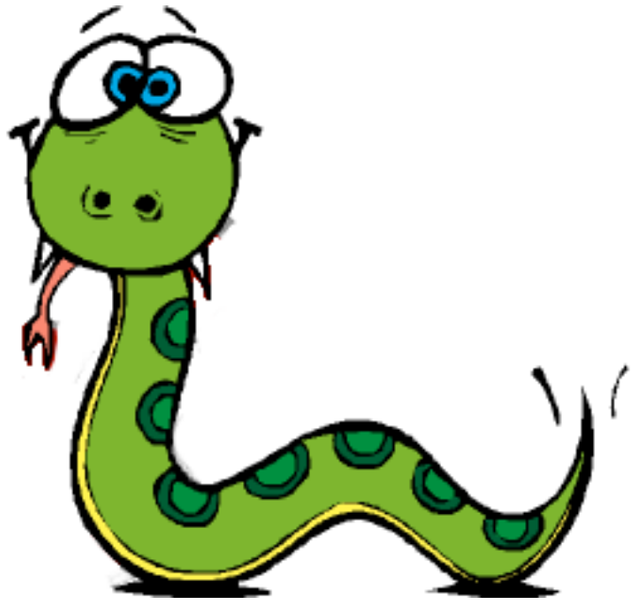
Combining Implicit and Explicit Topic Representations for Result Diversification

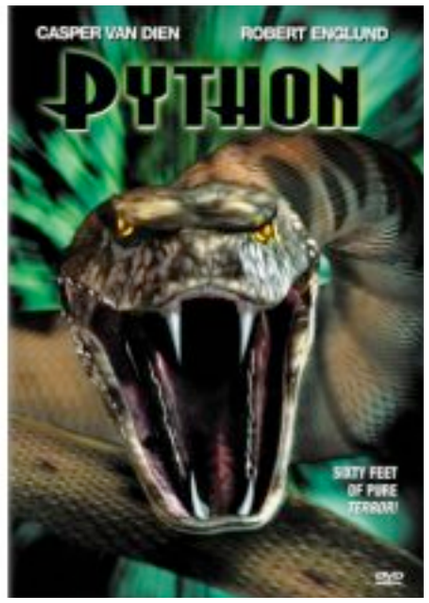
Jiyin He, Vera Hollink, Arjen de Vries
Centrum Wiskunde & Informatica

SIGIR 2012, Portland

Subtopics in result diversification

- Python





Implicit vs. explicit subtopics

- Intent, facets, subqueries, subtopics ...
- Many sources, different representations

