Table 3: The running time (sec) and error of PageRank computation by Pregel+ and CompreGel for large graphs.

Network		Pregel+	CompreGel @ε_1	CompreGel @ε_2
Arabic (AR) (1E-9, 1E-10)	Comm. Time	41.213 s	14.952 s	24.407 s
	Total Time	106.226 s	62.256 s	72.614 s
	Comp. Ratio	-	18.49	8.88
UK-2005 (UK) (1E-9, 1E-10)	Comm. Time	83.266 s	27.332 s	37.519 s
	Total Time	202.584 s	103.551 s	117.072 s
	Comp. Ratio	-	17.46	8.95
IT-2004 (IT) (1E-9, 1E-10)	Comm. Time	74.515 s	23.020 s	34.983 s
	Total Time	197.426 s	114.115 s	124.289 s
	Comp. Ratio	-	21.06	10.22
Friendster (FR) (1E-9, 1E-10)	Comm. Time	229.848 s	57.999 s	78.491 s
	Total Time	531.550 s	272.504 s	292.343 s
	Comp. Ratio	-	27.09	11.58
Webbase (WB) (1E-9, 1E-10)	Comm. Time	152.142 s	44.518 s	60.093 s
	Total Time	358.015 s	171.180 s	197.253 s
	Comp. Ratio	-	22.87	10.60

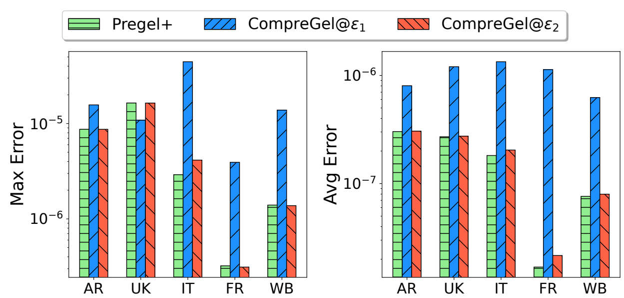


Figure 7: Comparison of Error in PageRank on Large Graphs

A Additional Experimental Results

A.1 Results for PageRank on Large Graphs

Table 3 shows that CompreGel runs faster than Pregel+ on large graphs. For instance, on the Friendster (FR) dataset, with $\varepsilon = 10^{-9}$ and 10^{-10}, the communication time of CompreGel is reduced by 4.0× and 2.9×, respectively. The accuracy metrics are shown in Figure 7 and Figure 8, where we can see that the results of Pregel+ and CompreGel@10^{-10} are almost the same on all datasets, and Kendall's τ and NDCG of CompreGel@10^{-9} are still comparable with Pregel+.

A.2 Results for PageRank with Local Push

Figure 9 shows the results of our local-push implementation for PageRank, where as the iteration number increases, the difference between the communication time of Pregel+ and CompreGel becomes larger. The difference in the running time shows a similar trend but the difference in the slopes is smaller since CompreGel is primarily designed to accelerate communication. For example, on the Webbase dataset, the communication achieves 4.5× speedup while the total running time is reduced by 2.8×.

