

Seung-Hee Bae

Department of Computer Science,
Western Michigan University,
Kalamazoo, MI, 49008-5466

Homepage: <http://shbae.cs.wmich.edu/>
E-mail: seung-hee.bae@wmich.edu
Phone: (269) 276-3113

CURRENT POSITION

Assistant Professor, (Aug. 2016 – current)
Computer Science Department, Western Michigan University, Kalamazoo, MI, U.S.A.

EDUCATION

Ph.D. in Computer Science, (Sep. 2005 – Feb. 2012)
Indiana University, Bloomington, IN, U.S.A.
Thesis Topic: *Scalable High Performance Multidimensional Scaling*.
Advisor: Prof. Geoffrey C. Fox

M.S. in Computer Science and Engineering, (Mar. 2002 – Feb. 2004)
Seoul National University, Seoul, Republic of Korea
Thesis Topic: *Adaptive Mutation Rates in the Context of Hybrid Genetic Algorithms*.
Advisor: Prof. Byung-Ro Moon

B.E. in Computer Science and Electrical Engineering, (Mar. 1995 – Feb. 2002)
Handong Global University, Pohang, Republic of Korea

RESEARCH INTERESTS

My research interests cover *parallel and distributed computing, large-scale data mining and machine learning, big data systems and applications, and optimization algorithms*. In specific, I would like to work on scalable parallel data mining algorithms which facilitate computational and scientific applications by dealing with large-scale data via employing efficient parallel computing strategies to them via appropriate parallel runtime models, such as MPI, GraphLab, and (iterative) MapReduce, on various parallel and distributed environments, i.e. multicore cluster systems and cloud systems. Currently, I am working on scalable parallel community detection algorithms for analyzing large-scale graph datasets. I have developed a distributed parallel community detection algorithm, called GossipMap, and it provides high quality outputs and shows high parallel efficiency and scalable to web-scale billion-edge graphs. In the near future, I would like to research on scalable and efficient parallel community detection algorithms for time-variant graph data analysis, which extends my recent research efforts, as well as other large-scale data analysis.

PROFESSIONAL EXPERIENCES and SERVICES

- **External Reviewer** (2016) IEEE Transactions on Big Data (TBD), IEEE
- **External Reviewer** (2016) Pattern Recognition, Elsevier
- **Postdoctoral Research Associate** (Sept. 2012 – Aug. 2016) University of Washington
 - Advisor: Dr. Bill Howe.
 - Scalable Community Detection Algorithms.
 - Sparse Matrix Operation on Big Data Systems.
- **Program Committee** (2015) Exploring Graphs At Scale (EGAS) Workshop at IEEE Vis 2015
- **External Reviewer** (2015) Data Mining and Knowledge Discovery, Springer
- **External Reviewer** (2012) Concurrency and Computation: Practice and Experience
- **External Reviewer** (2012) Journal of Parallel and Distributed Computing, Elsevier

- **External Reviewer** (2010) Computational Statistics
- **Part-time Research Staff** (Mar. 2012 – Jul. 2012) Indiana University, Bloomington
 - Supervisor: Prof. Geoffrey Fox.
 - Deterministic Annealing Multidimensional Scaling (DA-MDS)
 - MDS via adaptive Interpolation for dealing with huge amount of data
- **Research Assistant** (Jan. 2007 – Feb. 2012) Indiana University, Bloomington
 - Advisor: Prof. Geoffrey Fox.
 - Deterministic Annealing Multidimensional Scaling (DA-MDS)
 - MDS via Interpolation for dealing with huge amount of data
 - Parallel MDS for large high-dimensional data using Thread and MPI.
 - GIS data clustering with Indiana Census Data and visualize the clustering results on Google Earth and Microsoft's Virtual Earth.
 - Benchmarking Message passing overhead on multicore machines
- **Research Assistant** (Jan. 2006 – Dec. 2006) Indiana University, Bloomington
 - Advisor: Prof. Sun Kim.
 - Transcription Factor Binding Site Discovery in Human Genome using a Discriminative Pattern Analysis
- **Research Assistant** (Mar. 2002 – Feb 2004) Seoul National University
 - Advisor: Prof. Byung-Ro Moon.
 - Adaptive mutation rates in the Context of Hybrid Genetic Algorithms
- **Teaching Assistant** (Mar. 2002 – Jun. 2002) Seoul National University
 - Data Structure class
 - Teaching undergraduate student elementary Java language in Data Structure class
 - Maintaining the homework, exam, and grading for Data Structure class
- **Research Assistant** (Jan. 2001 – Dec. 2001) Handong Global University
 - Advisor: Prof. Youngsup Kim.
 - GIS data representation based on Scalable Vector Graphics (SVG).
- **Teaching Assistant** (Jan. 1998 – Feb. 1998) Handong Global University
 - Logic Design and Lab class
- **Internship Employee** (Jul. 1997) IBM Korea