External sources

Explicit topic labels

Internal sources

Implicit topic labels

Finding diverse subtopics from multiple sources

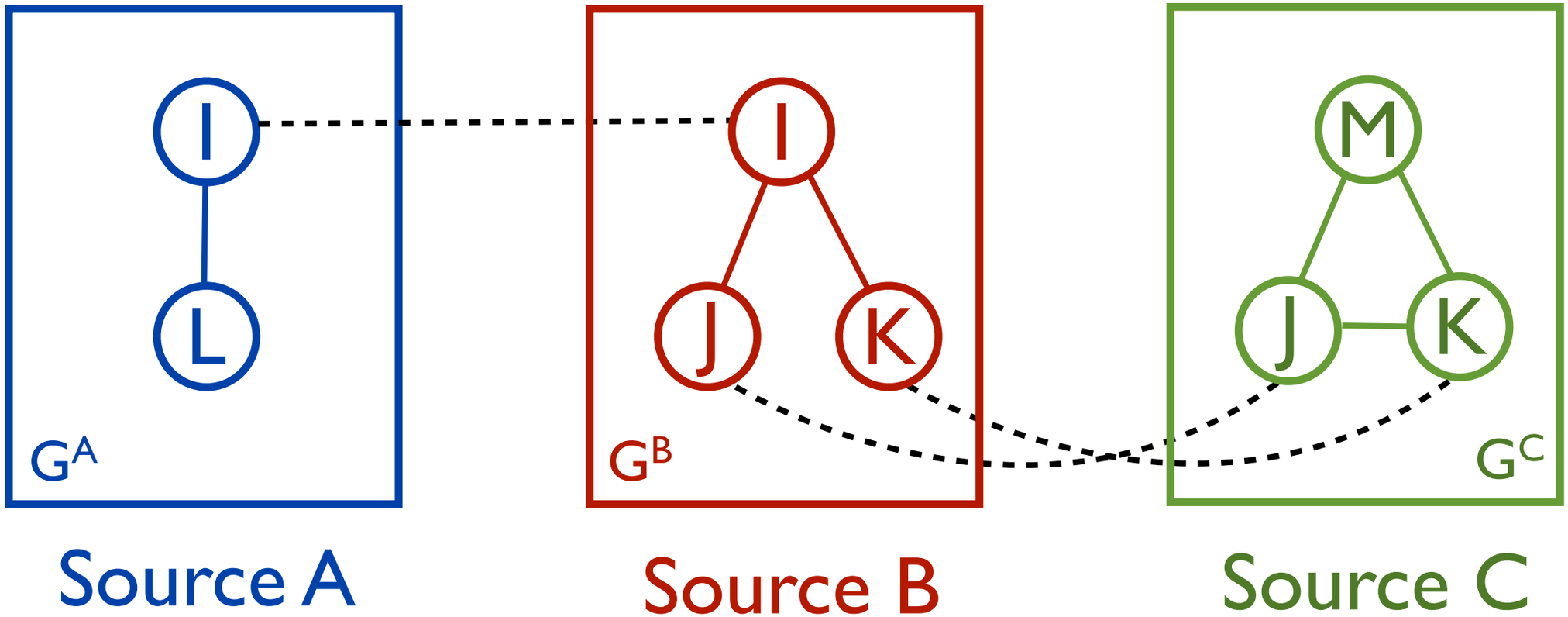
- Objectives
  - Can we make use of information from both implicit and explicit subtopics, and subtopics extracted from multiple sources?
- Potential benefits
  - Better coverage of search requests
  - Better coverage of subtopics of a search request

Finding diverse subtopics from multiple sources

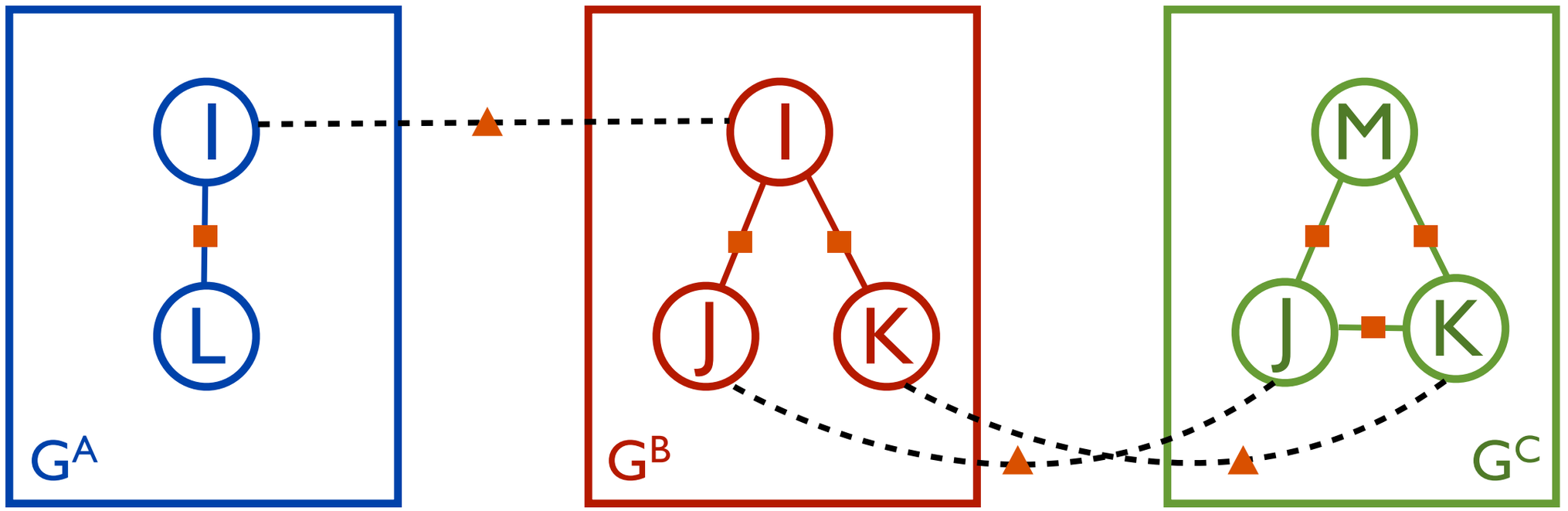
- Issues
  - Redundancy/overlaps of subtopics in different sources
    - Relation among subtopics needs to be modeled
  - Relation between subtopics in different resources may encode different semantic
    - e.g., co-clicks of urls in query logs vs. co-occurrences of anchor texts
  - Matching between different topic representations