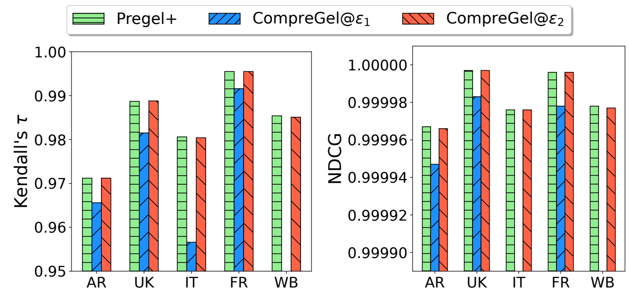
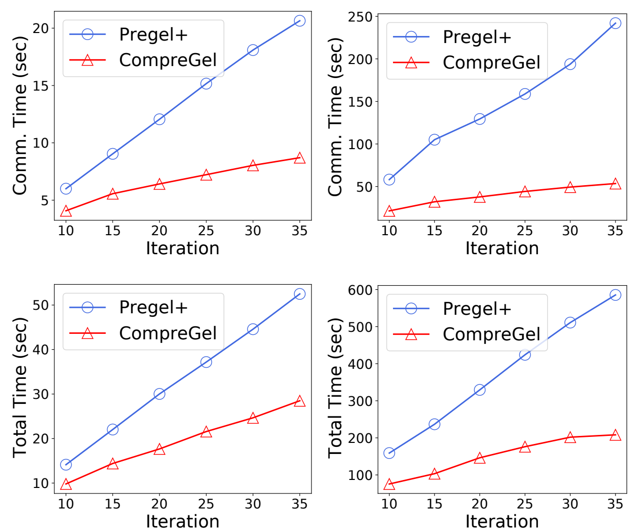


Figure 8: Comparison of Kendall's τ and NDCG on Large Graphs



(a) Indochina (b) Webbase

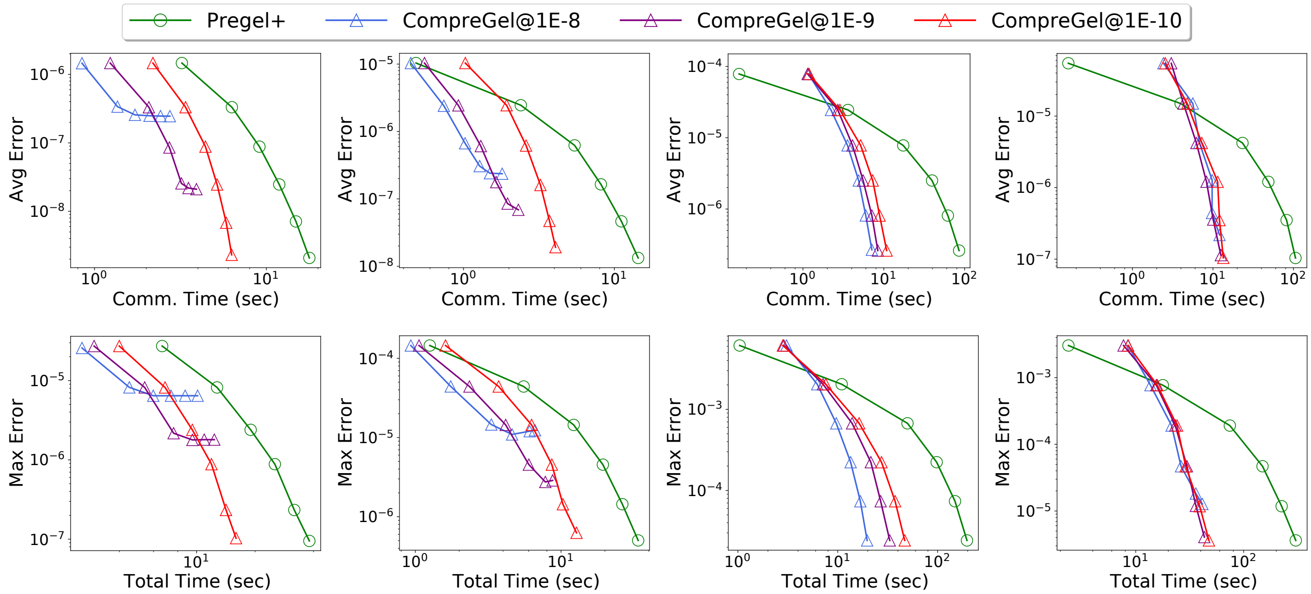
Figure 9: Time (sec) of Pregel+ and CompreGel for PageRank with Local Push

A.3 Time and Accuracy Tradeoff for PPR

We calculate the personalized PageRank algorithm with the local push method in Algorithm 1. Figure 10 shows the time-accuracy tradeoff on two small graphs and two large graphs (we omit results on the other graphs as they are similar). The horizontal axis in each subfigure in the first (resp. second) row is the communication time (resp. the running time), which includes both the communication time and the computation time. The vertical axis in each subfigure in the first (resp. second) row is the average error (resp. the maximum error). The lower the values are, the better the methods perform. In each subfigure, every line has six points (top-down) that correspond to the results when the iteration numbers are 10, 15, 20, 25, 30, and 35. We can see that CompreGel runs much faster than Pregel+ with all three values (10^{-8}, 10^{-9}, 10^{-10}) of the absolute error bound ε for SZ.

