Distance	Mean Comps	LL	LG	GL	GG
l_2	4006	959	2	2	3043
l_∞	3801	913	2	2	2884
l_1	3053	805	2	2	2244

Table 2: number of mean computations, and crossings for feature "bosnia" with threshold $r = 0.0025$ (simultaneous text arrival)

"crossings," i.e. updates $\mathbf{v}'$ of the mean $\mathbf{v}$ and location of $f(\mathbf{v})$ and $f(\mathbf{v}')$ relative to the surface $f(\mathbf{x}) = r$. The number of "crossings" is reported in the last four columns of the table. For example, the number of updates so that $f(\mathbf{v}) < r$ and $f(\mathbf{v}') < r$ is reported in column "LL" of Table 2.

Throughout the rest of the section we assume arrival of a new text at one node only at any given time. Under the assumption we obtain 4890 mean computations (see Table 4). An application of formula (4.14) yields 68460 messages. In both cases the required number of messages is significantly lower than the required number of messages reported in ([7], Figure 8).

We now focus on application Algorithm 4.1 equipped with three different norms. At time t_0 the four dimensional vectors

$$\mathbf{v}_i(t_0) = \begin{bmatrix} x_{11}(\mathbf{T}_i) \\ x_{12}(\mathbf{T}_i) \\ x_{21}(\mathbf{T}_i) \\ x_{22}(\mathbf{T}_i) \end{bmatrix}, \; i = 1, \ldots, m$$

the mean $\mathbf{v}(t_0)$, and the local constraint $\delta = \text{dist}(\mathbf{v}(t_0), \mathbf{Z}_{f-r})$ are computed (here $\mathbf{Z}_{f-r}$ is the zero set of the function $f(\mathbf{v}) - r$, i.e. $\mathbf{Z}_{f-r} = \{\mathbf{v} \; : \; f(\mathbf{v}) = r\}$). The local constraint δ is made available to all the nodes. The vectors $\mathbf{v}_i = \mathbf{v}_i(t_0)$ are remembered at each node.

As a new text arrives at node 1 at time t_1 the vector $\mathbf{v}_1(t_1)$ is computed (while $\mathbf{v}_i(t_1) = \mathbf{v}_i(t_0)$, $i = 2, \ldots, m$ remain unchanged), and inequality $|\mathbf{v}_1(t_1) - \mathbf{v}_1| < \delta$ is checked. If the inequality holds true, no updates of the mean $\mathbf{v}(t_0)$ and the local constraint δ are required, and the procedure is repeated for the nodes $i = 2, \ldots$ (see Algorithm 4.1). If the inequality fails the mean $\mathbf{v}(t_1)$ is updated by $\dfrac{\mathbf{v}_1(t_1) + \mathbf{v}_2(t_1) + \ldots + \mathbf{v}_m(t_1)}{m}$, the new local constraint $\delta = \text{dist}\,(\mathbf{v}(t_1), \mathbf{Z}_{f-r})$ is computed, and made available to each node. This process continues until the end of the stream.

Table 3, Table 4, and Table 5 present the results obtained with l_1, l_2 and l_∞ norms respectively. In all three cases the largest number of mean updates is required for the threshold value 0.00300. The results

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	2122	207	1	1	1912
0.00125	3739	826	1	1	2910
0.00250	3675	2034	12	12	1616
0.00300	**5247**	3812	2	2	1430
0.00600	3255	3050	6	7	191

Table 3: threshold, mean computations, and crossings computed with l_1 norm for feature "bosnia"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	2694	249	1	1	2442
0.00125	5010	1120	1	1	3887
0.00250	4890	2674	12	12	2191
0.00300	**7629**	5681	4	4	1939
0.00600	4289	4003	8	9	268

Table 4: threshold, mean computations, and crossings computed with l_2 norm for feature "bosnia"

show that l_2 is probably not the most convenient norm to be used if the number of mean updates is to be minimized. It appears that computation performed with l_1 norm requires smallest number of mean updates for selected threshold values. The results of the experiments with items "ipo" are collected in Table 6, Table 7, Table 8. The "febru" relevant results are presented in Table 9, Table 10, and Table 11. For all three features and five selected threshold values the l_1 norm requires the smallest number of mean updates.