Combining explicit subtopics from multiple sources

- A network constructed over subtopics of a query from multiple sources
  - Nodes: subtopics (related topics of the query)
  - Edges: weighted by similarity between subtopics



Random walk over the constructed network



- Two types of transitions:

  Assumption: the more similar two topics are, the more likely a transition can happen.

  ■ Within plane: $p_1^\theta(r_j|r_i) = w(i,j) / \sum_j w(i,j)$

  ▲ Between plane: $p_1^\beta(r_j|r_i) = \begin{cases} 0 \\ \beta_g \end{cases}$

- A one-step transition from i to j:

$$p_1(r_j|r_i) = \begin{cases} p_1^\theta(r_j|r_i) & \text{if } r_i, r_j \in G^g, \\ p_1^\beta(r_t|r_i)p_1^\theta(r_j|r_t) & \text{otherwise,} \end{cases}$$

- A walk of length t:

$$p_t(r_j|r_i) = \sum_k p_1(r_j|r_k)p_{t-1}(r_k|r_i)$$

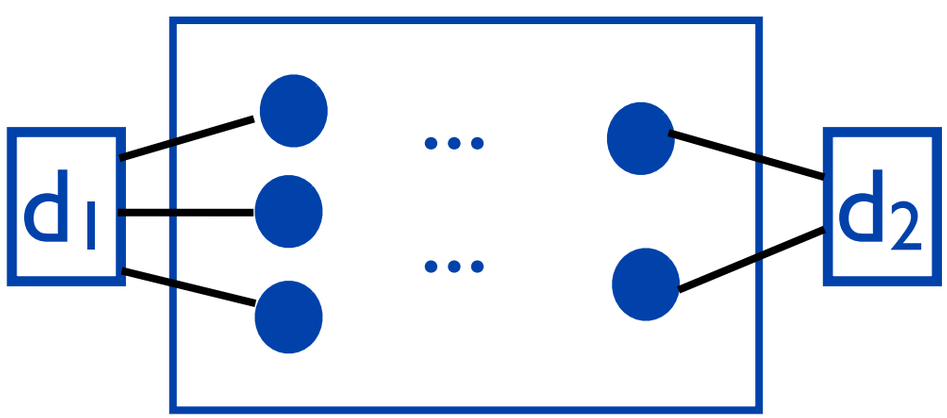
Combining explicit and implicit subtopics

- Regularized pLSA (Cai et al., 2008, Guo et al., 2011)

$$\mathfrak{L} = \mathcal{L} - \gamma \frac{1}{2} \sum_{k} \sum_{i,j} (P(z_k|d_i) - P(z_k|d_j))^2 p(d_i|d_j)$$

- From similarity between subtopics to similarity between documents

$$p(d_i|d_j) = \sum_{k,l} p(d_i|r_k) p_t(r_k|r_l) p(r_l|d_j)$$



Summary

- Random walk on a planed network constructed over (explicit) subtopics from multiple heterogeneous (external) resources
- Using resulting similarity between subtopics to regularize (implicit) topic models constructed (internally) from documents

External sources

Source	Nodes	Edge weights	Data
Click log (G^C)[1]	search queries	#co-clicked documents	MSN query log
Anchor texts(G^A)[2]	anchor texts	#co-occurrence in text passages	Anchor texts from ClueWeb09
Ngrams(G^N)[3]	Web ngrams	#co-occurrence in text passages	Bing Ngram service

[1] Radlinski et al., 2010; Guo et al., 2011; [2] Dang et al., 2010; [2,3] Dang et al., 2011