A.4 Discussion on Effect of CompreGel for PPR

Difference between the distributions of the values in the PPR vector π and the residue vector r. In Figure 11, we show the



(a) LiveJournal (b) Indochina (c) UK (d) Webbase

Figure 10: Time (sec) and Accuracy Tradeoff of Pregel+ and CompreGel on Various Datasets and Error Parameters (PPR)

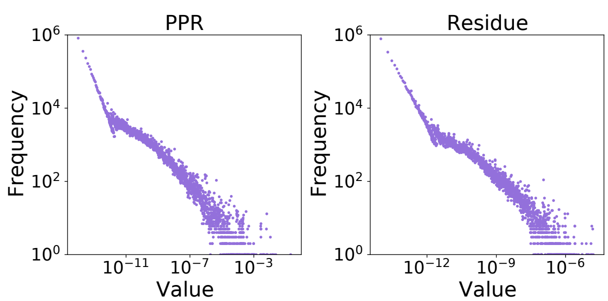
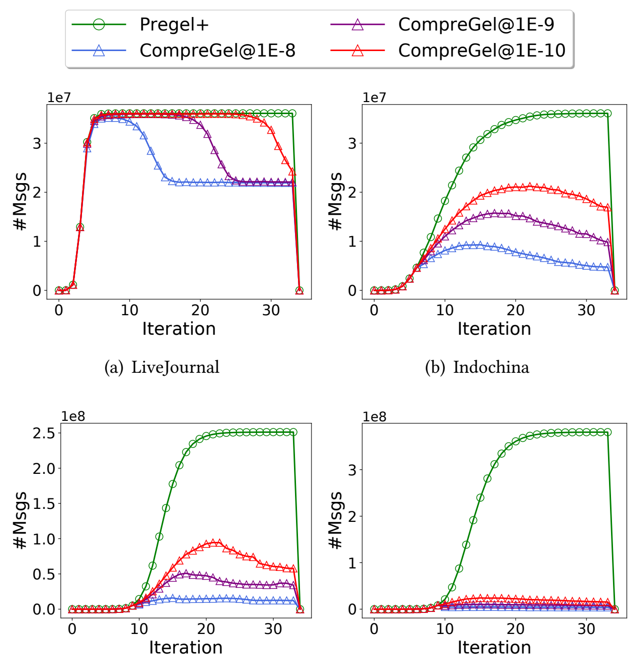


Figure 11: Distributions of Values in PPR Vector and Residue Vector

values in the PPR vector π ranging from 10^{-13} to 10^{-3}, and the values in the residue vector r ranging from 10^{-13} to 10^{-6}. Clearly, the variance of values in π is much larger than the value variance in r. In our experiments, the compression ratio reaches 200 for PPR with local push since the messages are computed from π (recall Line 7 of Algorithm 1), while recall that the compression ratio is only 10 for PageRank with the power method.

Comparison of the number of messages in Pregel+ and CompreGel. As is shown in Figure 12, taking the two small graphs and two large graphs as examples, the number of messages in Pregel+ becomes larger as the number of iterations increases, while in CompreGel, the number of messages shows an initial uprising and subsequently a downturn. For the two large graphs, in the last several iterations, the number of messages in CompreGel is smaller than Pregel+ by an order of magnitude. Note that our theoretical analysis in Section 5 does not consider the additional savings by the if-clause $r(u) > 0$ and assumes a value is transferred anyway.