6 Conclusion

Monitoring streams over distributed systems is an important and challenging problem with a wide range of applications. In this short note we propose a new approach for monitoring an arbitrary threshold functions, and focus on the number of time instances when

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	2368	210	1	1	2156
0.00125	4592	957	1	1	3633
0.00250	4737	2563	14	14	2146
0.00300	**7415**	5517	3	3	1892
0.00600	3954	3679	7	8	260

Table 5: threshold, mean computations, and crossings computed with l_∞ norm for feature "bosnia"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	3114	374	6	6	2727
0.00125	5899	1056	6	6	4830
0.00250	**15331**	4186	26	26	11092
0.00300	13712	8925	43	44	4699
0.00600	2820	2819	0	0	0

Table 6: threshold, mean computations, and crossings computed with l_1 norm for feature 'ipo"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	3987	476	6	6	3498
0.00125	7774	1360	6	6	6401
0.00250	**21109**	6178	26	26	14878
0.00300	19923	13138	48	49	6687
0.00600	3679	3678	0	0	0

Table 7: threshold, mean computations, and crossings computed with l_2 norm for feature 'ipo"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	3703	470	6	6	3220
0.00125	7333	1323	6	6	5997
0.00250	19598	5984	25	25	13563
0.00300	**19653**	13264	49	50	6289
0.00600	3256	3255	0	0	0

Table 8: threshold, mean computations, and crossings computed with l_∞ norm for feature "ipo"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	3595	2041	16	16	1521
0.00125	**4196**	2419	37	37	1702
0.00250	2591	2216	6	6	362
0.00300	1683	1438	5	5	234
0.00600	506	505	0	0	0

Table 9: threshold, mean computations, and crossings computed with l_1 norm for feature "febru"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	4649	2803	19	19	1807
0.00125	**5360**	3164	37	37	2121
0.00250	3140	2698	7	7	427
0.00300	1941	1659	5	5	271
0.00600	547	546	0	0	0

Table 10: threshold, mean computations, and crossings computed with l_2 norm for feature "febru"

Threshold	Mean Comps	LL	LG	GL	GG
0.00025	4426	2644	17	17	1747
0.00125	**5186**	3033	41	41	2070
0.00250	3044	2606	9	9	419
0.00300	1923	1634	5	5	278
0.00600	542	541	0	0	0

Table 11: threshold, mean computations, and crossings computed with l_∞ norm for feature "febru"

the global contingency table should be updated. The obtained preliminary results indicate that experiments with l_1 norm require fewer updates than those with l_∞ or l_2 norm. Identification of norms that are more appropriate for dealing with function f given by (2.3) is a future research direction [3].

Figure 5 inspection reveals that significant fraction of time the mean update causes the local constraint δ to grow. A particular possible communication saving strategy is to keep the coordinator silent if the updated local constraint δ grows. Investigation of various balancing procedures for the coordinator (see, e.g., [8]) may lead to a significant reduction in communication cost. This is an additional research direction that will be pursuit. Realistically verification of inequality $f(\mathbf{x}) - r > 0$ should be conducted with an error margin (i.e., the inequality $f(\mathbf{x}) - r - \epsilon > 0$ should be investigated, see [8]). A possible effect of an error margin on the required communication load is another direction of future research.

While the preliminary results appears to be promising additional research effort is needed to investigate effect of sliding window size, threshold and additional parameters of proposed algorithm performance. Finally we note that proposed approach allows to monitor values of threshold functions defined on vector functions other than mean.

References

[1] M Dilman and D. Raz. Efficient reactive monitoring.

In *INFOCOM '01*, pages 1012–1019. Proceedings of the Twentieth Annual Joint Conference of the IEEE Computer and Communication Societies, 2001.

[2] R.M. Gray. *Entropy and Information Theory*. Springer–Verlag, New York, 1990.

[3] D. Keren, I. Sharfman, A. Schuster, and A. Livne. Shape sensitive geometric monitoring. *IEEE Transactions on Knowledge and Data Engineering*, to appear.

[4] S. Madden and M.J. Franklin. An architecture for queries over streaming sensor data. In *ICDE 02*, page 555, Washington, DC, USA, 2002. IEEE Computer Society.

[5] A. Manjhi, V. Shkapenyuk, K. Dhamdhere, and C. Olston. Finding (recently) frequent items in distributed data streams. In *ICDE 05*, pages 767–778, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[6] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[7] I. Sharfman, A. Schuster, and D. Keren. A geometric approach to monitoring threshold functions over distributed data streams. *ACM Transactions on Database Systems*, 32(4):23:1–23:29, 2007.

[8] I. Sharfman, A. Schuster, and D. Keren. A Geometric Approach to Monitoring Threshold Functions over Distributed Data Streams. In M. May and L. Saitta, editors, *Ubiquitous Knowledge Discovery*, pages 163–186. Springer–Verlag, 2010.

[9] B.-K. Yi, Sidiropoulos N., Johnson T., H.V. Jagadish, C. Faloutsos, and Biliris A. Online datamining for co–evolving time sequences. In *ICDE 00*, page 13, Washington, DC, USA, 2000. IEEE Computer Society.

[10] Y. Zhu and D. Shasha. Statestream: Statistical monitoring of thousands of data streamsin real time. In *VLDB*, pages 358–369, 2002.