An example

Sample subtopic	Top 3 related subtopics					
anti-spy	windows defender	0.22	microsoft antispyware	0.12	defender	0.11
microsoft spyware	windows defender	0.22	microsoft antispyware	0.12	defender	0.11
antispyware	windows defender	0.22	microsoft antispyware	0.12	defender	0.11
microsoft beta	windows defender	0.22	microsoft antispyware	0.12	defender	0.11
windows defender	microsoft antispyware	0.12	defender	0.11	antispyware	0.09
space defender 1.0	star defender 4	0.12	star defender 3	0.12	star defender 2	0.12
defender industries	defender industries Inc	0.20	defender	0.11	windows defender	0.04
microsoft beta	windows defender	0.10	microsoft defender	0.05	microsoft s windows defender	0.05
a public defender	public defender	0.11	public defender's office	0.10	office of the public defender	0.10
tri state defender	chicago defender	0.10	the chicago defender	0.10	national legal aid defender association	0.03

A random sample of 5 subtopics related to the query "**defender**" from ***1 source (top)*** vs. ***2 sources (bottom)*** and the top 3 subtopics related to each of the sample subtopics. The scores are the result of a 5-step random walk on the corresponding graphs.

Experiments

- Goals
  - Does regularization with external explicit subtopics help to form better topic models?
  - How do various subtopics from external resources and their combinations compare in terms of diversification performance?
  - Do combinations of subtopics from different external resources achieve better diversification performance than that of single resources?
  - How sensitive is the performance of diversification based on regularized pLSA to the choice of number of topics (K)?

Experiments

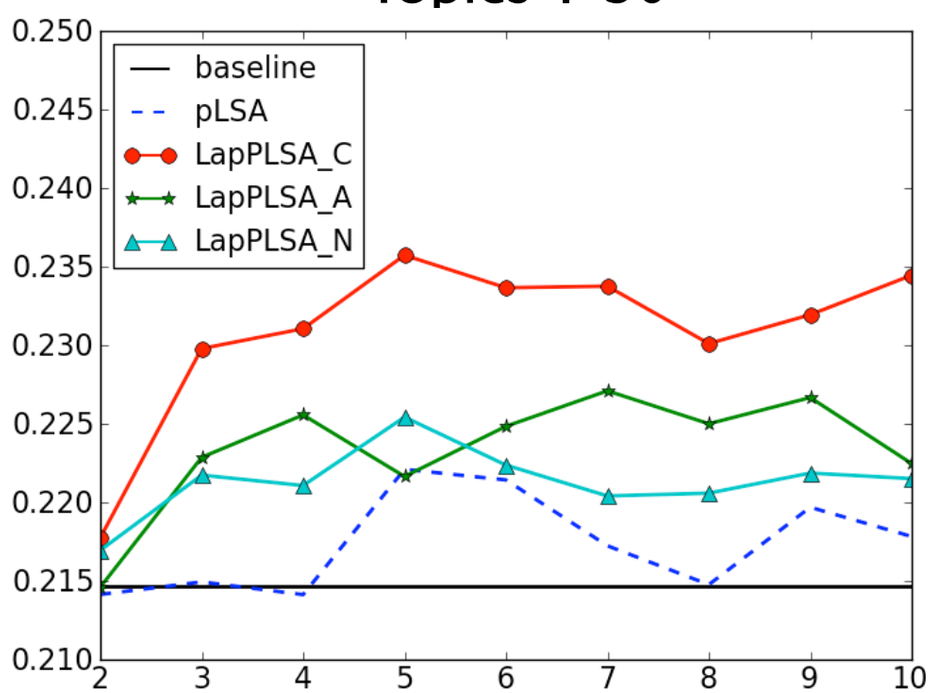
- Data
  - ClueWeb09
  - TREC diversity track topics 2009-2011
    - 2009/10: medium to high frequent queries
    - 2011: more obscure queries
- Diversification methods
  - IA-select*, xQuAD, MMR

