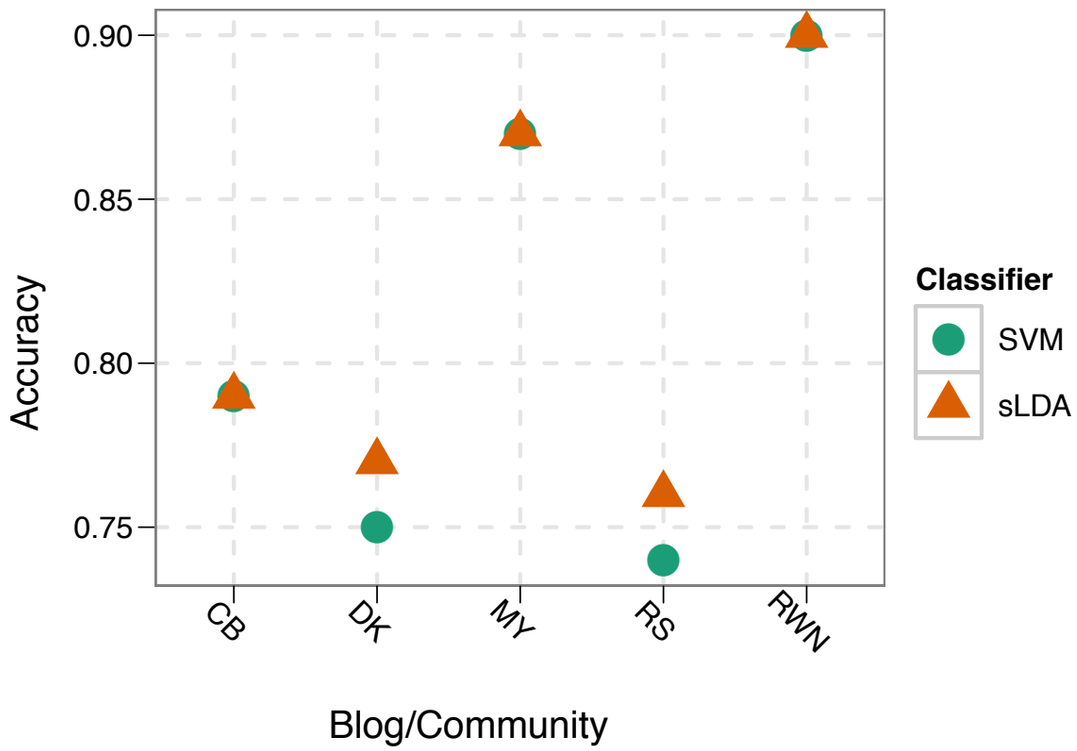


Balasubramanyan et al., ICWSM, 2012

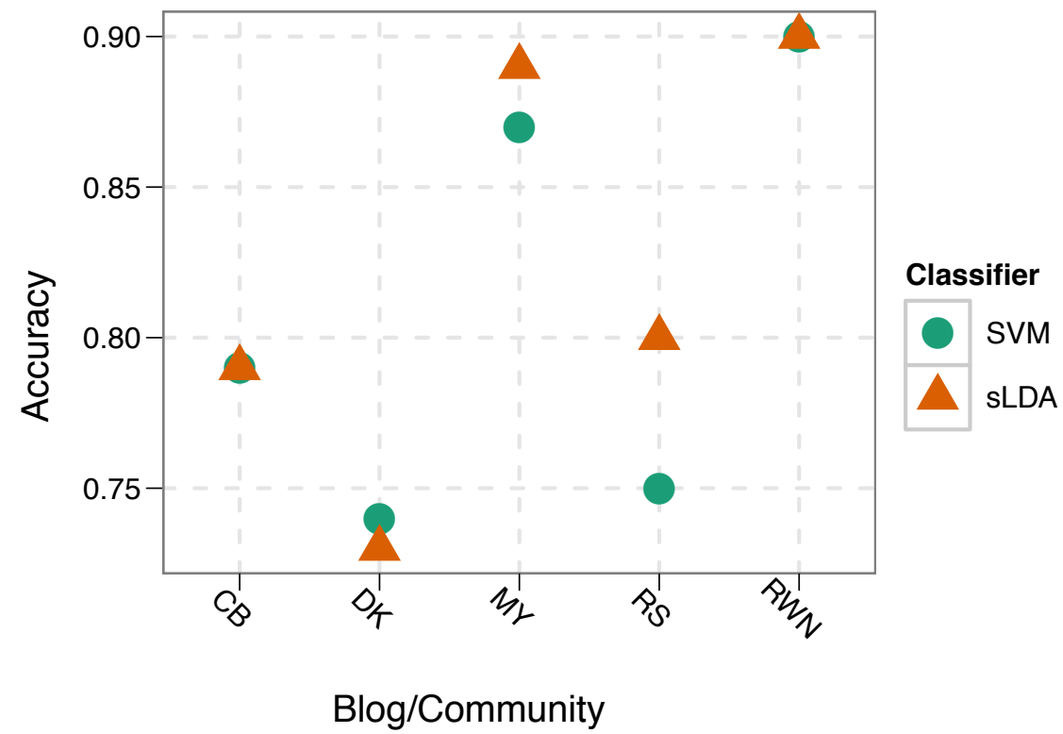
Datasets (Thanks Tae Yano & Noah Smith!)

Blog	# Posts
Carpetbagger	1201
Daily Kos	2597
Matthew Yglesias	1813
Red State	2357
Right Wing Nation	1184

Can we predict comment polarity?

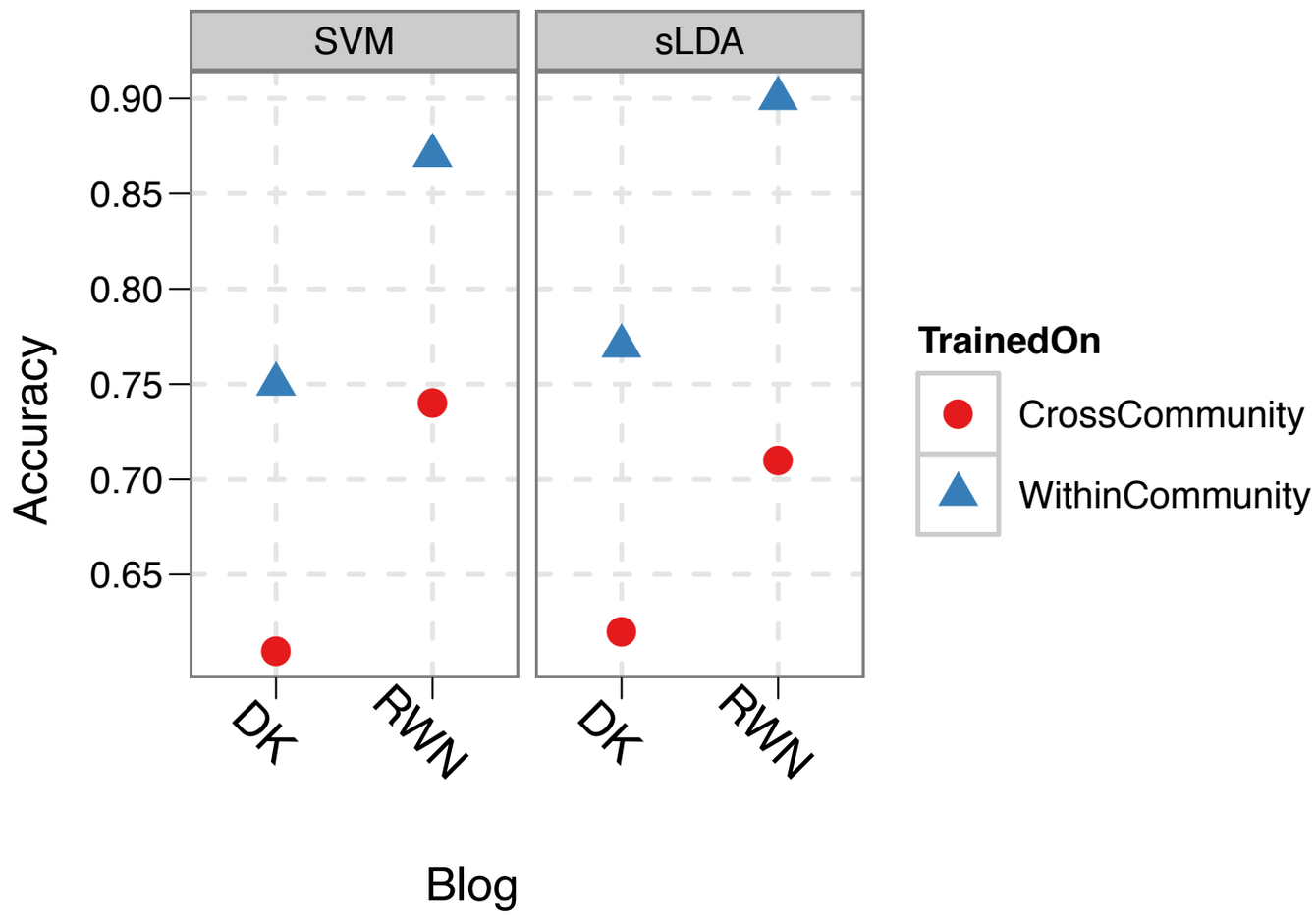


using blog posts



using comments

How important is it to be community-specific?



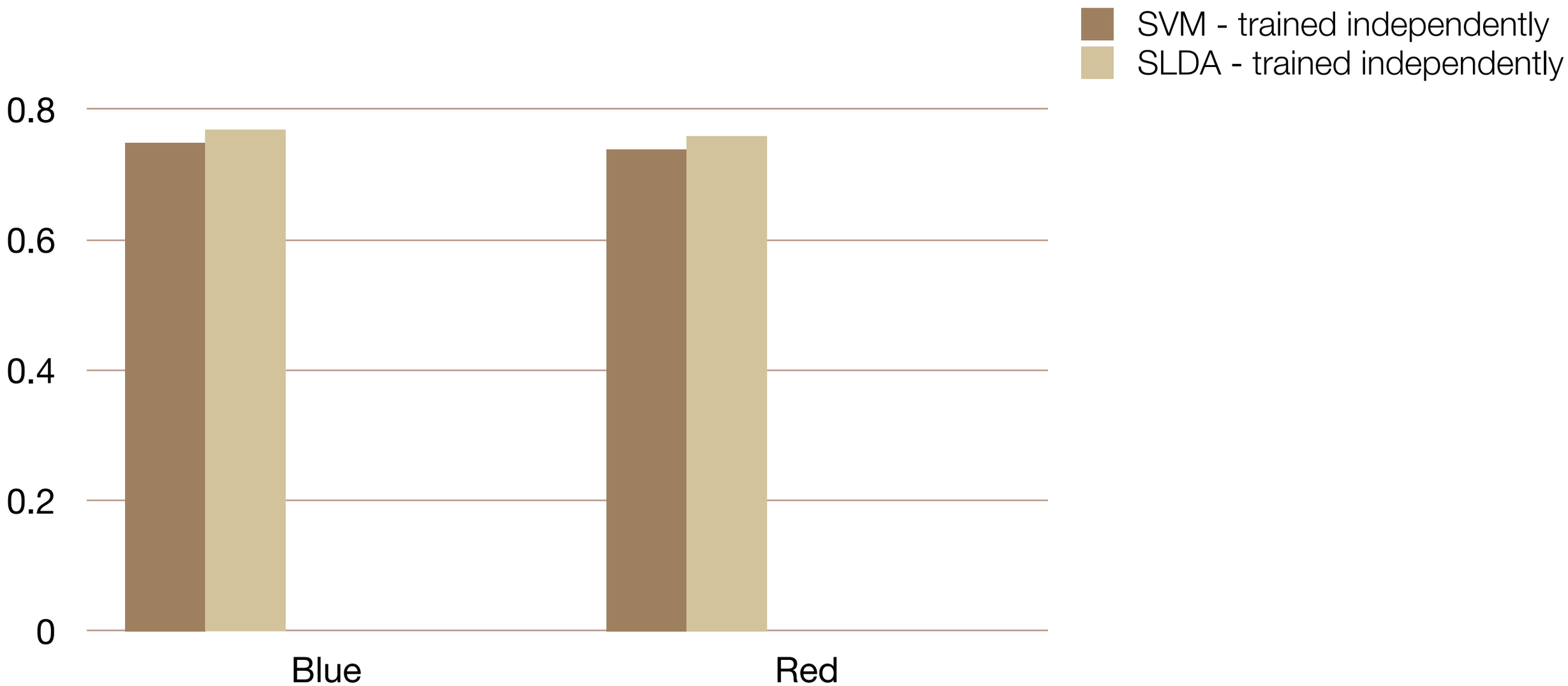
Multi Community Response LDA (MCR-LDA)

