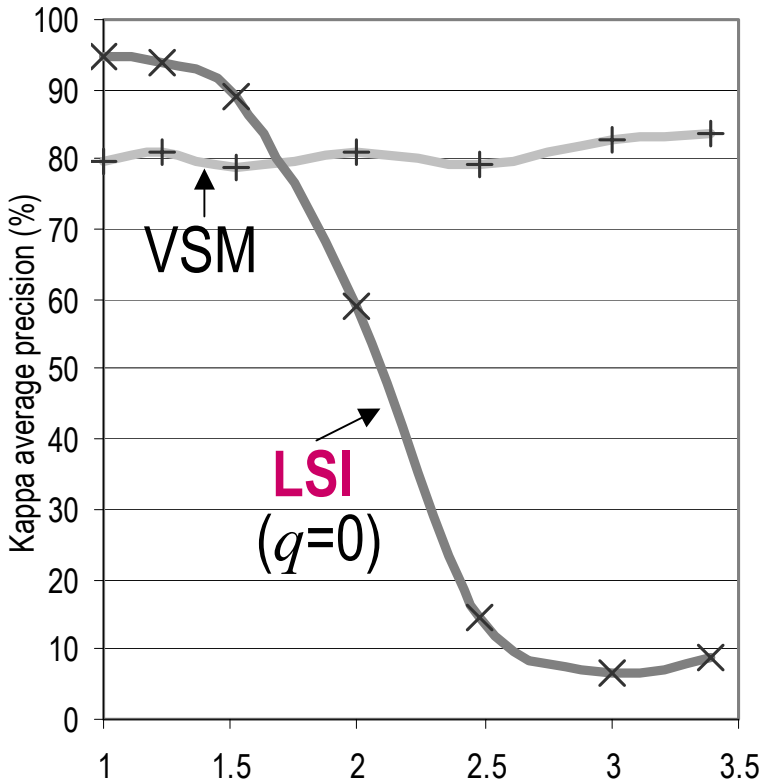
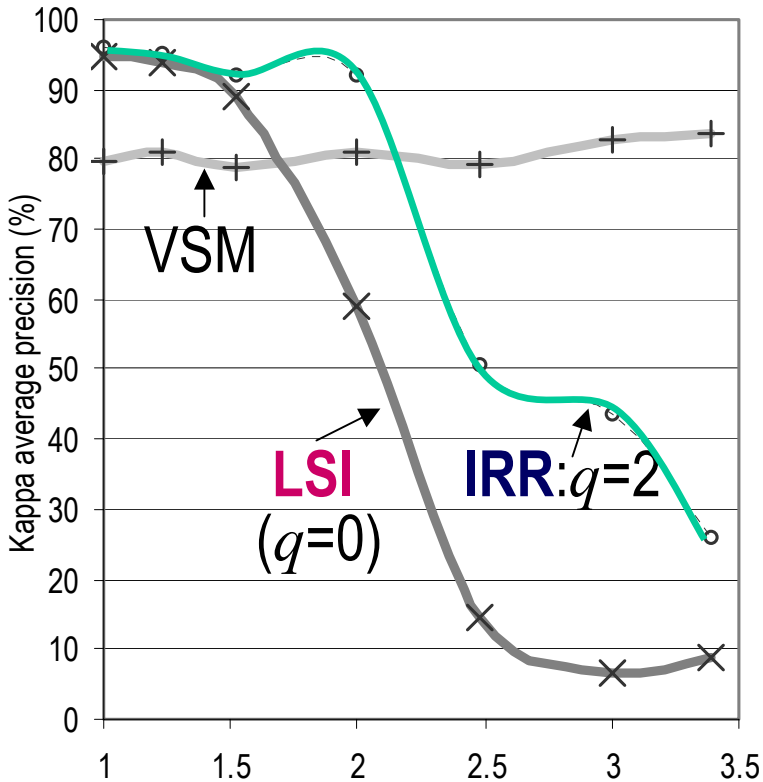
Controlled-distribution results: 2 topics, pair-wise kappa average precision



LSI's performance drops as the topic-doc distributions become **less uniform**.