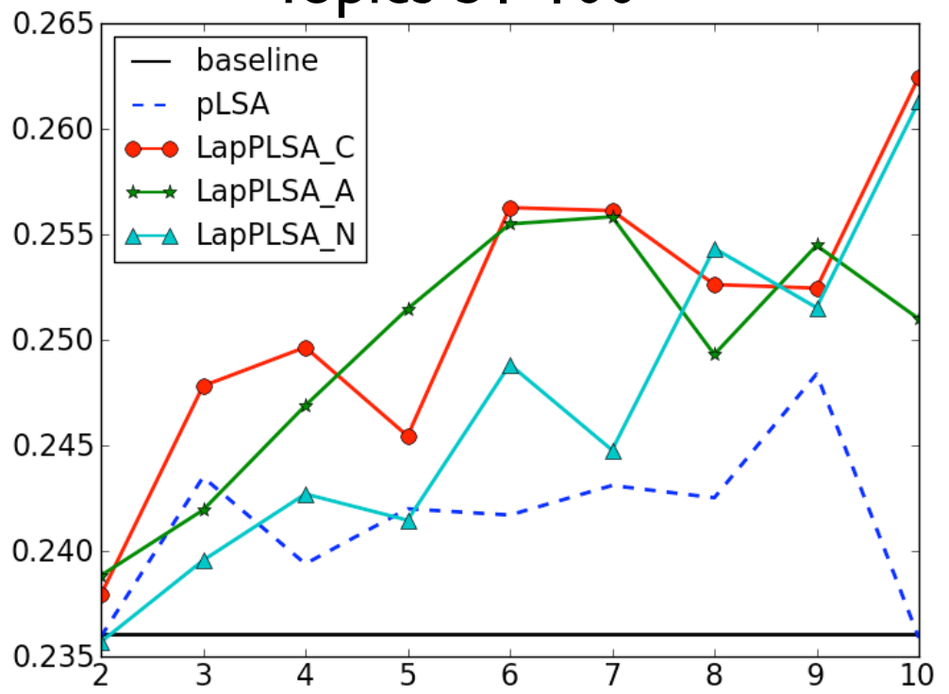
Coverage of the Web resources over the TREC topics

Graph	Coverage		
	1-50	51-100	101-150
G^C	39	37	21
G^A	48	47	25
G^N	48	45	34
G^{CA}	48	48	31
G^{CN}	50	48	39
G^{AN}	50	48	39
G^{CAN}	50	48	39

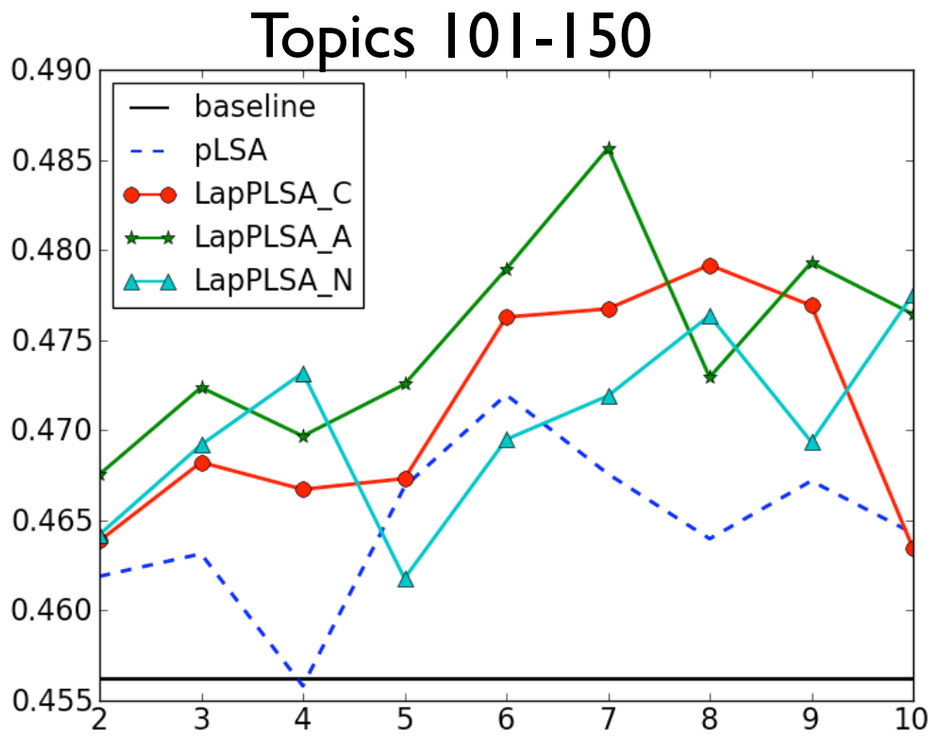
- More sources, higher coverage
- Difference between topic sets
- Implicit subtopics maybe useful when explicit sources does not provide any information