Predicting Comment Polarity

SVM - trained independently
SLDA - trained independently

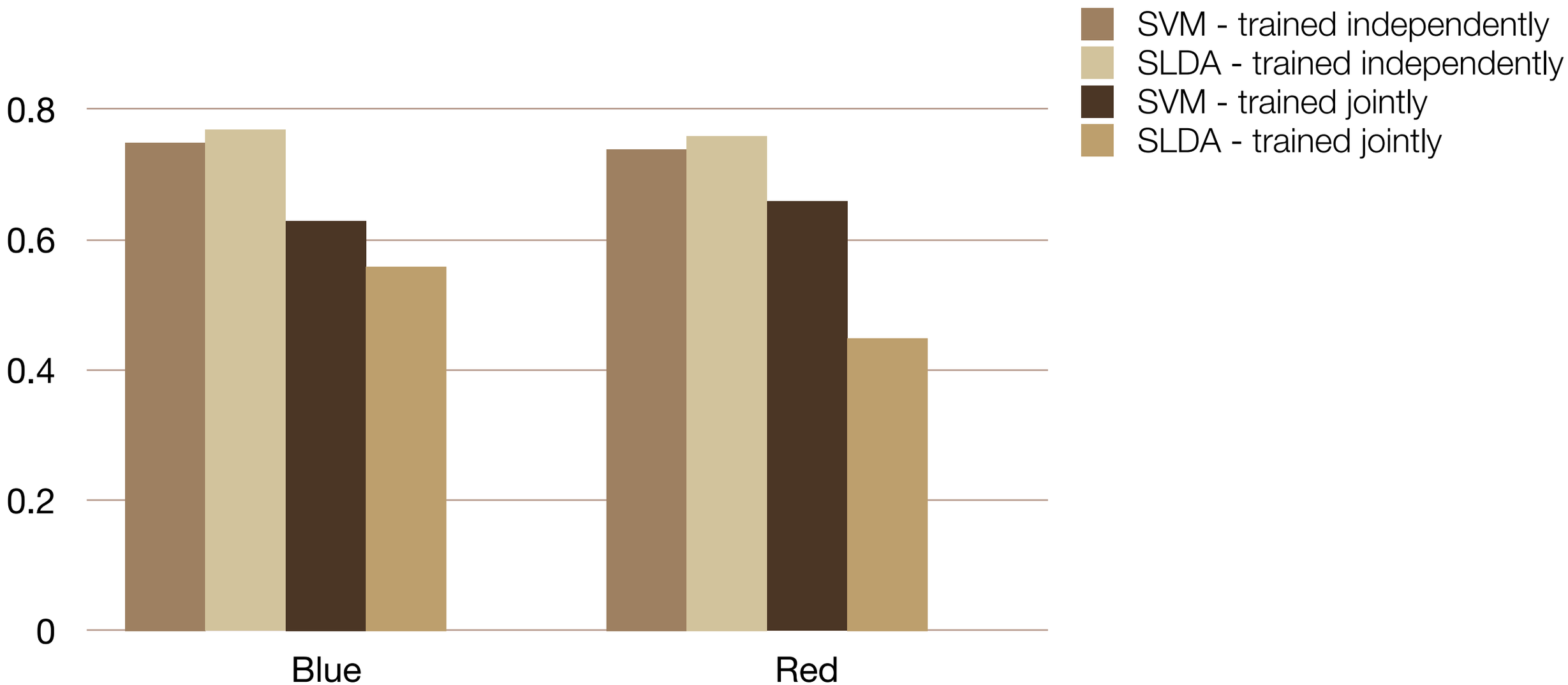
Multi Community Response LDA (MCR-LDA)

Predicting Comment Polarity



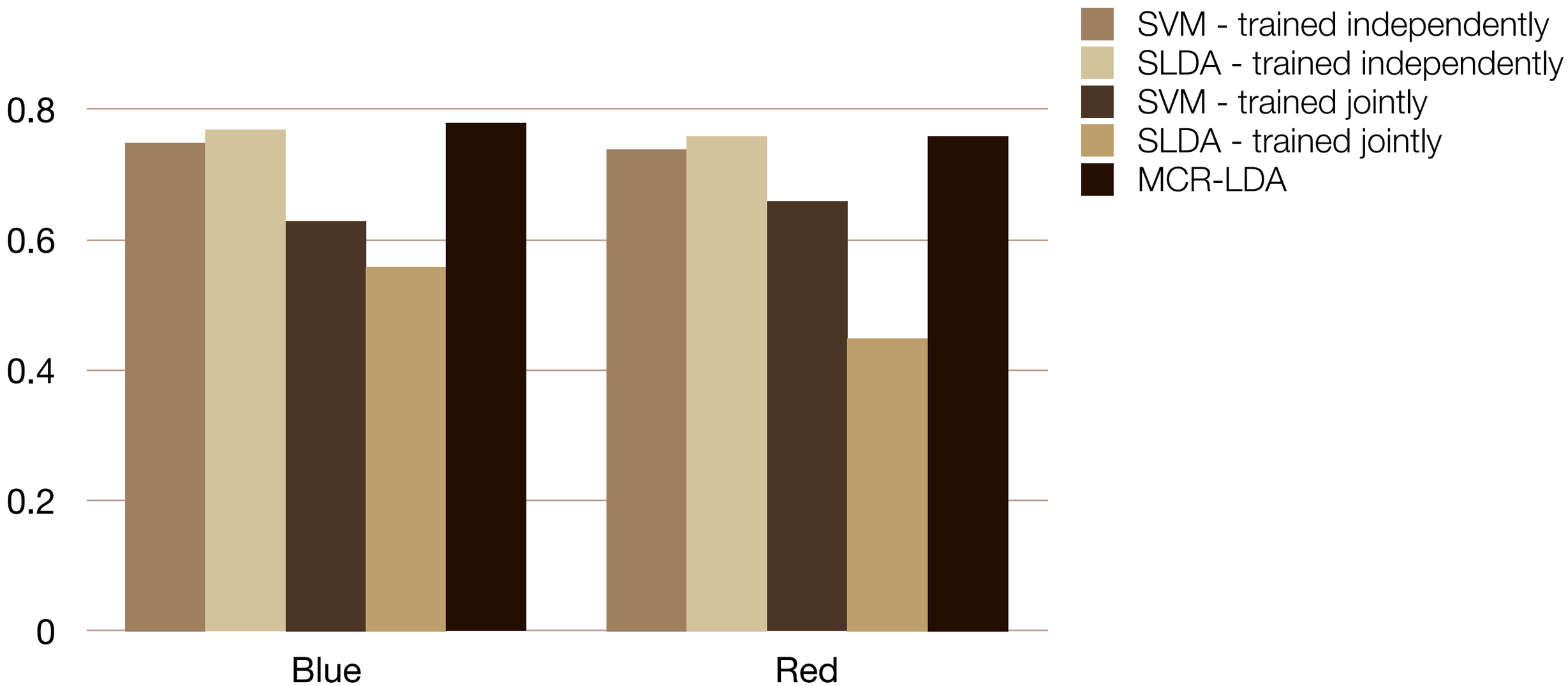
Multi Community Response LDA (MCR-LDA)

Predicting Comment Polarity



Multi Community Response LDA (MCR-LDA)

Predicting Comment Polarity



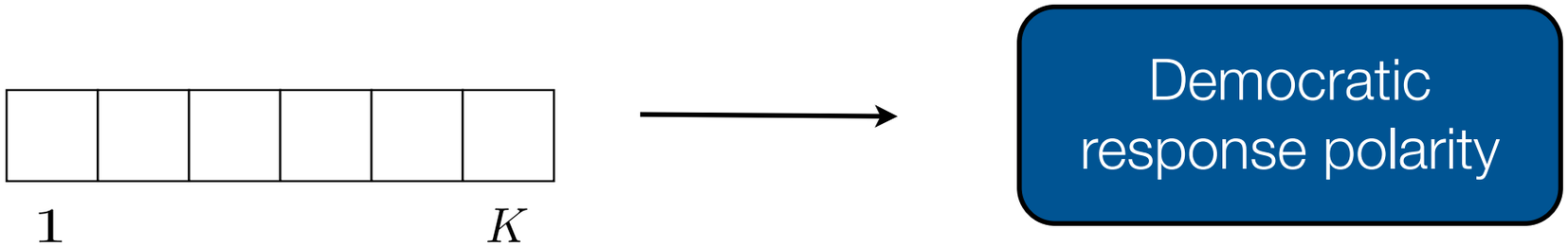
Multi Community Response LDA (MCR-LDA)

Predicting Comment Polarity

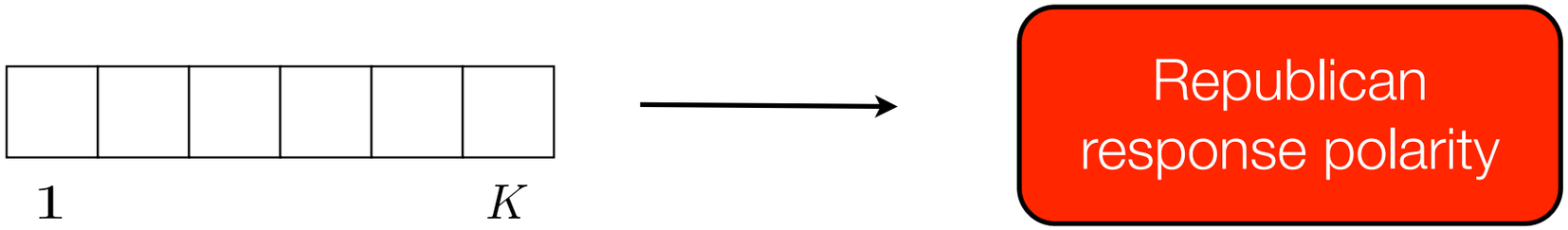
A. MCR-LDA matches the predictive performance of SVM/SLDA trained on a per-community basis

B. Helps identify polarizing and unifying topics - identified by sorting topics between Red & Blue comment polarity regression coefficients

Detecting polarizing topics



Regression co-efficients



Multi Community Response LDA (MCR-LDA)

Blue Topics

Union & Women's rights

Multi Community Response LDA (MCR-LDA)

Blue Topics

Energy & Environment

Union & Women's rights

Multi Community Response LDA (MCR-LDA)

Red Topics

Senate Procedures

Republican Primaries

Multi Community Response LDA (MCR-LDA)

Neutral Topics

Economy, taxes, social security

Mid term elections

chatter in the twitterverse

tweet categorization - by intent

- conversational - queries etc.
- status / daily chatter - state of mind, activities
- information sharing - retweets
- news - sports, events, weather, current headlines

tweet chatter detector

enables identification of content type

Combine the two	Topical	Not Topical
Not Chatter	✦ news	✦ spam?
Chatter	✦ information sharing with commentary	✦ conversational ✦ status updates

tweet chatter detector

enables identification of content type

Combi **ty**		cal
Not C		
Chatter	sharing with commentary	✦ conversational ✦ status updates

definition of chatter: “does the tweet present any personal input from the tweeter?”