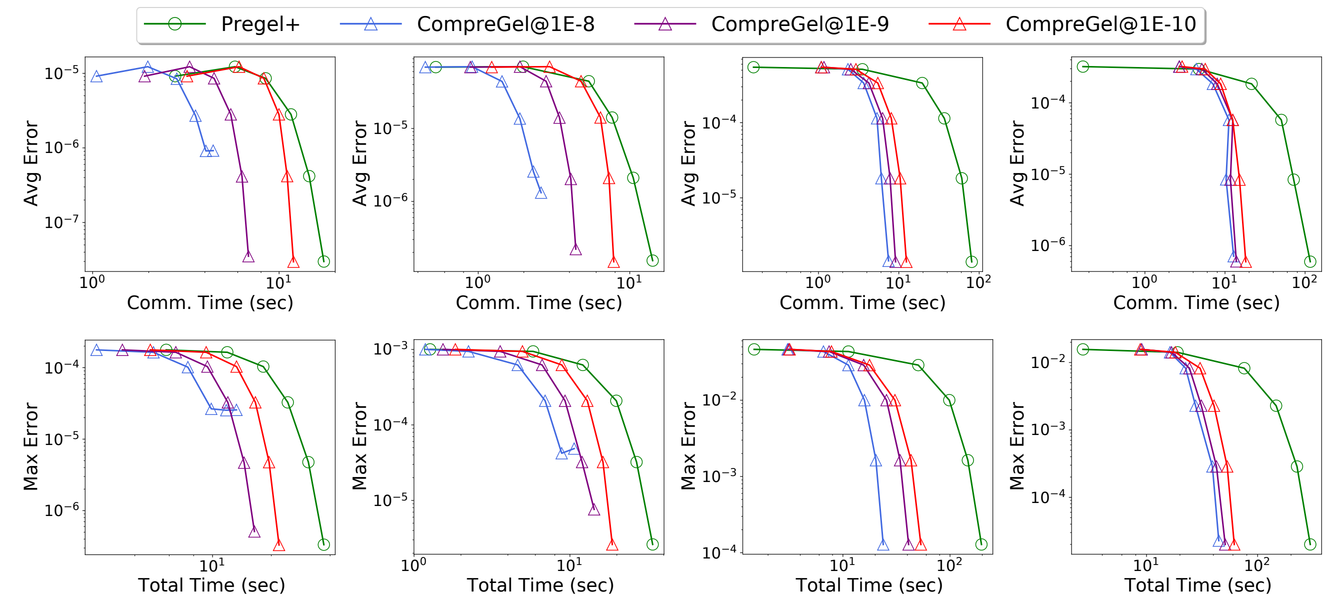


(a) LiveJournal (b) Indochina

(c) UK (d) Webbase

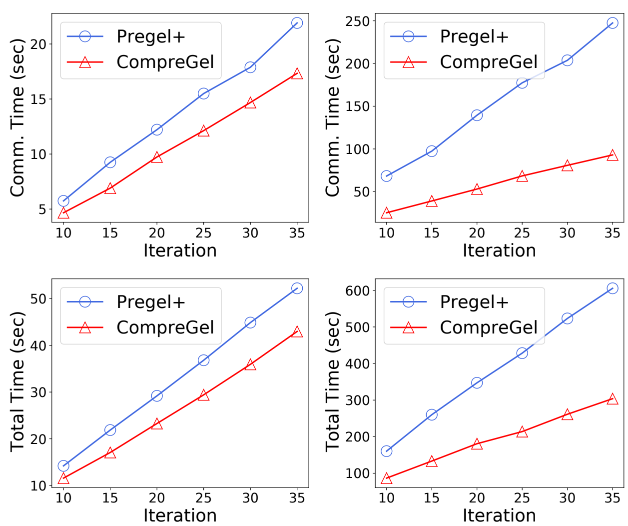
Figure 12: Number of Messages in Each Iteration