Results

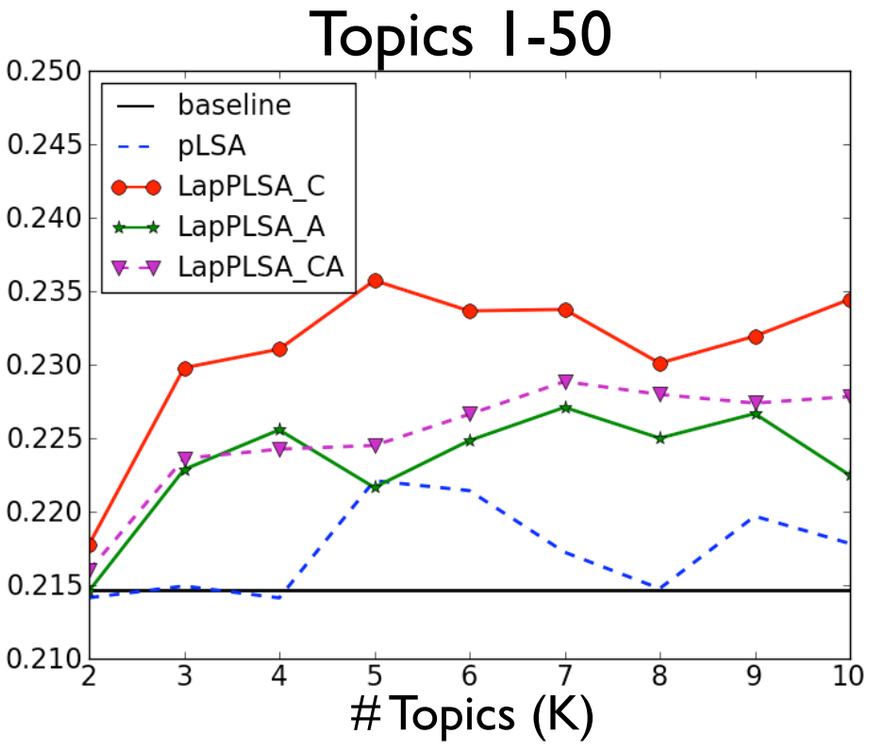


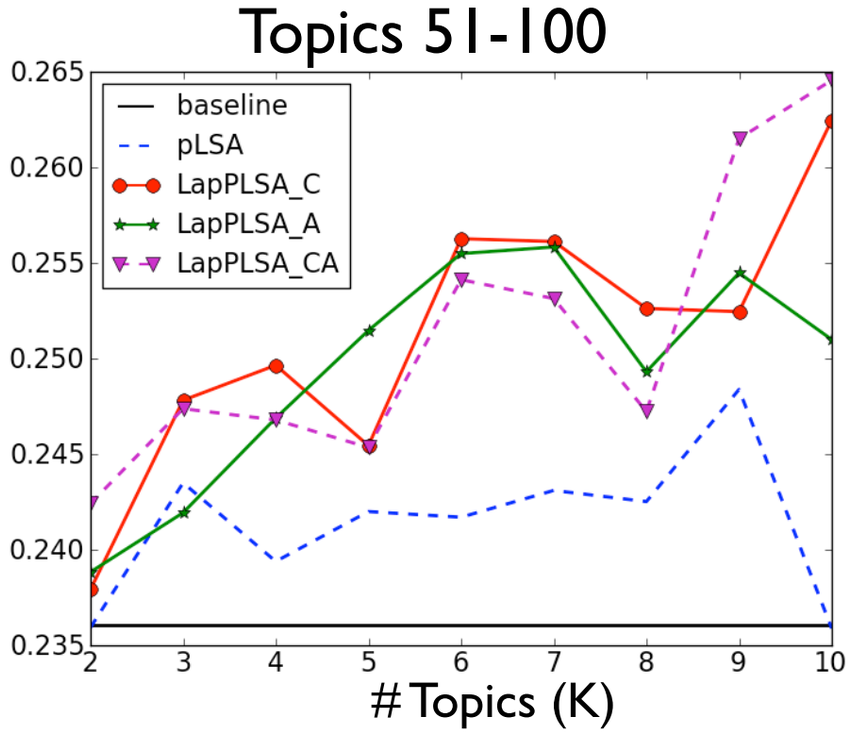


- Main findings (1)
  - Regularization with external subtopics often helps
  - Individual resource is effective in different cases