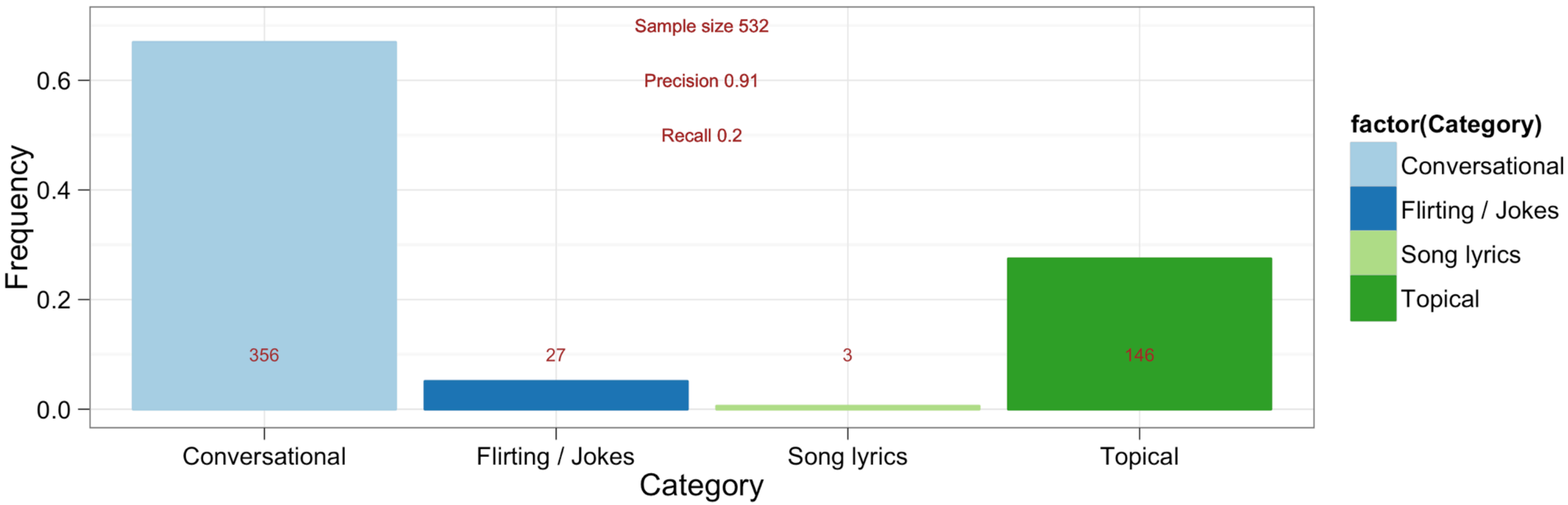
why?

- signal for search relevance
- ad-targeting
- provide filter options
- ...

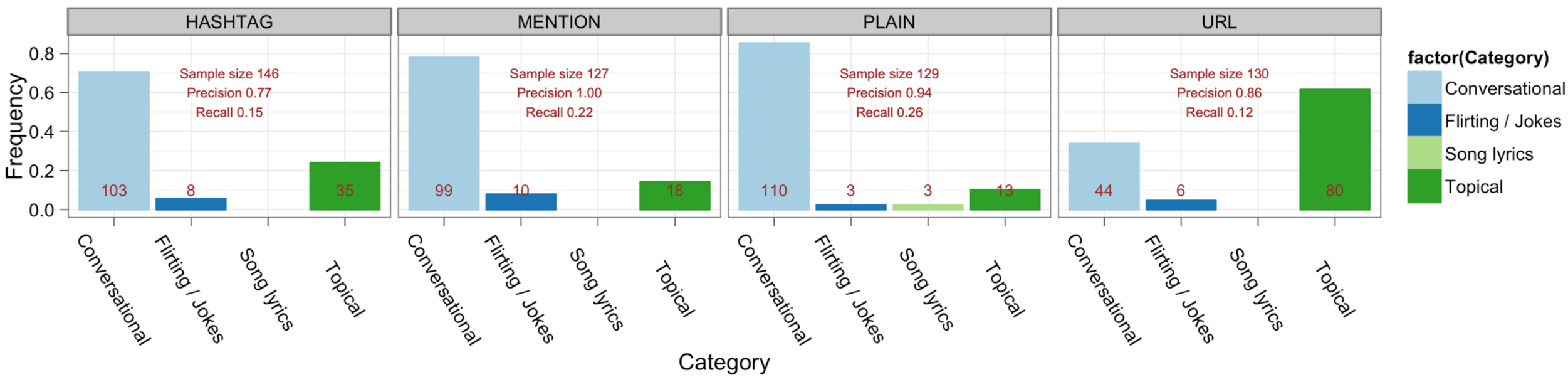
chatter prevalence evaluation using mturk

- 800 tweets randomly sampled
- broken into tweet-characteristic buckets
  - contains hashtag
  - contains @mentions
  - contains URLs
  - does not contain any of these
- valid responses for ~500 tweets

What fraction of tweets have chatter?



tweet type breakdown



tweets which are plain are more likely to be conversational
tweets with URLs are less likely to be conversational

chatter and engagement

Type	Reply	Retweet	Favorite
Hashtag	18.02	11.71	4.50
	11.43	17.14	5.71
URL	12.00	18.00	4.00
	6.25	12.50	0.075
Plain	15.51	24.14	7.76
	7.69	7.69	0
Mention	40.36	11.00	5.50
	27.77	0	0
All	**22.79**	**16.06**	**5.69**
	10.27	**11.69**	**5.48**

tl;dr - conversational tweets get replied to (2x) and retweeted

chatter and engagement

Type	Reply	Retweet	Favorite
Hashtag	18.02	11.71	4.50
	11.43	17.14	5.71
URL	12.00	18.00	4.00
	6.25	12.50	0.075
Plain	15.51	24.14	7.76
	7.69	7.69	0
Mention	40.36	11.00	5.50
	27.77	0	0
All	**22.79**	**16.06**	**5.69**
	10.27	**11.69**	**5.48**

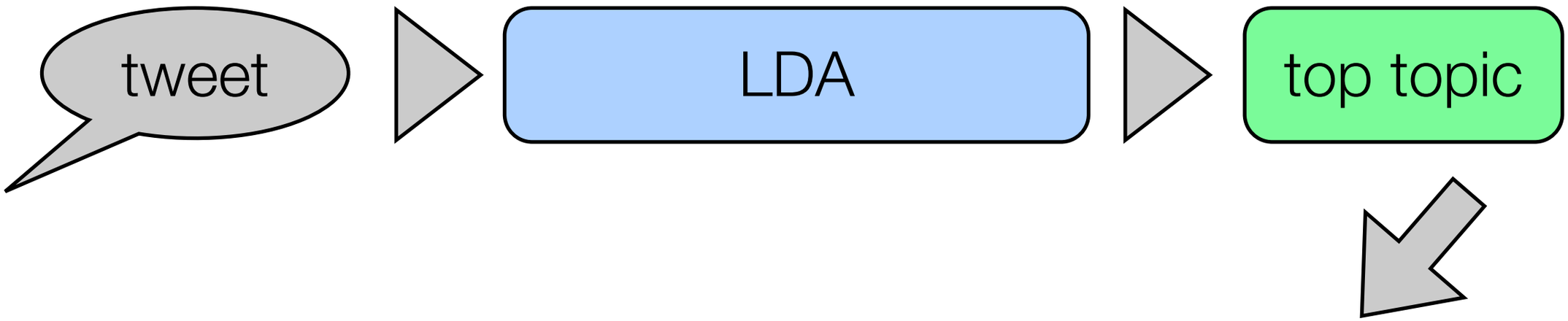
exception: conversational tweets get retweeted less than topical tweets

tl;dr - conversational tweets get replied to (2x) and retweeted

tl;dr

- 78% tweets are pure chatter - status updates and conversations
- 14% are news-like
- 8% are both i.e. offer commentary on news-like stories

how do we detect chatter?



uses a pre-judged list of chatter topics

if top topic is “chatter-like”, the tweet has chatter

Precision: 0.9
Recall: 0.2

a random sample of tweets labeled as chatter is used as training examples for a “chatter” category in the tweet classifier

chatter classifier - next version

- ✦ uses a decision tree trained on human labeled tweets
- ✦ features
  - ✦ morphological - exclamations, capitalization
  - ✦ twitter-specific - url present?, hashtag present?
  - ✦ network - #followers, #followees, ratio, tweepcred ...
  - ✦ LDA top topic
- ✦ similar to the previous version, use random sample

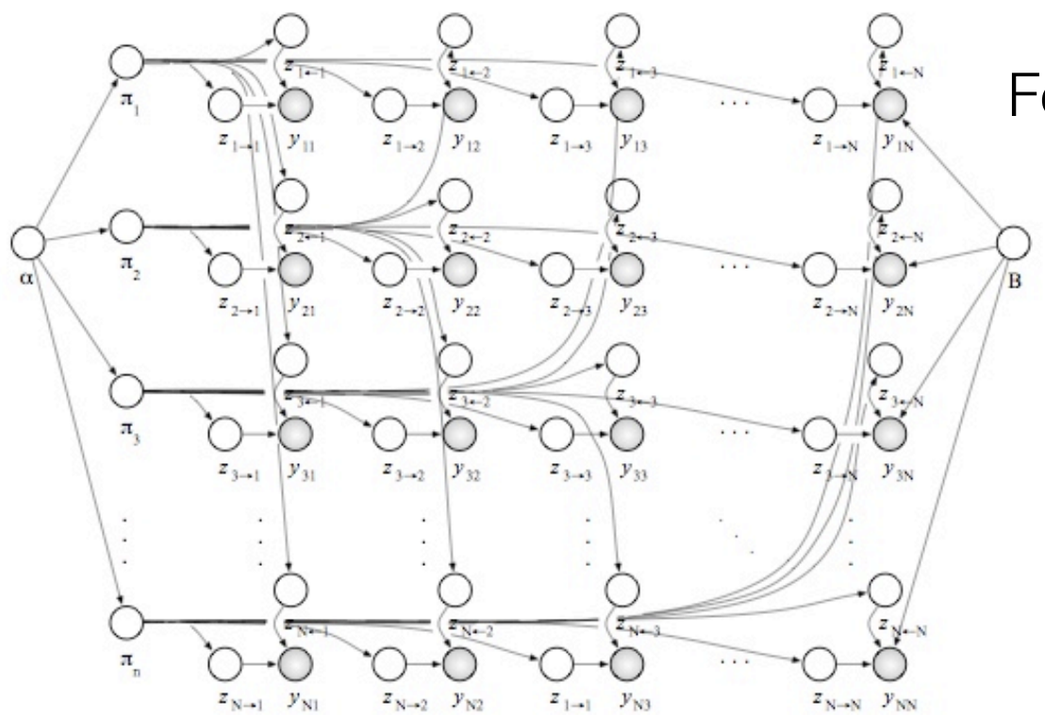
Performance in predicting chatter

Heuristic	Recall	Precision
Chatter-LDA	0.9	0.2
Chatter-DTree	0.87	0.83
MLR (threshold at 0.6616644)	1.00	0.03
MLR (threshold at 0.58)	0.99	0.28

