# Expanding Statistical Similarity Based Data Reduction to Capture Diverse Patterns

Dongeun Lee*, Alex Sim†, Jaesik Choi‡, and Kesheng Wu†

*Texas A&M University-Commerce
Commerce, TX 75428, USA
dongeun.lee@tamuc.edu

†Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
{asim,kwu}@lbl.gov

‡Ulsan National Institute of Science and Technology
Ulsan, 44919, Korea
jaesik@unist.ac.kr

We propose a new class of lossy compression based on locally exchangeable measure (LEM) that captures the distribution of repeating data blocks while preserving unique patterns. Our aim is to compress floating-point values produced from large scientific simulations or large experiments, which are known to be hard to compress. Existing techniques are designed to reduce the Euclidean distance between original data and reconstructed data, which heavily rely on continuity present in data. However, there are many applications where such continuity is not essential; instead, the most common feature in the data might be the apparent randomness.

Our technique has been demonstrated to reduce data volume by more than 100-fold on power grid monitoring data[1] where many data blocks can be characterized as following stationary probability distributions [1]. However, we also noted that our implementation named IDEALEM was not effective in compressing the phase angles which were not stationary and drifted over time. To capture data with more diverse patterns including non-stationary values, we propose two techniques to transform non-stationary time series into locally stationary blocks: residual transformation and delta transformation. These new ideas are incorporated into the IDEALEM software.[2] Tests show that the new IDEALEM can reduce the storage requirement by nearly 100 fold in many cases.

| **Data/Compression** | gzip | ZFP | ISABELA | SZ | original [1] | residual | delta |
|---|---|---|---|---|---|---|---|
| A6BUS1C1ANG | 2.42 | 8.76 | 5.36 | 45.84 | 2.04 | 86.89 | 99.19 |
| A6BUS1L1ANG | 2.45 | 8.78 | 5.38 | 159.34 | 1.68 | 84.32 | 38.21 |
| BANK514C1ANG | 2.39 | 8.76 | 5.36 | 49.18 | 2.45 | 96.39 | 99.21 |
| BANK514L1ANG | 2.45 | 8.78 | 5.38 | 148.22 | 1.69 | 85.05 | 62.99 |
| **Overall** | 2.43 | 8.77 | 5.37 | 72.50 | 1.92 | 87.91 | 64.30 |

## References

[1] D. Lee, A. Sim, J. Choi, and K. Wu, "Novel data reduction based on statistical similarity," in *Proc. Int'l Conf. Scient. Stat. Database Manag. (SSDBM '16)*, 2016, pp. 21:1–21:12.

[1] More information about the data could be found at http://powerdata.lbl.gov/.

[2] Code is available at http://datagrid.lbl.gov/idealem/.

2375-0359/17 $31.00 © 2017 IEEE
DOI 10.1109/DCC.2017.77