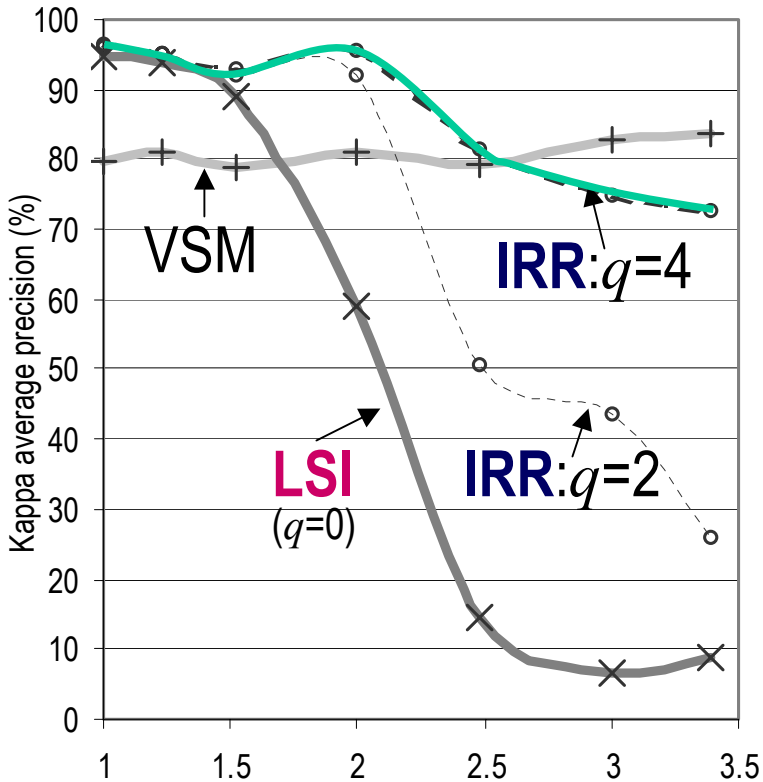
(Average over 10 sets; dim = # of topics.

Two topics, kappa average precision (cont)



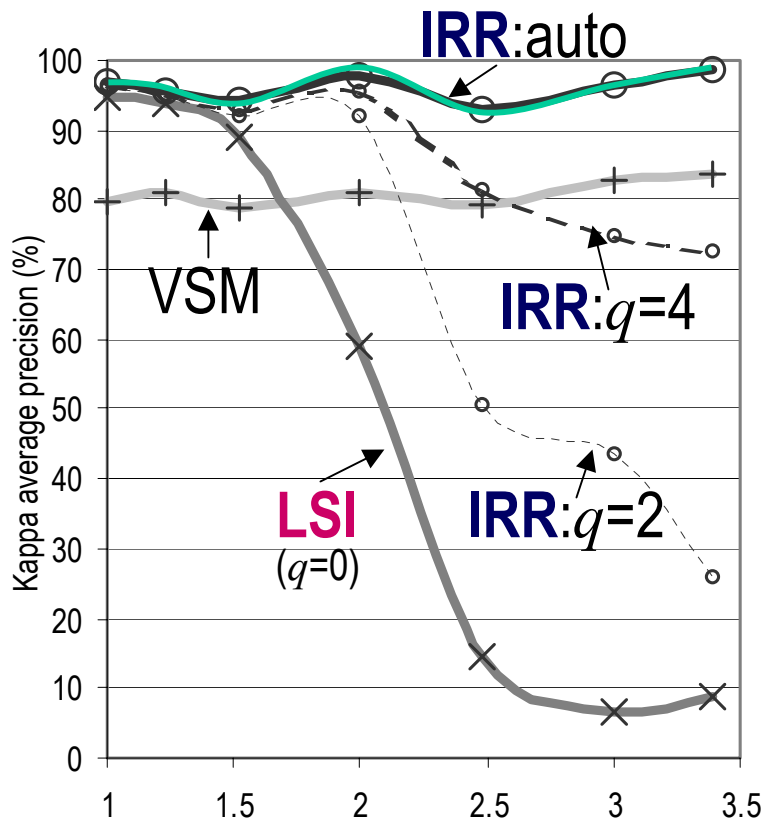
IRR with q=2 compensates for non-uniformity slightly.