Block-LDA: Joint Modeling Of Entity-entity Links & Entity-annotated text

SDM 2011 -
Phoenix, AZ

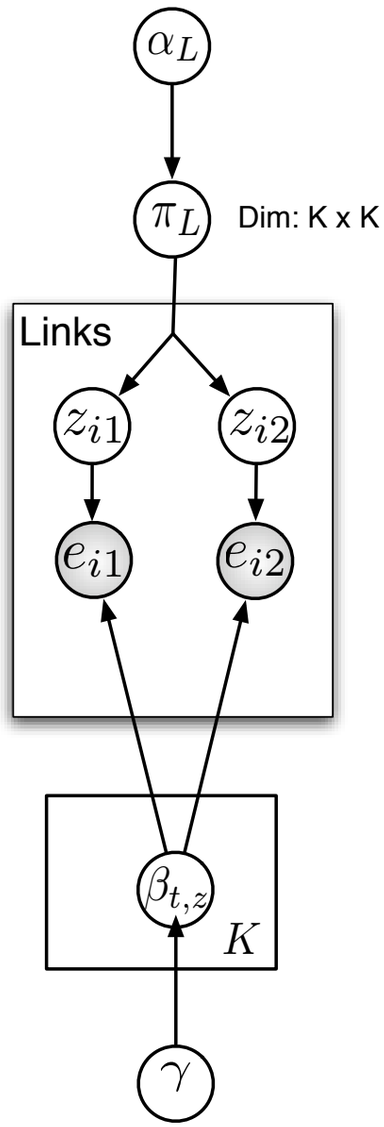
Mixed Membership Block Models (Airoldi et al., JMLR, 2008)



For each protein p,
    Draw a K dimensional mixed membership vector

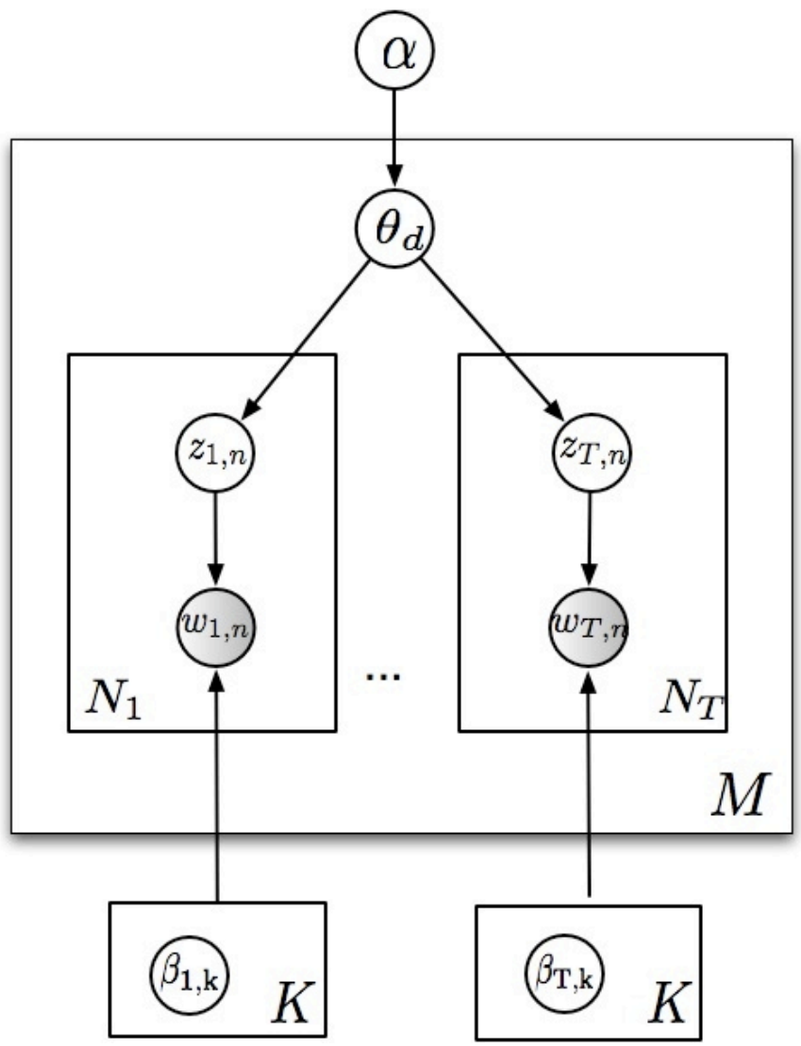
For each pair of nodes (p,q)
    Draw membership indicator $z_{p \to q}$ from π_p Multinomial
    Draw membership indicator $z_{q \to p}$ from π_q Multinomial
    Sample the value of their interaction Y(p,q) from
        Bernoulli($z_{q \to p}$ B $z_{p \to q}$)

Sparse Block Model - (Parkinnen et al, 2007)



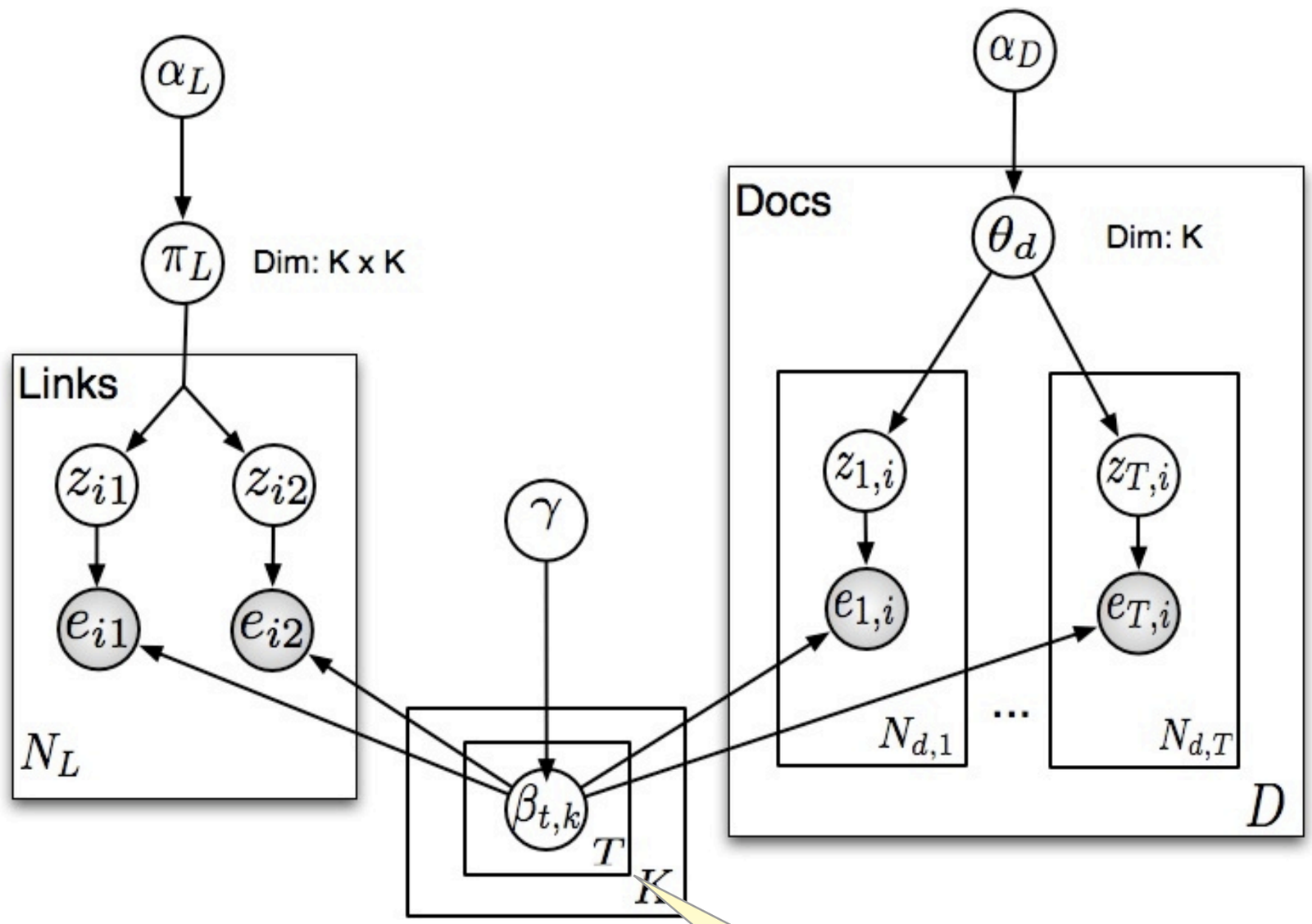
- More suitable for sparse matrices
- Easier to sample from

Modeling entity annotated text



Link LDA

Block-LDA: Jointly modeling links and text



sharing entity distributions

Gibbs Sampler

- entity entity links

Sampling the class pair for a link

$$p(\mathbf{z_i} = \langle z_1, z_2 \rangle | \langle e_{i1}, e_{i2} \rangle, \mathbf{z}^{\neg \mathbf{i}}, \langle \mathbf{e_1}, \mathbf{e_2} \rangle^{\neg \mathbf{i}}, \alpha_L, \gamma)$$

$$\propto \left(n^{L \neg i}_{\langle z_1, z_2 \rangle} + \alpha_L \right)$$

$$\frac{\left(n^{\neg i}_{z_1 t_l e_{i1}} + \gamma\right)\left(n^{\neg i}_{z_2 t_l e_{i2}} + \gamma\right)}{\left(\sum_e n^{\neg i}_{z_1 t_l e} + |E_{t_l}|\gamma\right)\left(\sum_e n^{\neg i}_{z_2 t_l e} + |E_{t_l}|\gamma\right)}$$