RESEARCH PROJECTS

- **Parallel Sparse Matrix Operations** (2015 – Present)
 - Research about sparse matrix operations on various big data systems as well as linear algebra libraries.
- **Scalable RelaxMap** (2012 – Present)
 - Infomap is one of the best community detection algorithms, but it cannot scale to large-scale graph datasets. I have developed parallel generalization of Infomap, called RelaxMap; it achieves high parallel efficiency in shared-memory environment via OpenMP.
 - Proposed a prioritized method of the RelaxMap, which is much faster than non-prioritized version.
 - In order to handle web-scale graphs with billions of edges, I have developed a novel and distributed parallel algorithm called GossipMap, which is a distributed community detection algorithm and able to handle billion-edge graphs with comparable quality to Infomap.
- **DA-MDS** (2009 – 2012)
 - I applied Deterministic Annealing (DA) optimization approach to MDS algorithms, for the purpose of avoiding well-known local optima problem in conventional MDS solutions, and originated DA-MDS algorithm. DA-MDS improves mapping quality in terms of the well-known objective function, called STRESS, compared to Expectation-Maximization (EM)-like majorization approach named SMACOF. In addition, DA-MDS turns out more efficient than

SMACOF algorithm for larger data sets. DA-MDS shows better performance in terms of mapping quality, and I parallelized the DA-MDS by using MPI.

- **MDS Interpolation and High-performance Parallel MDS for Large High-dimensional Data Visualization** (2008 – 2012)
 - Developed high-performance parallel MDS, which is a data-intensive application as well as computation-intensive application for large high-dimensional scientific data, by using Thread and MPI.
 - Since the time complexity of MDS is quadratic, it is still difficult and suffers from long running time to deal with millions of data, which is a normal data size nowadays, even if we parallelized the algorithm. Therefore, I invented a MDS interpolation algorithm to reduce time complexity based on majorizing method used in SMACOF.
- **Parallel Runtime Comparison** (2010)
 - Analyzed performance of different parallel runtimes, including Microsoft's Concurrency and Coordination Runtime (CCR), Task Parallel Library (TPL), and MPI.NET through various parallel applications.
- **GIS data clustering and Visualization** (2007)
 - Analyzed DA clustering results with Indiana Census Data, and visualized them on Google Earth and Microsoft's Virtual Earth.
- **Message Passing Overhead Benchmark** (2007)
 - Compared message passing overhead of Microsoft's Concurrency and Coordination Runtime (CCR) with various MPI parallel runtimes, for instance, MPICH2, mpiJava, and MPJ Express (MPJE), upon multicore systems based on simple message passing patterns.
- **dPattern: Transcription Factor Binding Site (TFBS) Discovery** (2006)
 - Devised an algorithm to discover TFBS in Human Genome by a discriminative pattern analysis. The given profile model is improved by a mixture model approach which combines the original profile model with the complimentary model which is a k-order Markov model with respect to the predefined alignment result in Genome Browser of UCSC.
- **Adaptive Mutation Rates for Hybrid GA** (2002 – 2004)
 - Investigated an adaptive method for mutation rates of Hybrid Genetic Algorithm (GA). Since hybrid GA combines normal GA procedure with strong local optimum algorithm, the level of perturbation decreases considerably. Thus, we tried to increase mutation rate over the generations to maintain the perturbation rate, in contrast to the conventions, and the experimental results supported the proposed strategy.
- **GIS Data Representation based on SVG** (2001)
 - Transformed eXtensible Markup Language (XML) type of GIS data into Scalable Vector Graphics (SVG) using eXtensible Stylesheet Language Transformation (XSLT).