Two topics, kappa average precision (cont)



IRR with q=4 compensates for non-uniformity more.

Two topics, kappa average precision (cont)



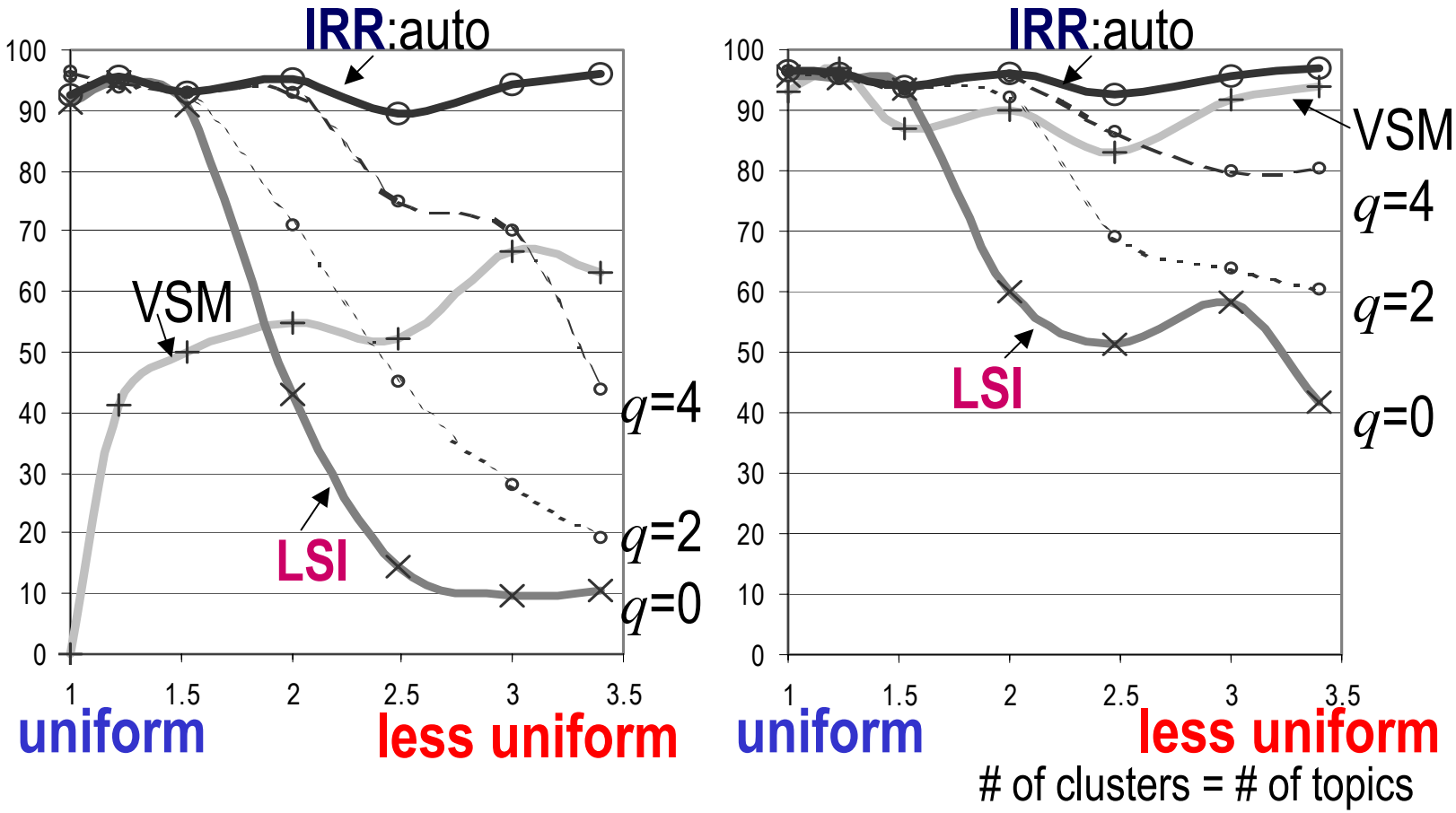
IRR with automatically selected q compensates for any non-uniformity.