probability of class pair in the link corpus

probability of the two entities in their respective classes

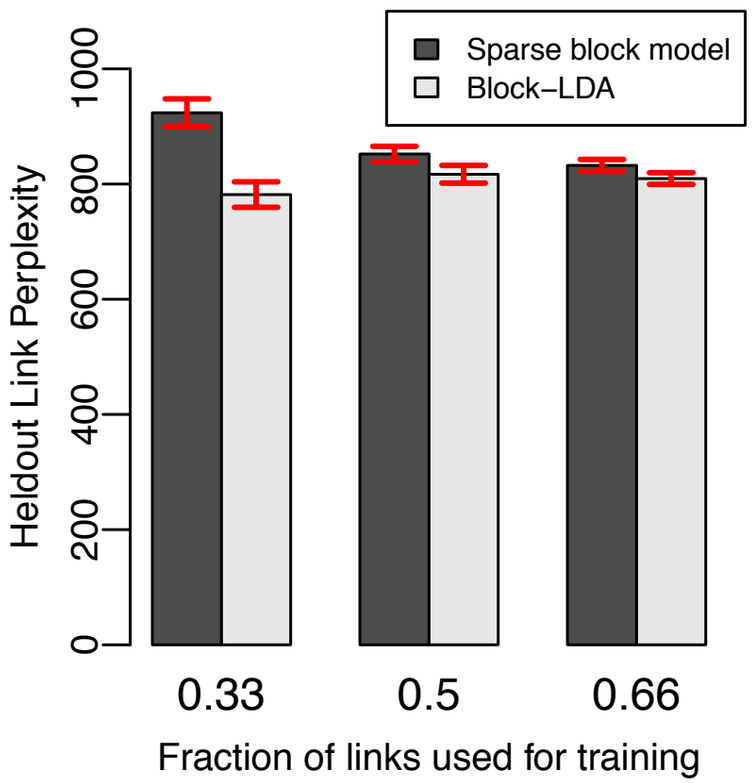
Enron corpus

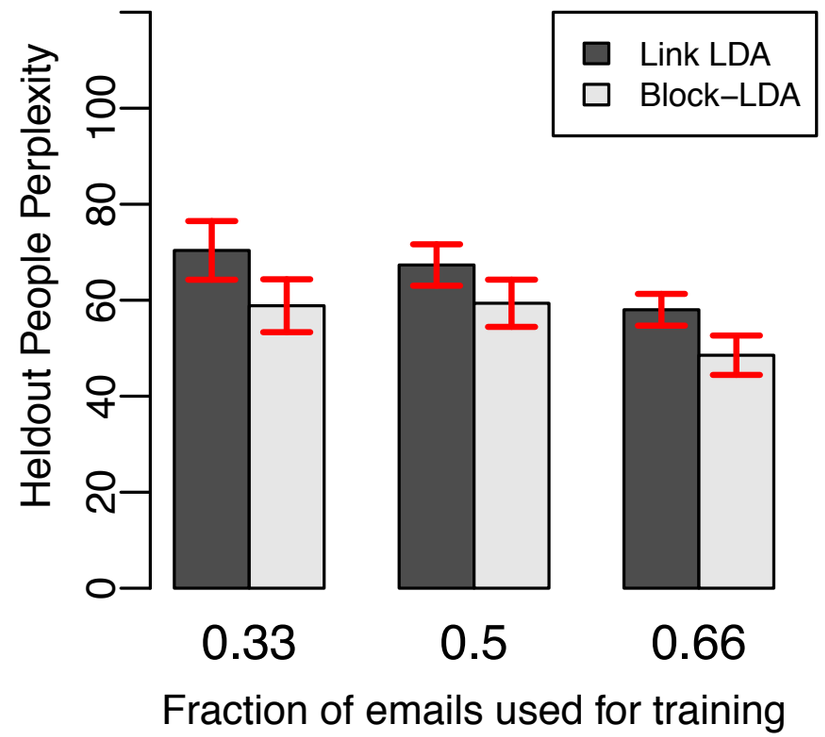
- 96,103 emails
- Link A -> B indicates person A sent an email to person B (either listed in the To or CC fields)
- Can we
- Identify interesting blocks of users?
- Use text of email in predicting links?

Examples of topics induced from the Enron email corpus

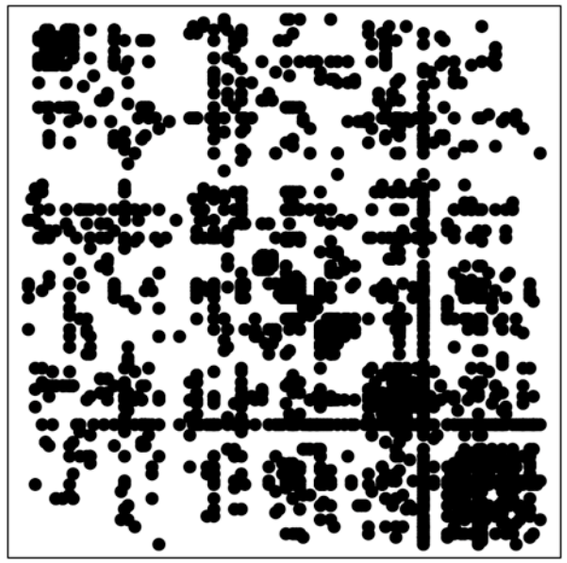
contract, party, capacity, gas, df, payment, service, tw, pipeline, issue, rate, section, project, time, system, transwestern, date, el, payment, due, paso *fossum, scott, harris, hayslett, campbell, geaccone, hyatt, corman, donoho, lokay* Notes: Geaconne was the executive assistant to Hayslett who was the Chief Financial Officer and Treasurer of the Transwestern division of Enron.	Financial Contracts
power, california, energy, market, contracts, davis, customers, edison, bill, ferc, price, puc, utilities, electricity, plan, pge, prices, utility, million, jeff *dasovich, stevies, shapiro, kean, williams, sanders, smith, lewis, wolfe, bass* Notes: Dasovitch was a Government Relations executive, Steffies the VP of government affairs, Shapiro, the VP of regulatory affairs and Haedicke worked for the legal department.	Energy Distribution
enron, business, management, risk, team, people, rick, process, time, information, issues, sally, mike, meeting, plan, review, employees, operations, project, trading *kitchen, beck, lavorato, delainey, buy, presto, shankman, mcconnell, whalley, haedicke* Notes: The people in this topic are top level executives: Kitchen was the President of Enron Online, Beck the Chief operating officer and Lavarato the CEO.	Strategy

Experiment with the Enron corpus

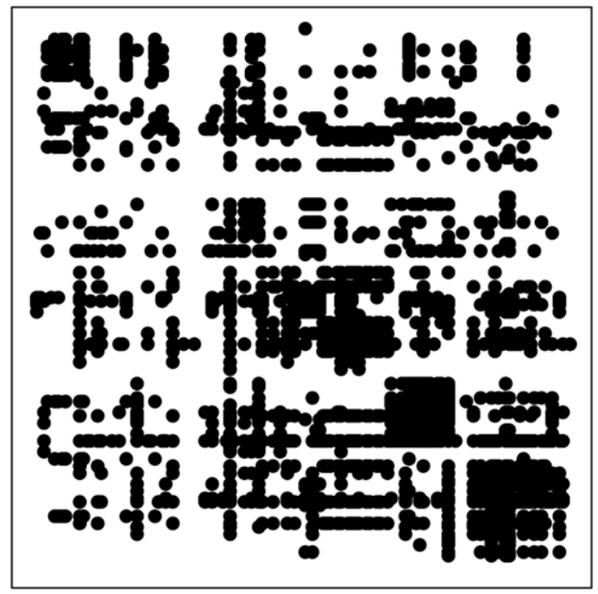




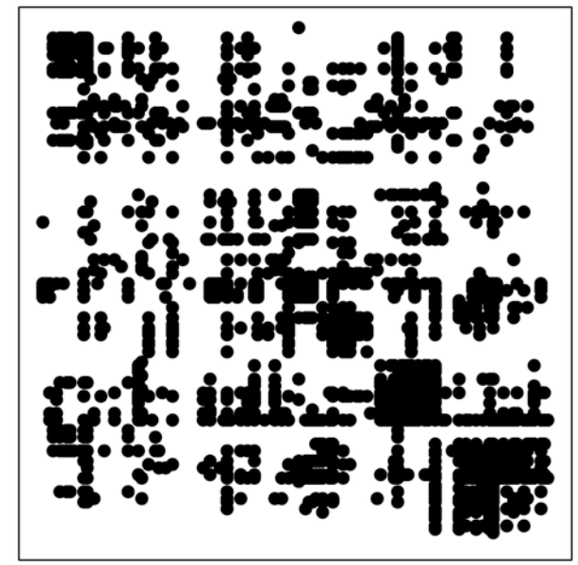
Enron corpus



Enron network



Sparse model



Block LDA

Annotated Text - *Saccharomyces* Genome Database

A scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae

- Database contains **protein annotations** in publications about yeast.
- We use 16K publications annotated with at least one protein present in the MIPS protein interactions.

Annotated Text - *Saccharomyces* Genome Database

A scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae

- Database contains **protein annotations** in publications about yeast.
- We use 16K publications annotated with at least one protein present in the MIPS protein interactions.

Vac1p coordinates Rab and phosphatidylinositol 3-kinase signaling in Vps45p-dependent vesicle docking/ fusion at the endosome.

The vacuolar protein sorting (VPS) pathway of Saccharomyces cerevisiae mediates transport of vacuolar protein precursors from the late Golgi to the lysosome-like vacuole. Sorting of some vacuolar proteins occurs via a prevacuolar endosomal compartment and mutations in a subset of VPS genes (the class D VPS genes) interfere with the Golgi-to-endosome transport step. Several of the encoded proteins, including Pep12p/Vps6p (an endosomal target (t) SNARE) and Vps45p (a Sec1p homologue), bind each other directly [1]. Another of these proteins, Vac1p/Pep7p/Vps19p, associates with Pep12p and binds phosphatidylinositol 3-phosphate (PI(3)P), the product of the Vps34 phosphatidylinositol 3-kinase (PI 3-kinase) .....

PEP7 VPS45 VPS34 PEP12 VPS21

Annotated Text - *Saccharomyces* Genome Database

A scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae

- Database contains **protein annotations** in publications about yeast.
- We use 16K publications annotated with at least one protein present in the MIPS protein interactions.

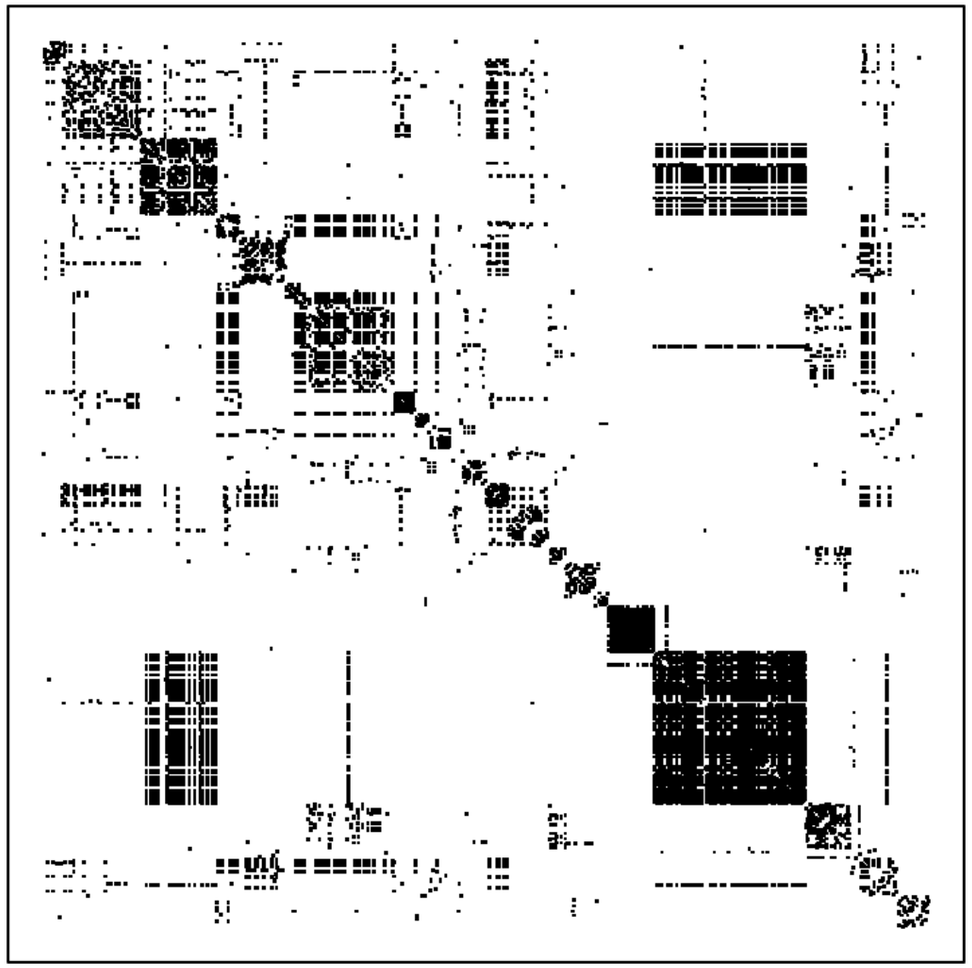
Vac1p coordinates Rab and phosphatidylinositol 3-kinase signaling in Vps45p-dependent vesicle docking/ fusion at the endosome.

