Local is Good: A Fast Citation Recommendation Approach
Haofeng Jia and Erik Saule
Department of Computer Science
UNC Charlotte
hjia1@uncc.edu, esaule@uncc.edu

Introduction
Context: Citation recommendation is now a common tool that academics rely on. It can be broadly defined as from who I am or what my interests are to provide relevant papers.

Target Systems: The query of the recommender system is expressed as a set of existing papers. TheAdvisor [3] uses explicit query, also passive recommender systems use past publications as a set of query paper.

Problem: There is a inherent trade-off between obtaining a good recommendation and a fast recommendation. Can we obtain a fast and good recommendation?

Collaborative Filtering [2]
For citation recommendation:
- Build a ratings matrix using the adjacency matrix of the citation graph
  - citing papers correspond to users
  - citations correspond to items.
- Add a pseudo target paper that cites all seed papers
- Computes the cosine similarity of all papers with the target paper
- Identify x peer papers, having the highest similarity to the target paper.
- Each paper is scored by summing the similarity of the peer paper that cites it.

Experimental setting
Dataset: a corpus of 2M Computer Science papers and 12M citations obtained by mapping:
- Microsoft Academic Graph
- CiteSeerX
- DBLP

Queries: 2,500 random hide-10% queries:
- Pick randomly a query paper q with 20 to 200 references and published between 2005 to 2010.
- Remove q and papers published after q
- Use a random 90% of the references of q as seed paper S
- Use the remaining 10% as hidden papers to discover

Hardware and Software:
- C++ code compiled with g++ 4.8.2 with -O3.
- Codes are run on 1 core of an Intel(R) Xeon CPU E-5-2623 @ 3.00GHz processor.

LocRank [ThisPaper]
Inspired from PaperRank:
Consider only the ego network of the seed papers.
- Remove all papers at a distance of 2 or more from a seed paper
- Retain all edges between these papers

Recall@k

Conclusion
LocRank is:
- Faster than both CF and PaperRank
- Higher recall than CF and same as PaperRank
LocRank only needs local context, missing information far from the seed will not impact performance. LocRank can not find loosely connected papers but they are hard to find in general [4]. How to integrate non-citation information?

RunTime
LocRank is:
- 15x faster than PaperRank
- 6x faster than CF

References

Acknowledgments
This work was partially supported by NSF under Grant No. 1652442.