Same trends on 3-topic and 5-topic data.

Two topics, document clustering performance

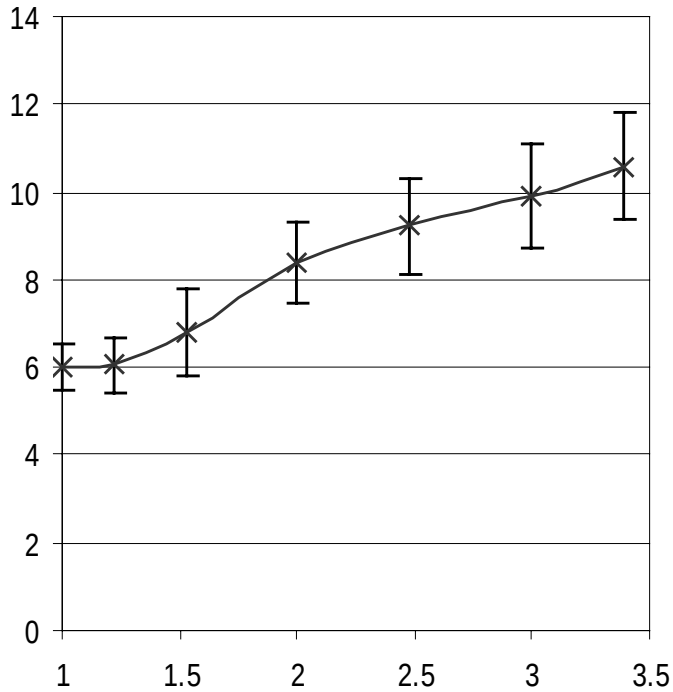
Floor
(worst over 6 clustering algs)

Ceiling
(best over 6 clustering algs)