The vacuolar protein sorting (VPS) pathway of Saccharomyces cerevisiae mediates transport of vacuolar protein precursors from the late Golgi to the lysosome-like vacuole. Sorting of some vacuolar proteins occurs via a prevacuolar endosomal compartment and mutations in a subset of VPS genes (the class D VPS genes) interfere with the Golgi-to-endosome transport step. Several of the encoded proteins, including Pep12p/Vps6p (an endosomal target (t) SNARE) and Vps45p (a Sec1p homologue), bind each other directly [1]. Another of these proteins, Vac1p/Pep7p/Vps19p, associates with Pep12p and binds phosphatidylinositol 3-phosphate (PI(3)P), the product of the Vps34 phosphatidylinositol 3-kinase (PI 3-kinase) .....

PEP7 VPS45 VPS34 PEP12 VPS21

Protein Annotations

Protein Protein Interaction Data



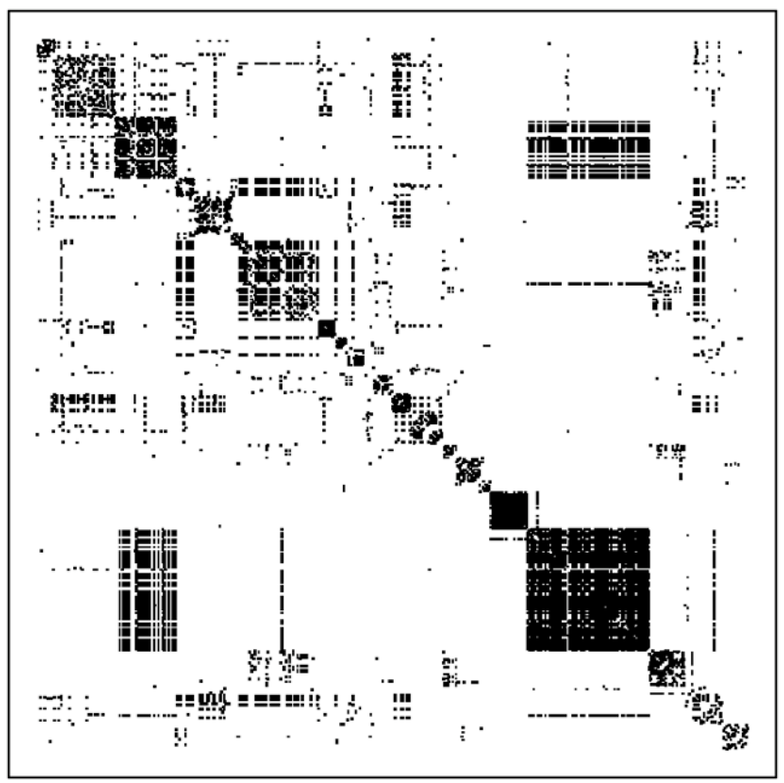
- Source: Munich Information Center for Protein Sequences (MIPS)
- 844 proteins identified by high throughput methods

Is there information about Protein interactions in text?

Let an abstract be annotated with n proteins P= {p_1, p_2, p_3 ... p_n}

We construct “interactions” by building a Cartesian product P x P resulting in links such as <p_1,p_1>, <p_1,p_2> ... <p_n, p_n> and applying a min frequency count threshold

MIPS interactions



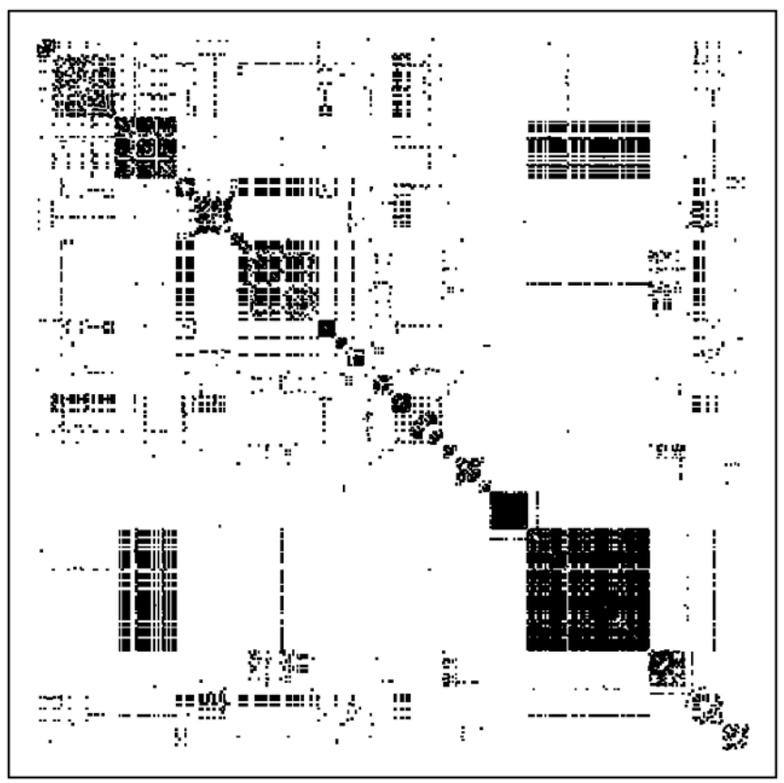
Text Co-occurences

Is there information about Protein interactions in text?

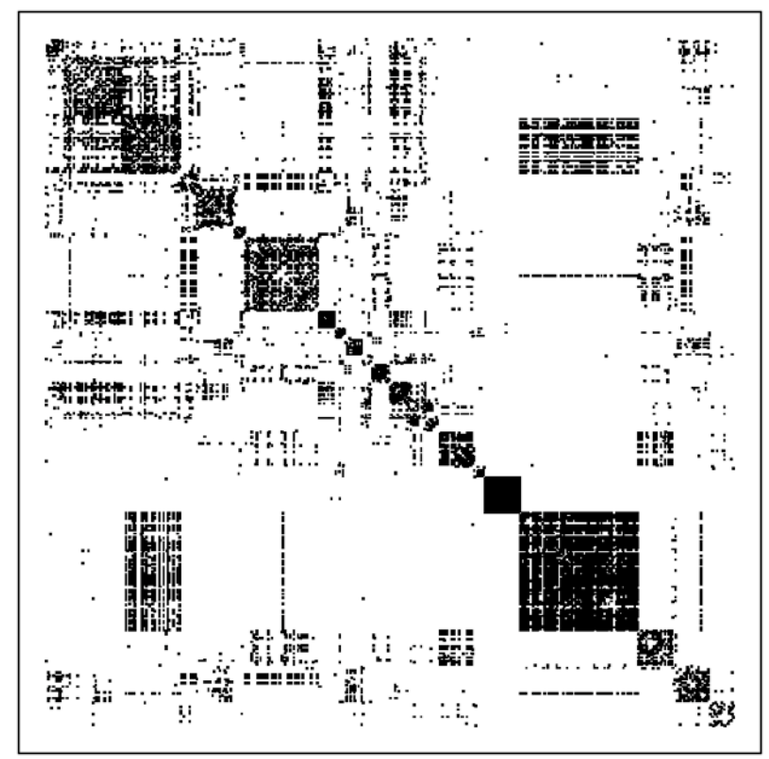
Let an abstract be annotated with n proteins P= {p_1, p_2, p_3 ... p_n}

We construct “interactions” by building a Cartesian product P x P resulting in links such as <p_1,p_1>, <p_1,p_2> ... <p_n, p_n> and applying a min frequency count threshold

