4. PERSISTENT CONCEPTS

We first ask what key concepts have persisted through the years. We find 24 concepts common to every snapshot, and compute a hierarchical clustering of their semantic vectors derived from the full corpus. The dendogram is cut to produce 10 clusters. The most frequent concept in each cluster is selected as the cluster's representative. This results in 10 concepts: **Algorithm, Data, Mathematical model, Equation, Computer program, Solution, Computer, Language, Design, Mathematics.** The list displays a concentration in mathematical computation, due to the effect of enforcing the requirement that the concepts must have appeared in all temporal snapshots. Newer concepts about computers and computation appearing later are not part of this persistent set.

5. EMERGING CONCEPTS

To follow the emergence of important new concepts, we select the most frequently occurring concepts from each snapshot that are accountable for the top ¼ of all concept mentions. This results in 99 concepts that are rendered in Figure 1 by the first two principal components of their semantic vectors derived from the full corpus.

Figure 1 shows that many key concepts were laid down during the formative years of the discipline. The newest were brought in during 1995-2004: **Support vector machine, Web service, Wireless sensor network,** and **XML**. The newer concepts mostly occupy new regions in the semantic space (lower right corner) far away from the conventional areas in mathematics and programming. An exception is **Support vector machine,** which (colored orange) resides semantically among the older concepts in mathematics and statistics. Figure 1 also shows concentrations in mathematics and statistics, natural/programming languages, networking, etc., with no clean separation. This reflects the continuity between many of the apparently disjoint themes in the discipline. Further analysis on the manifolds formed by the semantic similarities may reveal more detailed structure.

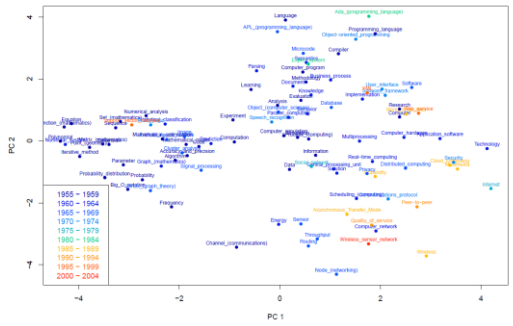

Figure 1: Key concepts colored by first appearance.

6. CHANGING SEMANTIC ASSOCIATION

The next question we ask is in what way concepts have changed in meaning over the years. While a formal definition of a concept is sometimes available in dictionaries (e.g. Wikipedia), its semantic content is often more fluid, and may change from time to time. We can follow such changes by observing the "semantic neighbors" of the concepts of concern in the per-snapshot embedding space. In Figure 2 we show changes (smoothed with a LOESS model) in association strengths of 30 concepts with the concept "**Computer network.**" The characteristic associations in each period are evident -- the rise of associations with Internet related concepts in the early 90's, followed by cloud computing, and later by mobile networks.

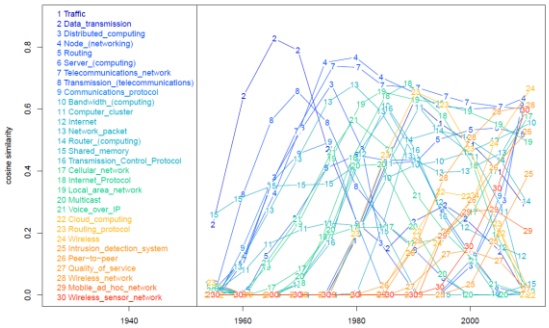

Figure 2: Changing associations with "Computer network."

7. CONCLUSIONS

We analyzed temporal changes in a corpus with a focus on the key concepts and their semantic associations. While we used the literature of computer science research as an example, we believe that the methodology is applicable to other time-stamped corpora.

8. ACKNOWLEDGMENTS

Our thanks to ACM, the curator of the collection of the abstracts, for sharing the corpus with us, and in particular Wayne Graves and Asad Ali for facilitating our access to the data. Our thanks also to Michele Franceschini and Livio Soares for their tools associated with the Concept Insights Service [8] of IBM Watson™ for processing the corpus and the annotations.

9. REFERENCES

[1] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. 2003. A neural probabilistic language model. *J. Machine Learning Research.* 3, 1137-1155.

[2] Cheng, X. and Roth, D. 2013. Relational inference for Wikification. *Proc. of the Conf. on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA, October 18–21, 2013), 1787-1796.

[3] Jatowt, A. and Duh, K. 2014. A framework for analyzing semantic change of words across time. *Proc. of the ACM/IEEE-CS Joint Conf. on Digital Libraries* (London, UK, September 8-12, 2014), 229-238.

[4] Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. 2015. Statistically significant detection of linguistic change, *Proc. of the Int'l World Wide Web Conf.* (Florence, Italy, May 18-22, 2015) , 625-635.

[5] Levy, O. and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. *Proc. of the Annual Conf. on Neural Information Processing Systems* (Montreal, Quebec, Canada, December 8-13, 2014), 2177-2185.

[6] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Proc. of the Annual Conf. on Neural Information Processing Systems* (Lake Tahoe, Nevada, USA, December 5-10, 2013), 3111-3119.

[7] Pennington, J., Socher, R., and Manning, C.D. 2014. Glove: Global vectors for word representation. *Proc. of the Conf. on Empirical Methods in Natural Language Processing* (Doha, Qatar, October 25–29, 2014), 1532-1543.

[8] http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/concept-insights.html.