PUBLICATIONS

Journals, Book Chapters, and Thesis

- **Seung-Hee Bae**, Daniel Halperin, Jevin West, Martin Rosvall, and Bill Howe, "Scalable and Efficient Flow-Based Community Detection for Large-Scale Graph Analysis," *ACM Transactions of Knowledge Discovery from Data (TKDD)*, 2016. [accepted]
- **Seung-Hee Bae**, "Scalable High Performance Multidimensional Scaling," *Ph.D. Dissertation*, Indiana University, Bloomington, 2012.
- Adam Hughes, Yang Ruan, Saliya Ekanayake, **Seung-Hee Bae**, Qunfeng Dong, Mina Rho, Judy Qiu, and Geoffrey Fox, "Interpolative Multidimensional Scaling Techniques for the Identification of Clusters in Very Large Sequence Sets," Special Issue of *BMC Bioinformatics*, vol. 13, no. Suppl 2, p. S9, 2012.
- Judy Qiu and **Seung-Hee Bae**, "Performance of Windows Multicore Systems on Threading and MPI," *Concurrency and Computation: Practice and Experience (CCPE)*, Vol 24(1): 14-28, 2012.

- Jong Youl Choi, **Seung-Hee Bae**, Judy Qiu, Bin Chen, and David Wild, "Browsing Large Scale Cheminformatics Data with Dimension Reduction," *Concurrency and Computation: Practice and Experience (CCPE)*, Vol 23(17): 2315-2325, 2011.
- Thilina Gunarathne, Tak-Lon Wu, Jong Youl Choi, **Seung-Hee Bae**, and Judy Qiu, "Cloud Computing Paradigms for Pleasingly Parallel Biomedical Applications," *Concurrency and Computation: Practice and Experience (CCPE)*, Vol 23(17): 2338-2354, 2011.
- Judy Qiu, Jaliya Ekanayake, Thilina Gunarathne, Jong Youl Choi, **Seung-Hee Bae**, Hui Li, Bingjing Zhang, Tak-Lon Wu, Yang Ruan, Saliya Ekanayake, Adam Hughes, and Geoffrey Fox, "Hybrid cloud and cluster computing paradigms for life science applications," *BMC Bioinformatics*, vol. 11, no. Suppl 12, p. S3, 2010.
- Judy Qiu, Jaliya Ekanayake, Thilina Gunarathne, Jong Youl Choi, **Seung-Hee Bae**, Yang Ruan, Saliya Ekanayake, Stephen Wu, Scott Beason, Geoffrey Fox, Mina Rho, and Haixu Tang, "Data Intensive Computing for Bioinformatics," in *Data Intensive Distributed Computing*, IGI Publishers, 2010.
- Geoffrey Fox, **Seung-Hee Bae**, Jaliya Ekanayake, Xiaohong Qiu, and Huapeng Yuan "Parallel Data Mining from Multicore to Cloudy Grids" in *High Speed and Large Scale Scientific Computing*, IOS Press, Amsterdam, ISBN:978-1-60750-073-5, 2009
- **Seung-Hee Bae**, Haixu Tang, Jing Wu, Jun Xie, and Sun Kim, "dPattern: Transcription Factor Binding Site (TFBS) Discovery in Human Genome using a Discriminative Pattern Analysis," *Bioinformatics*, Vol 23(19): 2619-2621, 2007.
- **Seung-Hee Bae** "An Adaptive Method for the Mutation Rates of Hybrid Genetic Algorithms," *M.S. Thesis*, Seoul National University, 2004

Conferences and Workshops

- **Seung-Hee Bae** and Bill Howe, "GossipMap: A Distributed Community Detection Algorithm for Billion-Edge Directed Graphs," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '15)*, Austin, Texas, Nov. 2015.
- **Seung-Hee Bae**, Daniel Halperin, Jevin West, Martin Rosvall, and Bill Howe, "Scalable Flow-Based Community Detection for Large-Scale Network Analysis," in *Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW 2013)*, Dallas, Texas, Dec. 2013.
- Yang Ruan, Saliya Ekanayake, Mina Rho, Haixu Tang, **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, "DACIDR: Deterministic Annealed Clustering with Interpolative Dimension Reduction using a Large Collection of 16S rRNA Sequences," in *Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM BCB)*, Orlando, Florida, Oct. 2012.
- **Seung-Hee Bae**, Judy Qiu, and Geoffrey C. Fox, "Adaptive Interpolation of Multidimensional Scaling," in *Proceedings of 12th International Conference on Computational Science (ICCS 2012)*, Omaha, Nebraska, Jun. 2012.
- **Seung-Hee Bae**, Judy Qiu, and Geoffrey C. Fox, "Multidimensional Scaling by Deterministic Annealing with Iterative Majorization algorithm," in *Proceedings of 6th IEEE e-Science Conference*, Brisbane, Australia, Dec. 2010.
- **Seung-Hee Bae**, Jong Youl Choi, Judy Qiu, and Geoffrey Fox, "Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation," in *Proceedings of the ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, Chicago, Illinois, Jun. 2010.
- Jong Youl Choi, **Seung-Hee Bae**, Judy Qiu, Geoffrey Fox, Bin Chen, and David Wild, "Browsing Large Scale Cheminformatics Data with Dimension Reduction," in *Proceedings of Emerging Computational Methods for the Life Sciences Workshop of ACM HPDC 2010*, Chicago, Illinois, Jun. 2010.
- Jaliya Ekanayake, Hui Li, Bingjing Zhang, Thilina Gunarathne, **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, "Twister: A Runtime for Iterative MapReduce," in *Proceedings of the First International Workshop on MapReduce and its Applications of ACM HPDC 2010*, Chicago, Illinois, Jun. 2010.