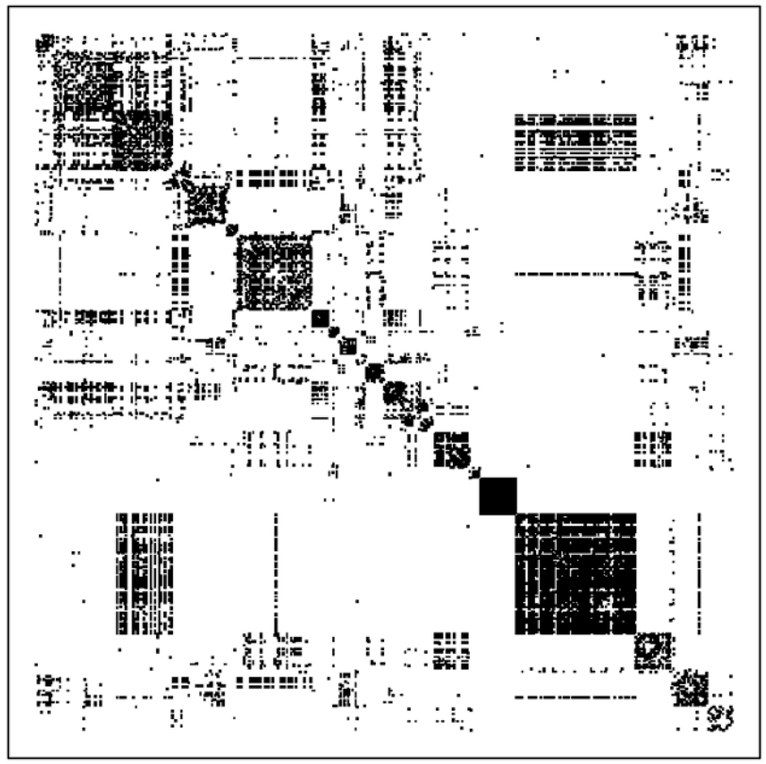
MIPS interactions

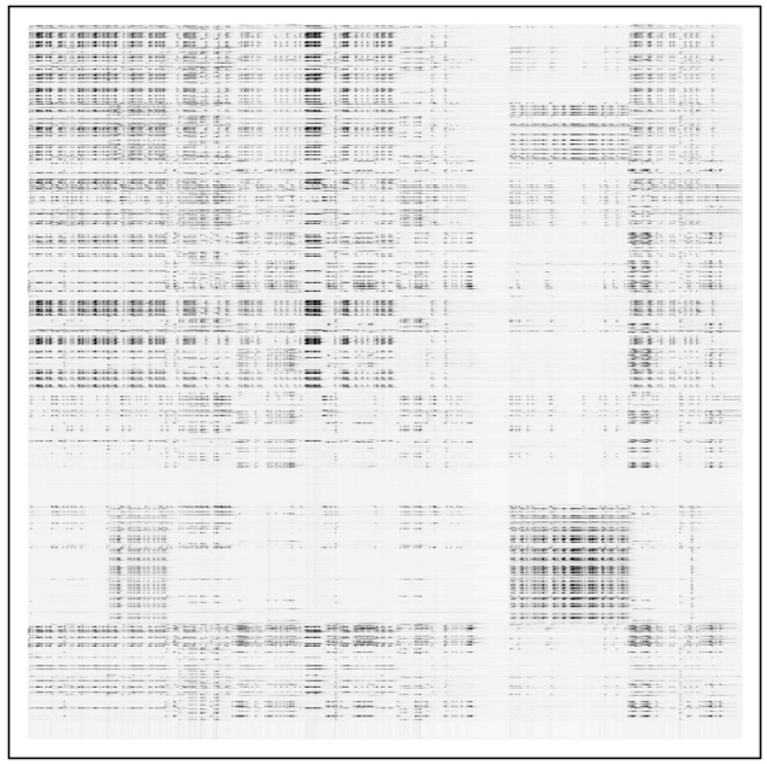


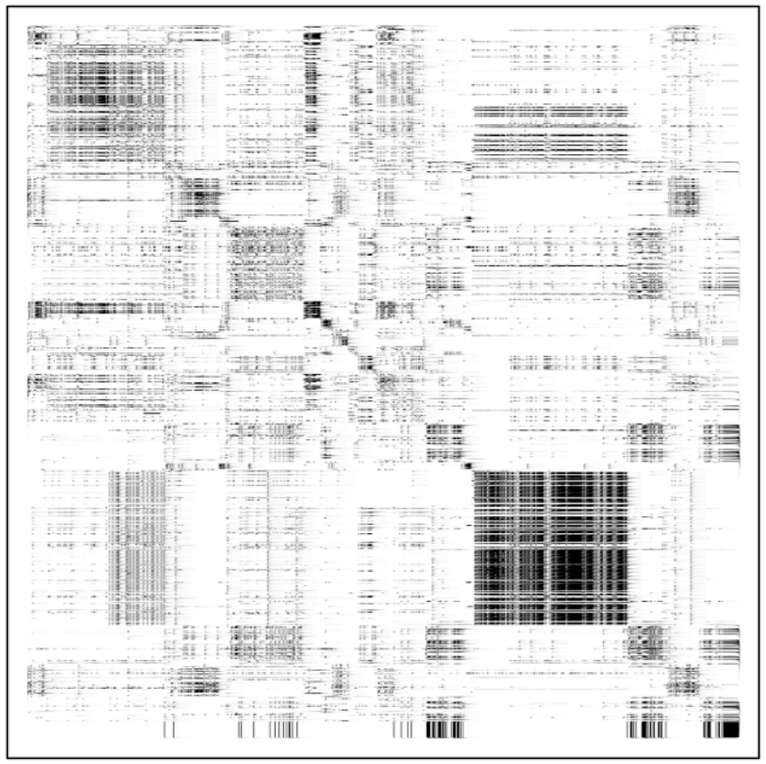
Text Co-occurences



Recovering the interaction matrix

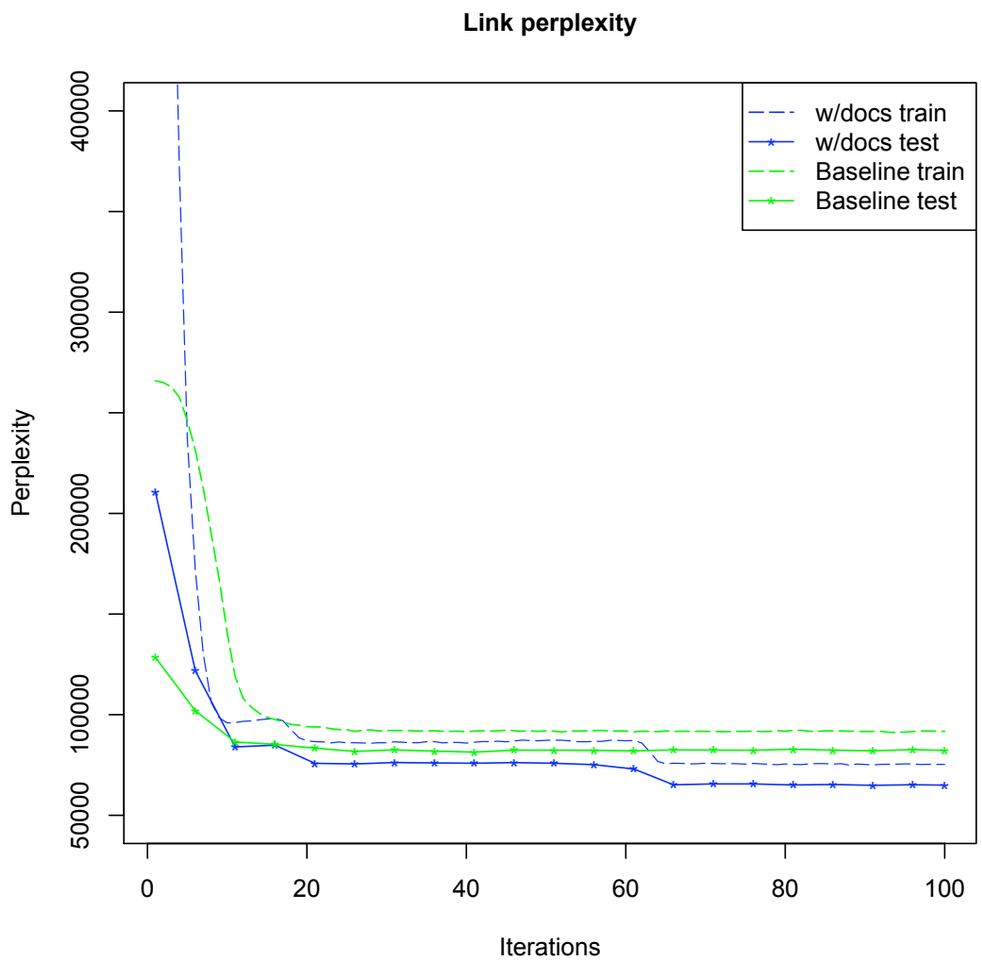






MIPS interactions Sparse Block model Block-LDA

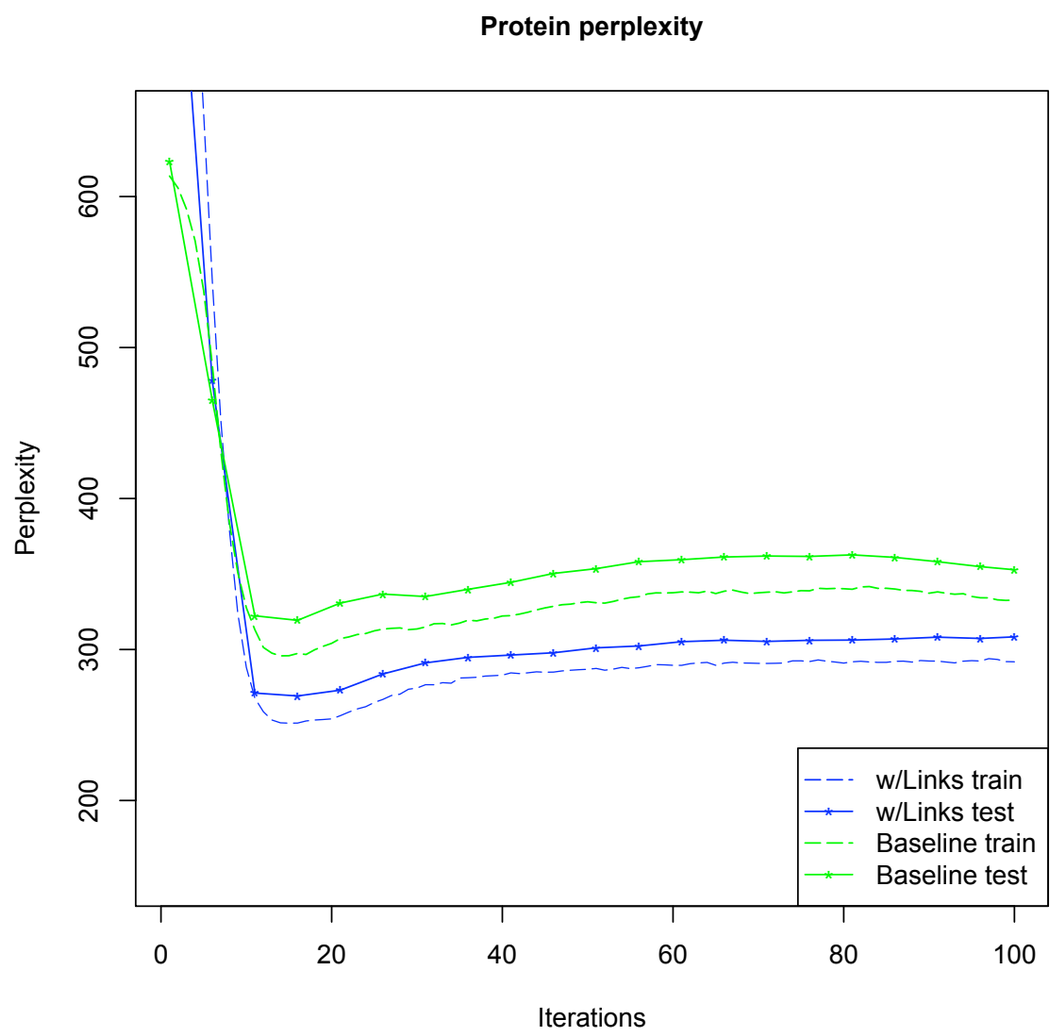
Evaluation using Link Perplexity



1/3 of links + all text used for training

2/3 of links used for testing

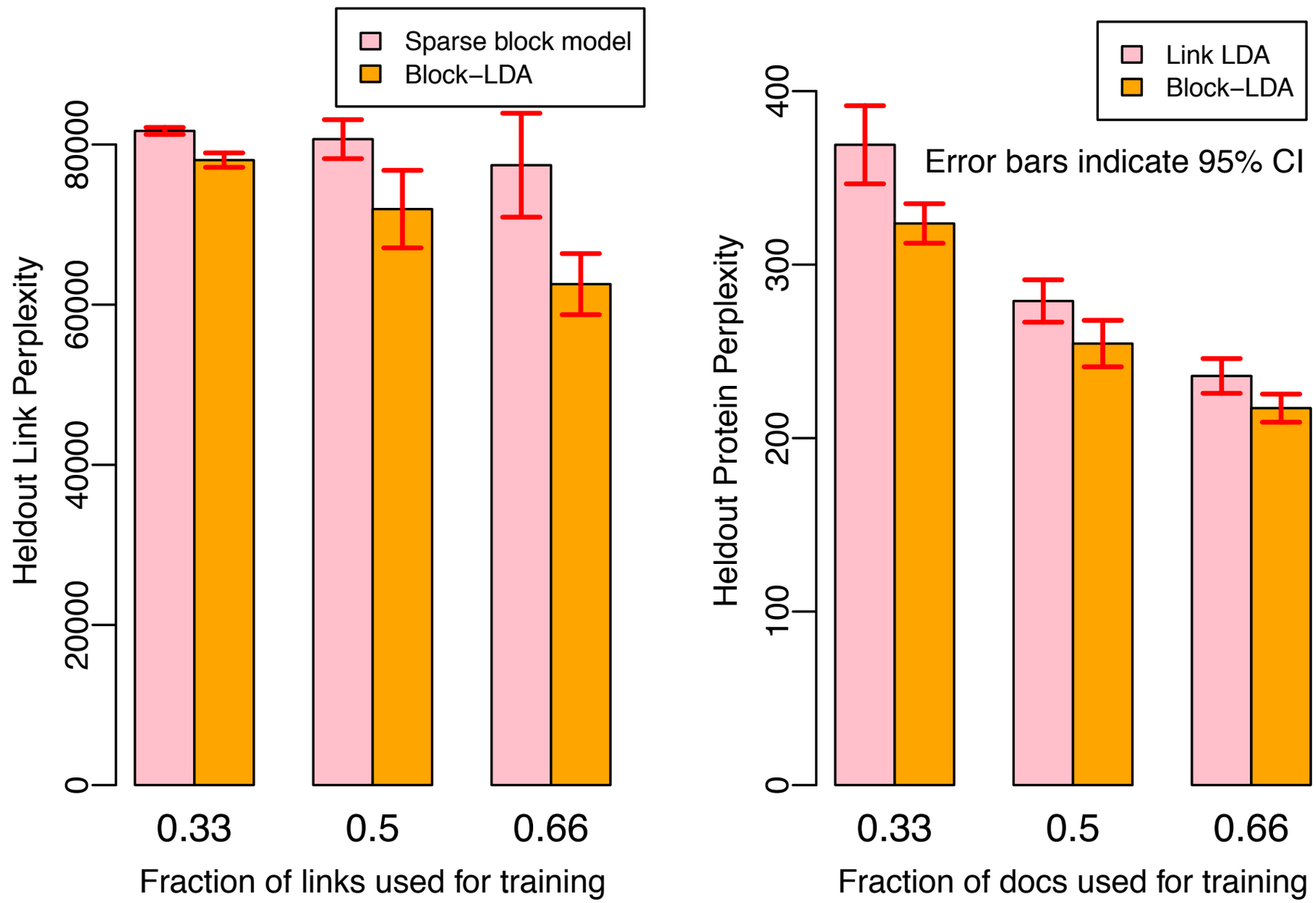
Evaluation using Protein Perplexity in text



1/3 of docs + all links used for training

2/3 of text used for testing

Varying Training Data



Sample topics

mutant mutants gene cerevisiae growth type mutations saccharomyces wild mutation strains strain phenotype genes deletion temperature resistance sensitive albicans wall defect sensitivity defects phenotypes candida	rpl20b rpl5 rpl16a rps5 rpl39 rpl18a rpl27b rps3 rpl23a rpl1b rpl32 rpl17b rpl35a rpl26b rpl31a rpp2a rpp0 rpl7a rpl10 rpl20a rpl34b rpp1b rpl24a rpl40b rpl38	klis_fm bussey_h miyakawa_t toh-e_a heitman_j perfect_jr ohya_y moye-rowley_ws sherman_f latge_jp schaffrath_r duran_a sa-correia_i liu_h subik_j kikuchi_a chen_j goffeau_a tanaka_k kuchler_k calderone_r nombela_c popolo_l jablonowski_d

A common experimental procedure is to induce random mutations in the "wild-type" strain of a model organism (e.g., saccharomyces cerevisiae) and then screen the mutants for interesting observable characteristics (i.e. phenotype). Often the phenotype shows slower growth rates under certain conditions (e.g. lack of some nutrient). The RPL* proteins are all part of the larger (60S) subunit of the ribosome. The first two biologists, Klis and Bussey's research use this method.

Sample topics (contd)

binding	rps19b	naider_f
domain	rps24b	becker_jm
terminal	rps3	leulliot_n
structure	rps20	van_tilbeurg h_h
site	rps4a	melki_r
residues	rps11a	velours_j
domains	rps2	graille_m
interaction	rps8a	quevillon-
region	rps10b	cheruel_s
subunit	rps6a	janin_j
alpha	rps10a	zhou_cz
amino	rps19a	blondeau_k
structural	rps12	ballesta_jp
conserved	rps9b	yokoyama_s
atp	rps28a	bousset_l
beta	rps30b	vershon_ak
motif	rps18a	bowler_be
complex	rps23b	zhang_y
sequence	rps26a	arshava_b
interactions	rps14b	buchner_j
sites	rps0b	wickner_rb
subunits	rps29a	steven_ac
form	rps15	wang_y
terminus	rps16a	zhang_m
function	rps31	forgac_m
		brethes_d

Protein structure is an important area of study. Proteins are composed of amino-acid residues, functionally important protein regions are called domains, and functionally important sites are often "converved" (i.e., many related proteins have the same amino-acid at the site). The RPS* proteins all part of the smaller (40S) subunit of the ribosome. Naider, Becker, and Leulliot study protein structure.

Sample topics (contd)