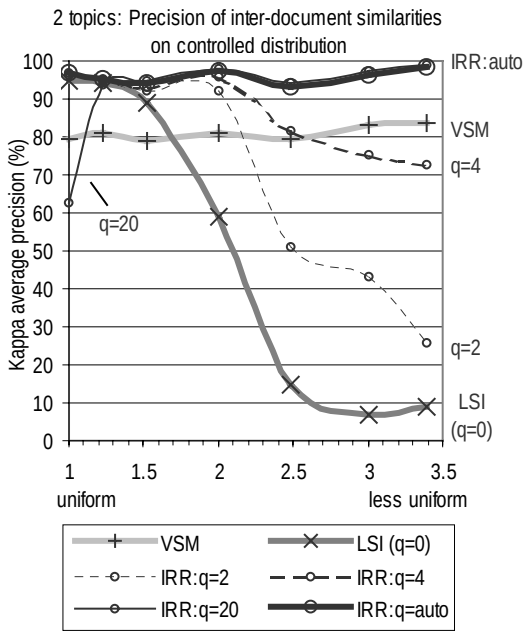
Adjustment for Non-Uniformity

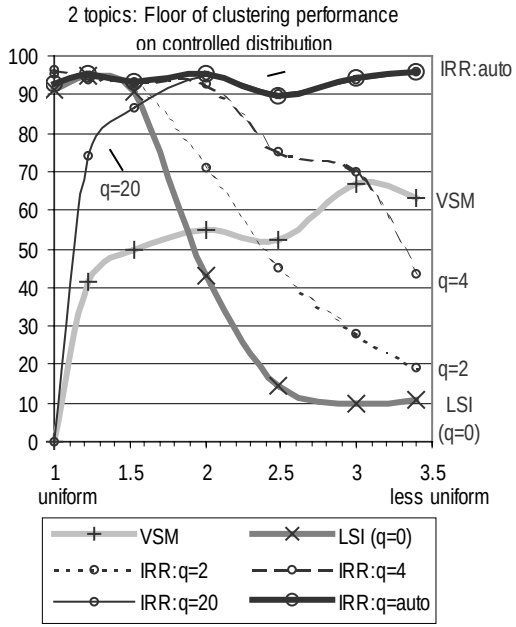
Selected scaling factor, two topics

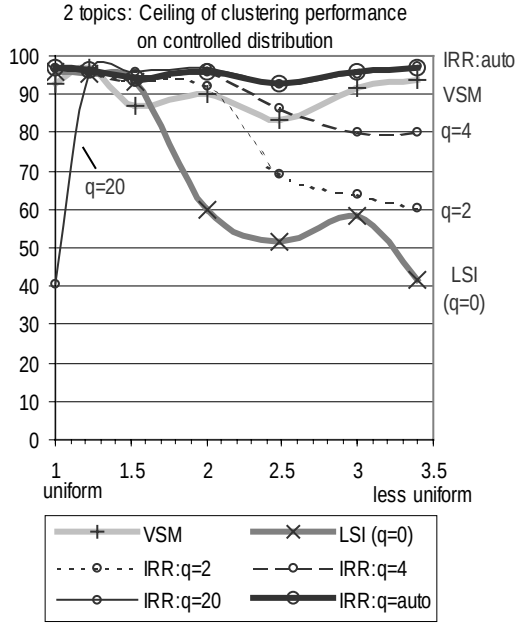


uniform **less uniform**

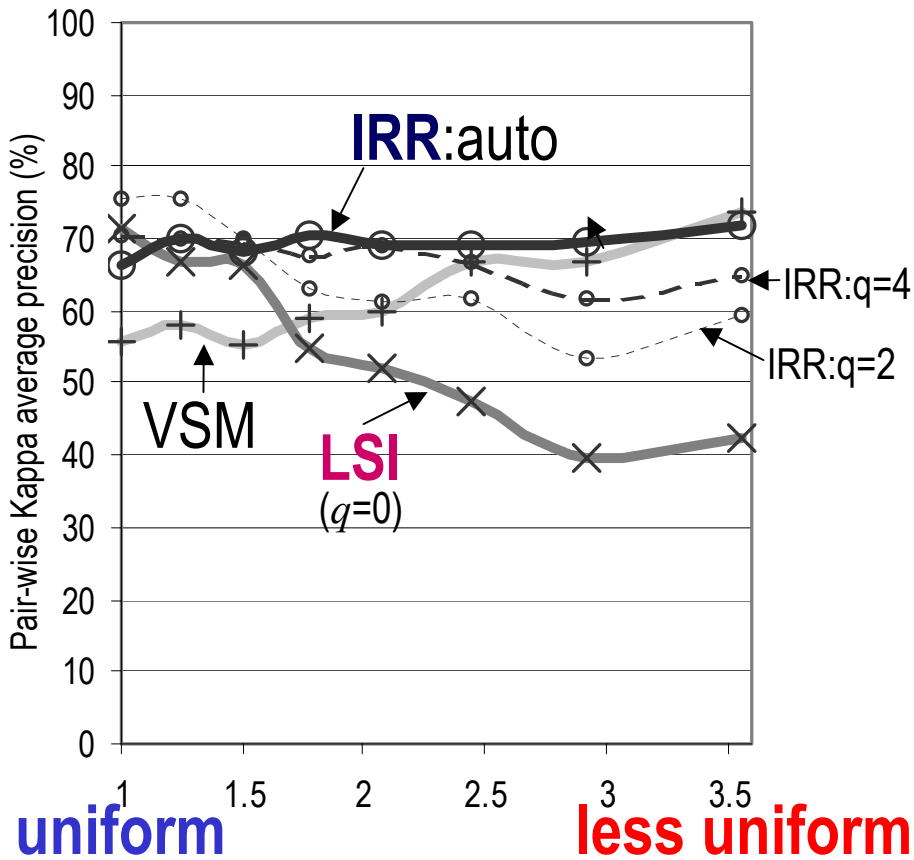
Large q Considered Harmful



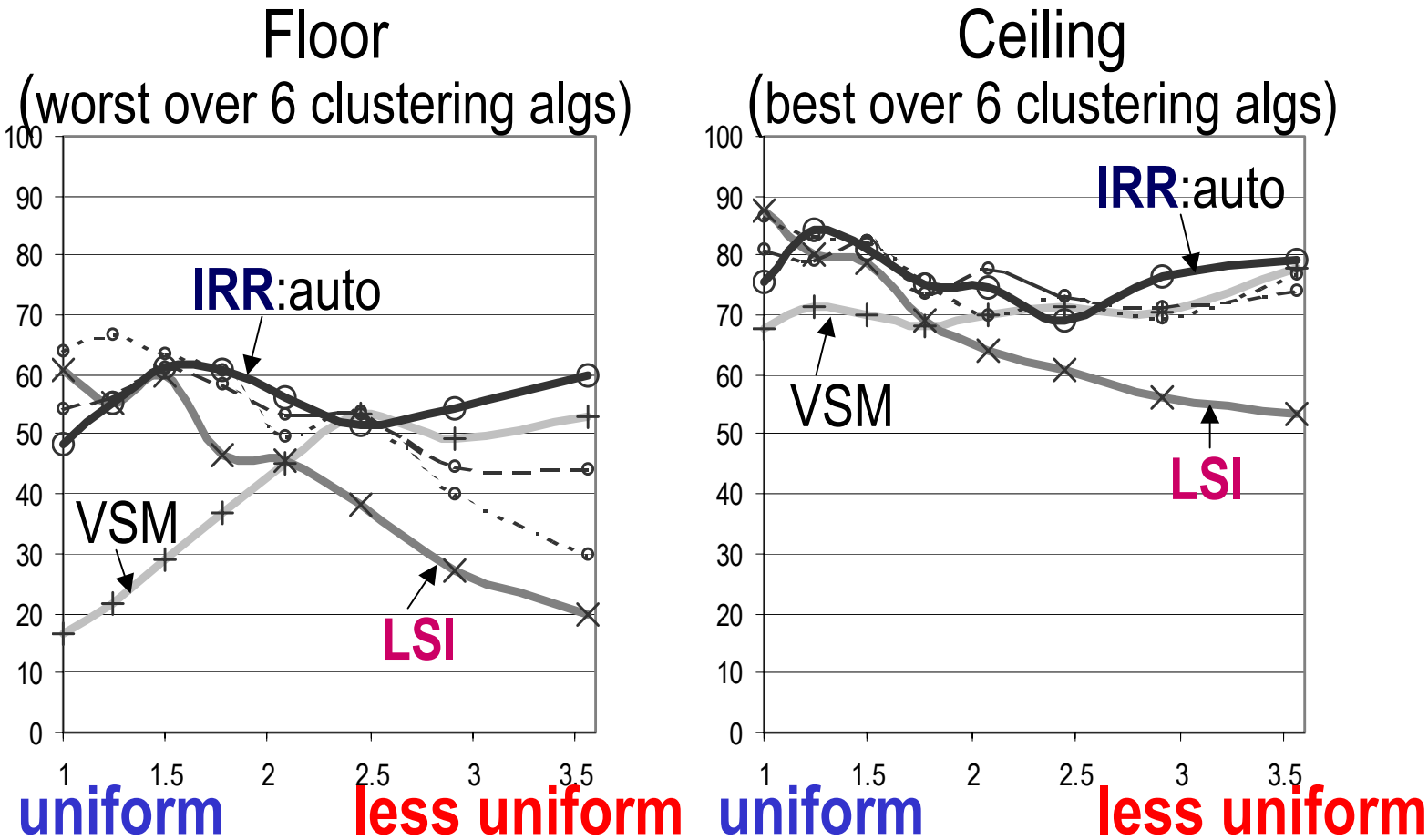




5 topics, pair-wise kappa average precision



Five topics, document clustering performance



# of clusters = # of topics

Conclusions

- A new analysis relating LSI's performance to the uniformity of the underlying topic-document distribution
- A new algorithm --- Iterative Residual Rescaling --- that automatically compensates for non-uniformity
- Experimental results showing IRR's effectiveness in comparison to LSI