Conclusions

The application and use of large scale bibliographic coupling networks has been hindered by the computational and storage resources required for the creation of these networks. Alternative networks based on direct citations have been used in large scale analysis. The new graph messaging algorithm proposed in this paper provides an opportunity to produce the large scale networks through the application of different selection functions at the level of individual cited references. The experiments with different functions show that references at the lower or higher end of the indegree distribution play a different role in the citation network. Focussing on the bottom results in a network that approximates most of the strong links but is more likely to ignore the weaker ones. Shifting the focus to the other end creates the inverse effect: a higher recall but worse for the identification of strong links. The choice for a particular set of selection function thus depends on the actual objectives for the creation of these BC-networks. If global clustering is the goal then the upper end of the distribution is the right path while if the objective is only to delineate a set of documents closest related to a particular sample the lower end of the indegree is most relevant. Future research will investigate the applicability of this graph based nearest neighbour search algorithm for lexical similarity between scientific documents.

References

1. Johnson, W. B. & Lindenstrauss, J., (1984). Extensions of Lipschitz mappings into a Hilbert space. Contemporary Mathematics. 26, 189–20
2. Karapiperis , D. & Verykios, V.S. (2016). A fast and efficient Hamming LSH-based scheme for accurate linkage. Knowledge and Information Systems, 49 (3), 861-884.
3. Rajaraman, A. & Ullman, J. (2010). "Mining of Massive Datasets, Ch. 3." URL: http://infolab.stanford.edu/~ullman/mmds.html
4. Ravichandran. D., Pantel. P. & Hovy. E. (2005). Randomized algorithms and nlp: using locality sensitive hash function for high speed noun clustering. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. 622-629.
5. Malewicz, G., Austern, M.H., Bik, A.J.C., Dehnert, J.C., Horn, I, Leiser, N. & Czajkowski, G. (2010). Pregel: a system for large-scale graph processing. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 135-146
6. Valiant, L.G. (1990). A bridging model for parallel computation, Communications of the ACM, 33 (8), 103-111.