Discussion on PPR with power method. To verify the effectiveness of the local-push method, we implement the PPR computation with the power method. Figure 13 shows the results where the speedup ratio by CompreGel becomes much smaller as compared with the local-push method (c.f., Figure 4). Nevertheless, CompreGel



(a) LiveJournal (b) Indochina (c) UK (d) Webbase

Figure 15: Time (sec) and Accuracy Tradeoff of Pregel+ and CompreGel on Various Datasets and Parameters (HKPR)



(a) Indochina (b) Webbase

Figure 13: Time (sec) for PPR with Power Method

still achieves a notable acceleration, and the difference between the communication time of Pregel+ and CompreGel becomes larger as the number of iterations increases. For example, on the Webbase dataset, CompreGel is 2.7× and 2.0× faster than Pregel+ concerning the communication time and the running time, respectively.

Discussion on the compression ratio of CompreGel. Figure 14 reports the compression ratio of CompreGel for the computation of PageRank and PPR on the Arabic dataset. We can see that the compression ratio of CompreGel with the local-push method is much larger than CompreGel with the power method. Comparing

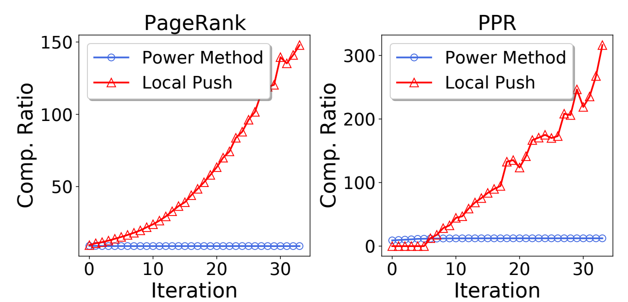
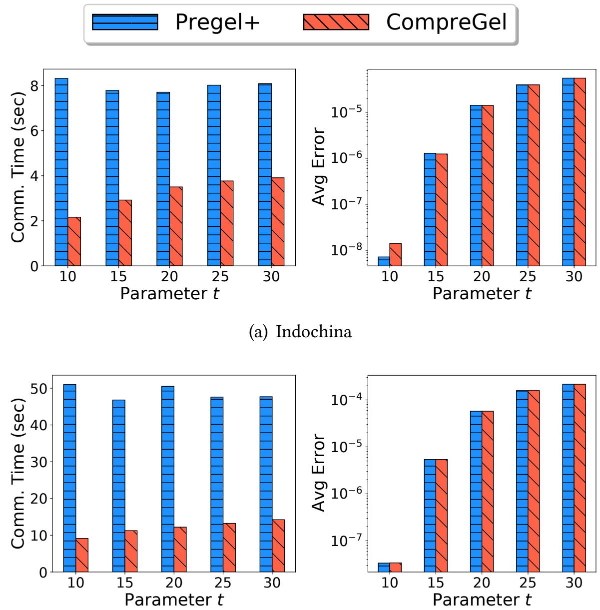


Figure 14: Comparison of Compression Ratio of Power Method and Local Push

the computation of PageRank and PPR implemented with the local-push method, the compression ratio of PPR is obviously larger, since the initial residue values of vertices in PPR are 0 except for the source vertex, while in PageRank the residue values of all vertices are nonzero initially.

A.5 Time and Accuracy Tradeoff for HKPR

Figure 15 shows the time-accuracy tradeoff on two small and two large graphs. The horizontal and vertical axes have the same meanings as in Figure 10 for PPR. The six points on each line in each subfigure indicate the time and error when we are at iterations 10, 15, 20, 25, 30, and 35. The parameter t is set to 20 by default in Figure 15. We can see that similar to the results of PPR, the acceleration for both communication and computation is quite remarkable. For all four datasets, CompreGel with parameter $\varepsilon = 10^{-10}$ is enough to obtain approximately the same accuracy results as with Pregel+.



