sition, and the y-axis is for the total execution time by that position. There are five curves in each figure, corresponding to five different frequencies of reporting prominent streaks. For instance, LLPS-1 means that, whenever a new data entry comes, all the prominent streaks so far are reported; LLPS-16 means the prominent streaks are requested at every 16 data entries. As discussed in Section 4, LLPS-1 is identical to NLPS (Algorithm 3), and LLPS-n is identical to LLPS (Algorithm 5), where n is the sequence length when it does not evolve anymore. Figures 5(a) and 5(b) clearly show that the total execution time of LLPS-i increases as the reporting frequency increases (i.e., reporting interval i decreases). Figures 6(a) and 6(b) further show how the total execution time changes along different reporting intervals. We can see that the execution time drops rapidly at the beginning and quickly reaches near-optimal value even when the frequency is still pretty high (e.g., reporting the prominent streaks at every 16 entries.)

6. CONCLUSION

In this paper, we study the problem of discovering prominent streaks in sequence data. A prominent streak is a long consecutive subsequence consisting of only large (small) values. We propose efficient methods based on the concept of local prominent streak (LPS). We prove that prominent streaks are a subset of LPSs and the number of LPSs is less than the length of a data sequence. Our linear LPS-based method guarantees to consider only local prominent streaks, thus achieving significant reduction in candidate streaks. The results of experiments over multiple real datasets verified the effectiveness of the proposed methods.

Acknowledgements: We thank Jun Yang for discussion of the initial ideas of this paper when Chengkai Li and Jun Yang were both visiting HP Labs in Beijing, China in the summer of 2010.

7. REFERENCES

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. *Foundations of Data Organization and Algorithms*, pages 69–84, 1993.

[2] R. Agrawal, K. ip Lin, H. S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *VLDB*, pages 490–501, 1995.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, 1995.

[4] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[5] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator. In *Proceedings of the 17th International Conference on Data Engineering*, pages 421–430, 2001.

[6] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang. Skyline with presorting. In *Proceedings of the International Conference on Data Engineering*, pages 717–719, 2003.

[7] S. Cohen, C. Li, J. Yang, and C. Yu. Computational journalism: A call to arms to database researchers. In *Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR)*, pages 148–151, 2011.

[8] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *SIGMOD*, pages 419–429, 1994.

[9] H.T.Kung, F.Luccio, and F.P.Preparata. On finding the maxima of a set of vectors. *Journal of the ACM*, 22(4):469 – 476, 1975.

[10] B. Jiang and J. Pei. Online interval skyline queries on time series. In *Proceedings of the 25th International Conference on Data Engineering*, pages 1036–1047, 2009.

[11] D. Kossmann, F. Ramsak, and S. Rost. Shooting stars in the sky: an online algorithm for skyline queries. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 275–286, 2002.

[12] T. W. Liao. Clustering of time series data–a survey. *Pattern Recognition*, 38(11):1857 – 1874, 2005.

[13] T. Oates, L. Firoiu, and P. Cohen. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, pages 17–21, 1999.

[14] D. Papadias, Y. Tao, G. Fu, and B. Seeger. Progressive skyline computation in database systems. *ACM Transactions on Database Systems (TODS)*, 30(1):41–82, 2005.

[15] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: the prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, 2004.

[16] J. Pei, Y. Yuan, X. Lin, W. Jin, M. Ester, Q. Liu, W. Wang, Y. Tao, J. X. Yu, and Q. Zhang. Towards multidimensional subspace skyline analysis. *ACM Trans. Database Syst.*, 31(4):1335–1381, 2006.

[17] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[18] Y. Shin and D. Fussell. Parametric kernels for sequence data analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 1047–1052, 2007.

[19] P. Smyth. Clustering sequences with hidden Markov models. In *Advances in neural information processing systems*, 1997.

[20] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology (EDBT)*, pages 1–17, 1996.

[21] K.-L. Tan, P.-K. Eng, and B. C. Ooi. Efficient progressive skyline computation. In *Proceedings of the 27th International Conference on Very Large Data Bases*, pages 301–310, 2001.

[22] M. Wang and X. Wang. Finding the plateau in an aggregated time series. In *Advances in Web-Age Information Management (WAIM)*, pages 325–336, 2006.

[23] W.-K. Wong. *Data mining for early disease outbreak detection*. PhD thesis, Pittsburgh, PA, USA, 2004.

[24] T. Xia and D. Zhang. Refreshing the sky: the compressed skycube with efficient support for frequent updates. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 491–502, 2006.

[25] X. Yan, J. Han, and R. Afshar. CloSpan: Mining closed sequential patterns in large datasets. In *Proceedings of SIAM International Conference on Data Mining*, pages 166–177, 2003.

[26] B. Yi, H. Jagadish, and C. Faloutsos. Efficient retrieval of similar time sequences under time warping. In *Proceedings of the 14th International Conference on Data Engineering*, pages 201–208, 1998.

[27] M. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1):31–60, 2001.