- Jong Youl Choi, **Seung-Hee Bae**, Xiaohong Qiu, and Geoffrey Fox, "High Performance Dimension Reduction and Visualization for Large High-dimensional Data Analysis," in *Proceedings of the 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2010)*, Melbourne, Australia, May 2010.
- Judy Qiu, Scott Beason, **Seung-Hee Bae**, Saliya Ekanayake, and Geoffrey Fox, "Performance of Windows Multicore Systems on Threading and MPI," in *Proceeding of Frontiers of GPU, Multi- and Many-Core Systems Workshop of 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing Conference (CCGrid 2010)*, Melbourne, Australia, May 2010.
- Xiaohong Qiu, Jaliya Ekanayake, Thilina Gunarathne, **Seung-Hee Bae**, Jong Youl Choi, Scott Beason, and Geoffrey Fox, "Using MapReduce Technologies in Bioinformatics and Medical Informatics," in *Proceedings of Using Clouds for Parallel Computations in Systems Biology workshop at SC09*, Portland, Oregon, Nov. 2009.
- **Seung-Hee Bae**, "Parallel Multidimensional Scaling Performance on Multicore Systems," in *Proceedings of Advanced in High-Performance E-Science Middleware and Applications Workshop at eScience 2008*, Indianapolis, IN, USA, Dec. 2008.
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "Parallel Data Mining on Multicore Clusters" in *Proceedings of 7th International Conference on Grid and Cooperative Computing (GCC08)*, Shenzhen, China, 2008.
- Geoffrey Fox, **Seung-Hee Bae**, Jaliya Ekanayake, Xiaohong Qiu, and Huapeng Yuan "Parallel Data Mining from Multicore to Cloudy Grids" in *Proceedings of HPC 2008 High Performance Computing and Grids workshop*, Cetraro, Italy, Jul. 2008
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "Performance of Multicore Systems on Parallel Data Clustering with Deterministic Annealing," in *Proceedings of International Conference on Computational Science (ICCS 2008)*, LNCS 5101, pp. 407-416, Krakow, Poland, Jun. 2008.
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "Parallel Clustering and Dimensional Scaling on Multicore Systems," in *Proceedings of The 2008 High Performance Computing & Simulation Conference (HPCS 2008)*, Nicosia, Cyprus, Jun. 2008.
- Geoffrey Fox, **Seung-Hee Bae**, Rajarshi Guha, Marlon E. Pierce, Xiaohong Qiu, David J. Wild, H. Yuan, "High Performance Robust Datamining for Cheminformatics," PAPER ID: 1168842 at Division of Chemical Information session on *Cheminformatics: From Teaching to Research* Tuesday, April 8, 2008 from 11:35 AM to 12:00 PM at *Spring 2008 American Chemical Society National Meeting & Exposition*, Apr. 6-10, 2008, New Orleans, LA, USA
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "High Performance Data Mining," at *Shanghai Many-Core Workshop*, Shanghai, China, Mar. 2008.
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "High Performance Multi-Paradigm Messaging Runtime Integrating Grids and Multicore System," in *Proceedings of The 3rd IEEE International conference on e-Science and Grid Computing (eScience 2007)*, Bangalore, India, Dec. 2007.
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "Web 2.0 Grids and Cyberinfrastructure," *Web 2.0 Workshop at The 21st Open Grid Forum – OGF21*, Seattle, Washington, Oct. 2007.
- **Seung-Hee Bae** and Byung-Ro Moon, "Mutation Rates in the Context of Hybrid Genetic Algorithms," in *Proceedings of Genetic and Evolutionary Computation Conference (GECCO)*, LNCS 3103, pp. 381-382, Jun. 2004.
- Hwa-Jung Lee, **Seung-Hee Bae**, and Youngsup Kim. "Research for National Space Data Application based on XML", in *Proceedings of The Korean Association of Geographic Information Studies [KAGIS] Annual Spring Conference*, May, 2001