(a) Indochina

(b) Webbase

Figure 16: Time and Accuracy with Different Parameter t

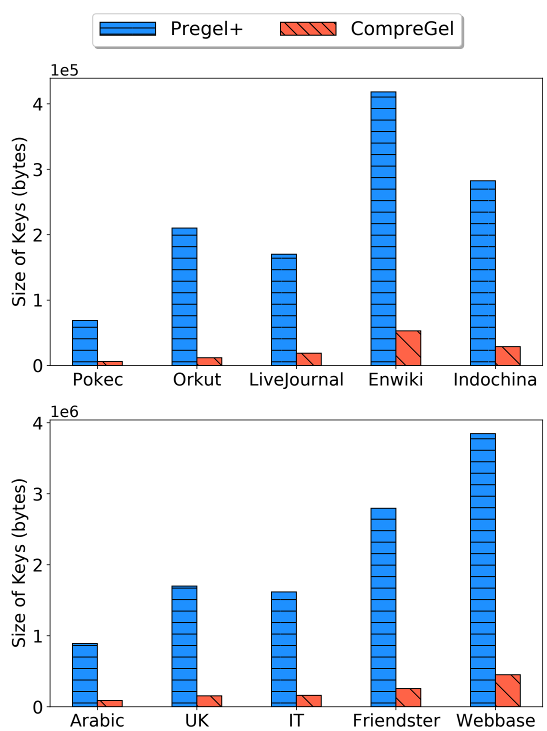
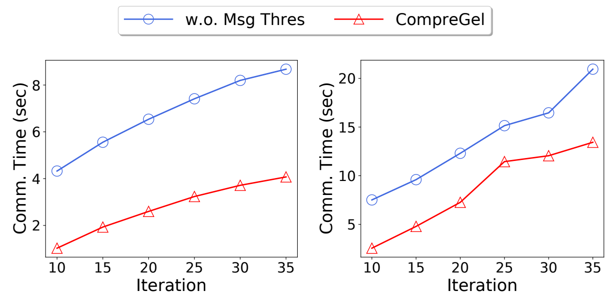


Figure 17: Size of Keys on Different Datasets (PageRank)

A.6 Effect of Parameter t

We also study how the performance of Pregel+ and CompreGel is impacted as the parameter t varies, and the results are shown in Figure 16. We set the error bound ε as 10^{-9} and the number of iterations as 25. We choose the values 10, 15, 20, 25, and 30 for parameter t. The first row in Figure 16 is the communication time and average error on Indochina and the second row is the results on Webbase. We can see that the communication time does not change a lot as t varies, and CompreGel outperforms Pregel+ on both datasets. Meanwhile, Pregel+ and CompreGel have nearly the same average errors for all t, while the average error varies a lot with different t since HKPR converges faster with a smaller t [9].



(a) Indochina

(b) Webbase

Figure 18: Communication Time with and without Message Threshold (PPR)

A.7 Additional Ablation Study

The sizes of keys on different datasets are presented in Figure 17. We can see that the size of the keys is considerably reduced with the lossless bitmap compression. For example, on the Arabic dataset, the sizes of keys with and without compression are 8.8×10^4 bytes and $8.9{\times}10^5$ bytes, respectively; in other words, the bitmap compression provides a compression ratio of more than 10×.

We also conducted an ablation study to verify the effectiveness of the threshold τ_{msg} on the number of messages in a stream, above which message compression is enabled for sending. The results for the personalized PageRank implemented with the local push method are presented in Figure 18, where we use "w.o. Msg Thres" to denote CompreGel without the message threshold. As Figure 18 shows, the communication time is smaller when applying the message threshold. For example, when the number of iterations is 10, the communication time obtained with and without the message threshold is 2.55 s and 7.51 s on Webbase. For PPR, usually, the case when the number of messages is smaller than τ_{msg} only occurs in the first several iterations. Therefore, the difference between the two lines representing the communication time remains fairly stable as the number of iterations increases from 10 upwards.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009