transcription ii histone chromatin complex polymerase transcription al rna promoter binding dna silencing h3 factor genes gene complexes vivo pol specific tbp factors required dependent promoters	rpl16b rpl26b rpl24a rpl18b rpl18a rpl12b rpl6b rpp2b rpl15b rpl9b rpl40b rpp2a rpl20b rpl14a rpp0 rpl32 rpl37b rpl40a rpl1b rpl7a rpl27b rpl16a rpl9a rpl36a rpl3	workman_jl struhl_k winston_f buratowski_ s tempst_p erdjument- bromage_h kornberg_rd sentenac_a svejstrup_jq peterson_cl berger_sl grunstein_m stillman_dj cote_j cairns_br shilatifard_a hampsey_m allis_cd young_ra thuriaux_p zhang_z sternglanz_r krogan_nj weil_pa pillus_l

In transcription, DNA is unwound from histone complexes (where it is stored compactly) and converted to RNA. This process is controlled by transcription factors, which are proteins that bind to regions of DNA called promoters. The RPL* proteins are part of the larger subunit of the ribosome, and the RPP proteins are part of the ribosome stalk. Many of these proteins bind to RNA. Workman, Struhl, and Winston study transcription regulation and
the interaction of transcription with the restructuring of chromatin (a combination of DNA, histones, and otherproteins that comprises chomosomes).

Protein Functional Category prediction

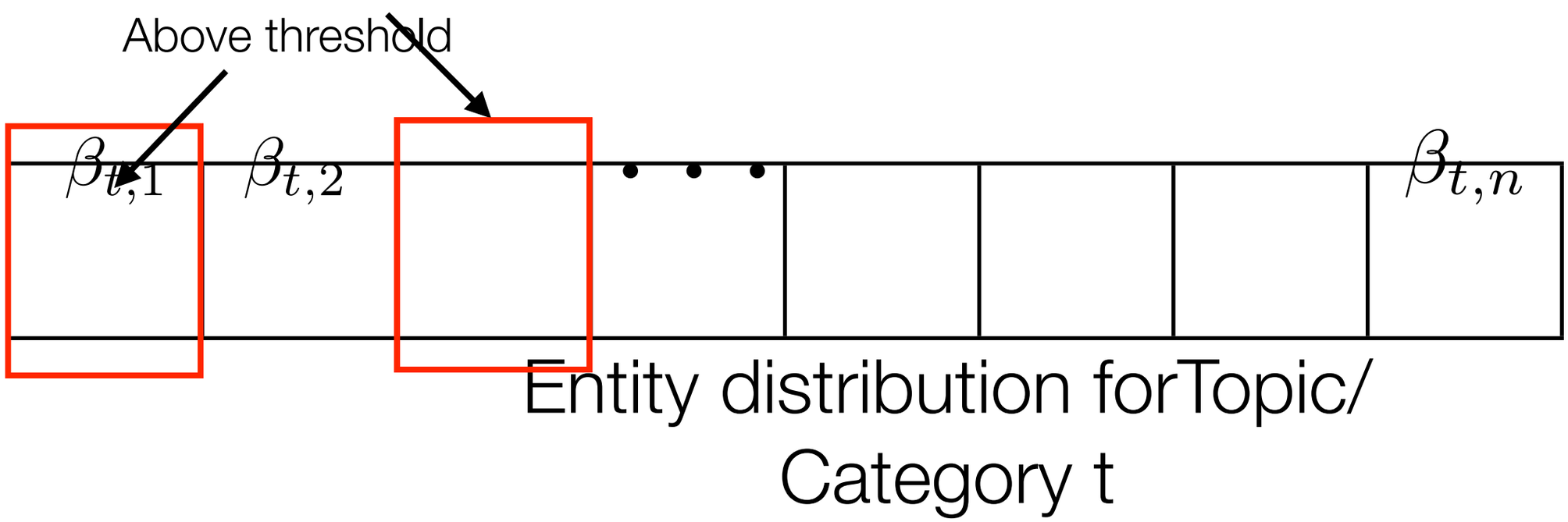
- METABOLISM
  - amino acid metabolism
    - amino acid biosynthesis
      - biosynthesis of the aspartate family
        - biosynthesis of lysine
      - biosynthesis of the cysteine-aromatic group
        - biosynthesis of serine
    - nitrogen and sulfur utilization
- ENERGY
- METABOLISM
- CELL COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
- CELL RESCUE, DEFENSE AND VIRULENCE
- CELL FATE
- CELLULAR ENVIRONMENT / INTERACTION WITH CELLULAR ENVIRONMENT
- ENERGY
- CONTROL OF CELLULAR ORGANIZATION
- CELL CYCLE AND DNA PROCESSING
- SUBCELLULAR LOCALISATION
- TRANSCRIPTION
- PROTEIN SYNTHESIS
- PROTEIN ACTIVITY REGULATION
- TRANSPORT FACILITATION
- PROTEIN FATE (folding, modification, destination)
- CELLULAR TRANSPORT AND TRANSPORT MECHANISMS

MIPS Functional Category Tree - 15 top level nodes, 255 leaf nodes. We consider only top level categories

Proteins on average associated with 2.5 top level nodes

Protein Functional Category prediction

- Train Block LDA with 15 topics (the number of top level categories)
- Map topics to functional categories using the Hungarian algorithm to find best mapping.
- For each functional category / topic, entities with probability above threshold are deemed as having that function



Performance

Method	F1	Precision	Recall
Block-LDA	**0.249**	0.247	0.25
Sparse Block Model	0.161	0.224	0.126
Link LDA	0.152	0.150	0.155
MMSB	0.165	0.166	0.164
Random	0.145	0.155	0.137

Related Work

- Link PLSA LDA: Nallapati et al., 2008 - Models linked documents
- Nubbi: Chang et al., 2009, - Discovers relations between entities in text
- Topic Link LDA: Liu et al, 2009 - Discovers communities of authors from text corpora

Conclusions

- Not surprisingly, additional sources of information helps (with the usual caveats)
- We present a technique to blend two different kinds of information - networks and text together
- The method shows demonstrable improvements across two different domains with both internal and external evaluation.

thanks!