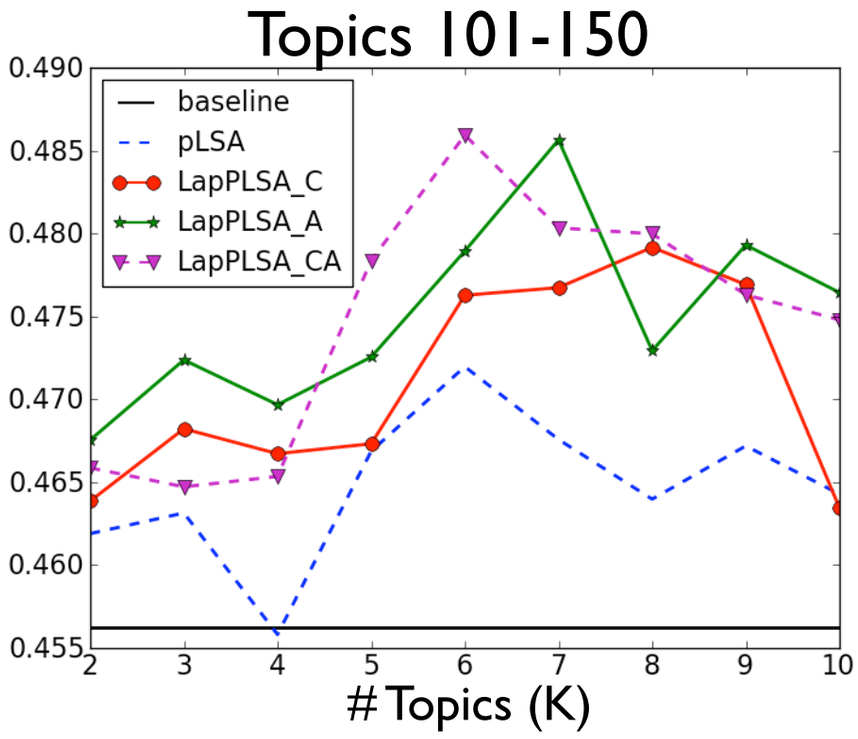


Results

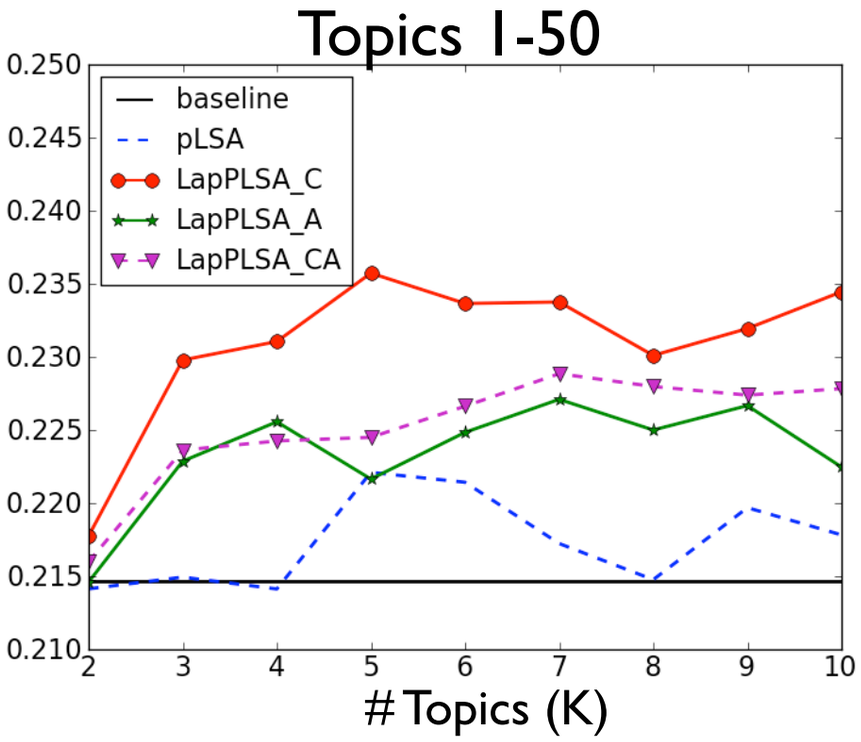


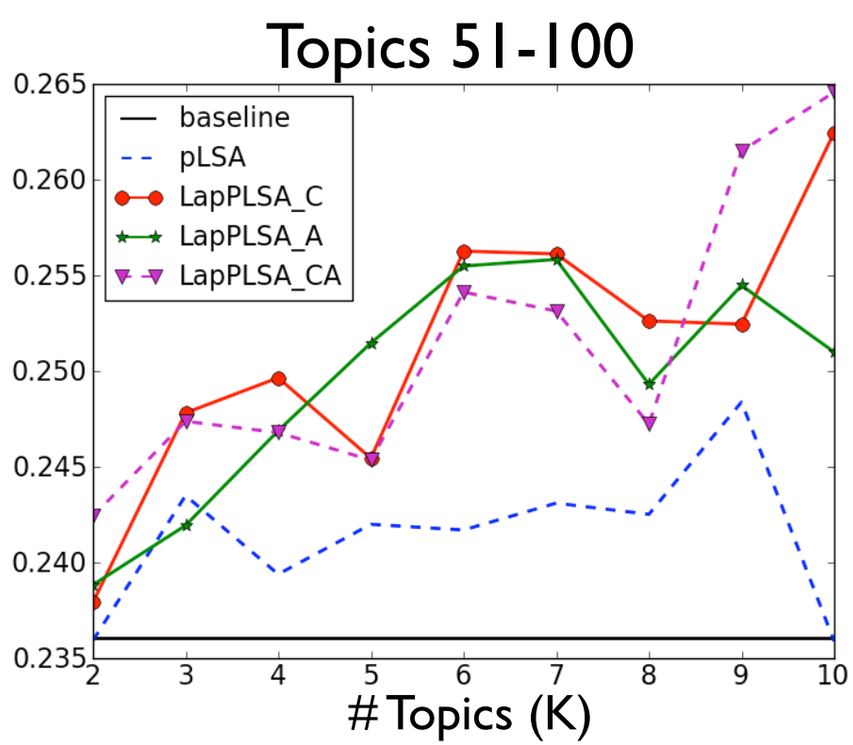


- Main findings (2)
  - Combination of sources does not always lead to optimal results

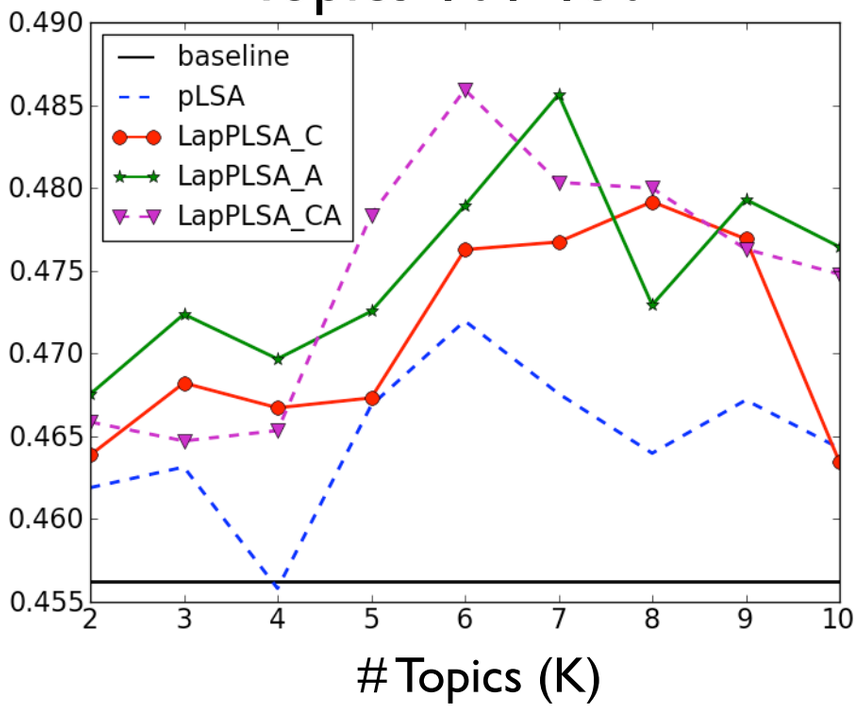


Results





- Main findings (3)
  - Results are sensitive to K
  - A wilcoxon ranksum test confirms that with random K, diversification with
    - regularized pLSA is likely to outperform that of pLSA
    - combined sources is likely to outperform that of the worst individual source



Conclusions

- Combining subtopics of a query from multiple sources and in different representations
- A transparent approach
- Flexible for incorporating different types of subtopics
- Enables intuitive comparisons of resources
- Leads to more robust diversification results
- Source code available online: http://code.google.com/p/mss-rw/