Posters & Demonstration

- **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, “Scalable Dimension Reduction for Large Abstract Data Visualization,” Poster at *IEEE Cluster 2011*, Austin, Texas, Sept. 2011.
- Judy Qiu, Jong Youl Choi, **Seung-Hee Bae**, Thilina Gunarathne, Geoffrey Fox, Bin Cao, and David Wild, “Browsing Large Scale Cheminformatics Data with Dimension Reduction,” Demonstration at *10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2010)*, May 17-20, 2010, Melbourne, Australia
- Xiaohong Qiu, Scott Beason, **Seung-Hee Bae**, Jaliya Ekanayake, Jong Youl Choi, Ruan Yang, presented by Geoffrey Fox, "Parallel Data Analysis from Multicore to Cloudy Grids," at *Microsoft External Research Symposium*, Mar. 31, 2009, Seattle, WA, USA.
- Geoffrey Fox, Xiaohong Qiu, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, Henrik Frystyk Nielsen, Rajarshi Guha, David Wild, Haixu Tang, and Neil Devadasan, "Service Aggregated Linked Sequential Activities: High Performance Data Mining on Multi-core Systems," at *Microsoft All Hands Meeting*, Mar. 5-8, 2008, Seattle, WA, USA.
- Xiaohong Qiu, Geoffrey Fox, Huapeng Yuan, **Seung-Hee Bae**, George Chrysanthakopoulos, and Henrik Frystyk Nielsen, "Performance of a Multi-Paradigm Messaging Runtime on Multicore Systems," *The 8th IEEE/ACM International Conference on Grid Computing (Grid 2007)*, Austin, Texas, Sep. 2007.
- **Seung-Hee Bae**, Haixu Tang, Jing Wu, Jun Xie, and Sun Kim, "A Mixture Model Approach to Identification of Interferon-Stimulated Response Element," *Third Annual Indiana Bioinformatics Conference*, May, 2006.

Technical Reports

- **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, “Visualization of Large High-Dimensional Data via Interpolation Approach of Multidimensional Scaling,” *Technical Report*, Jul. 2012.
- **Seung-Hee Bae**, Judy Qiu, and Geoffrey Fox, “High Performance Multidimensional Scaling for Large High-Dimensional Data Visualization,” *Technical Report*, Apr. 2012.
- Jaliya Ekanayake, Thilina Gunarathne, Judy Qiu, Geoffrey Fox, Scott Beason, Jong Youl Choi, Yang Ruan, **Seung-Hee Bae**, and Hui Li, “Applicability of DryadLINQ to Scientific Applications,” *Technical Report*, Jan. 30 2010.

Presentations

- “GossipMap: A Distributed Community Detection Algorithm for Billion-Edge Directed Graphs,” at *the International Conference on High Performance Computing, Networking, Storage and Analysis (SC '15)*, Austin, Texas, Nov. 2015.
- “Scalable Flow-Based Community Detection for Large-Scale Network Analysis,” at the 3rd *IEEE ICDM Workshop on Data Mining in Networks*, Dallas, Texas, Dec. 2013.
- “Adaptive Interpolation of Multidimensional Scaling,” at *12th International Conference on Computational Science (ICCS 2012)*, Omaha, Nebraska, Jun. 2012.
- “Multidimensional Scaling by Deterministic Annealing with Iterative Majorization algorithm,” at *6th IEEE e-Science Conference*, Brisbane, Australia, Dec. 2010.
- “Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation,” at *The ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, Chicago, Illinois, Jun. 2010.
- “Parallel Multidimensional Scaling Performance on Multicore Systems,” at *e-Science Workshop on Advanced in High-Performance E-Science Middleware and Applications Workshop*, Indianapolis, IN, USA, Dec. 11, 2008.

REFERENCES

Available upon request