Scalable and High Performance Computing for Big Data Analytics in Understanding the Human Dynamics in the Mobile Age

Xuan Shi, Ph.D.
Department of Geosciences, University of Arkansas

As the National Science Foundation (NSF) indicated – "*Theory and experimentation have for centuries been regarded as two fundamental pillars of science. It is now widely recognized that computational and data-enabled science forms a critical third pillar*." Spatiotemporal computation is thus the third pillar of our transdisciplinary research project entitled "IBSS: Spatiotemporal Modeling of Human Dynamics across Social Media and Social Networks" sponsored by NSF. This IBSS large interdisciplinary project will study human dynamics across social media and social networks and focus on the spatiotemporal modeling of information diffusion and the inter-correlation between online activities and real world human behaviors.

Different from the traditional attribute-value data, the vast amount of social media data are largely unstructured, noisy, distributed, and dynamic. Consequently this research has to develop effective and accessible data processing, visualization, and analytical tools for social scientists to study human dynamics and information diffusion through analyzing the dynamic changes of spatiotemporal patterns with two scenarios of human dynamics (disaster warnings/alerts and referendum/propositions of controversial social topics) using computational predictive methods and agent-based modeling (ABM) approaches. When intensive data processing and analytics are inevitably expected, exploring scalable and high performance computing solutions over emerging advanced computer architecture and systems are indispensable to support the tasks of big data analytics and modeling in this project.

With the exponential growth of geospatial and social media data, the challenge of scalable and high performance computing for big data analytics become *urgent* because many research activities are constrained by the *inability* of software or tool that even could *not* complete the computation process. In the problem-solving and decision-making processes, the performance of spatiotemporal computation is severely limited when massive datasets are processed. Heterogeneous geospatial data integration and analytics obviously magnify the complexity and operational time frame. Many large-scale geospatial problems may be not processable at all if the computer system does not have sufficient memory or computational power.

Emerging computer architectures, such as Intel's Many Integrated Core (MIC) Architecture and Graphics Processing Unit (GPU), and advanced computing technologies provide promising solutions to employ massive parallelism and hardware resources to achieve scalability and high performance for data intensive computing over large spatiotemporal and social media data. Exploring novel algorithms and deploying the solutions in massively parallel computing environment to achieve the capability for scalable data processing and analytics over large-scale, complex, and heterogeneous spatial and social media data with consistent quality and high-performance is the central theme of this research task.

New multi-core architectures combined with application accelerators hold the promise to achieve scalability and high performance by exploiting task and data levels of parallelism that are not

supported by the conventional computing systems. Such a parallel or distributed computing environment is particularly suitable for large-scale spatiotemporal computation over big spatial data as proved by our prior works [1-7], while the potential of such advanced infrastructure remains unexplored in this domain. Within this research, we will exploit multicore CPUs, GPUs, and MICs, and potentially clusters of CPUs, GPUs and MICs, to accelerate the computation for timely delivery of result of analyzing the dynamic changes of spatiotemporal patterns with two scenarios of human dynamics using computational predictive methods and agent-based modeling (ABM) approaches.

Given the example in our current initiative of "Understanding and detecting wildfire events from the chaotic social media through supervised learning", we designed the workflow to clarify whether a tweet message is really about a true wildfire event or not. The workflow consists of three sequential procedures as 1) Apply cosine similarity calculation to examine the distance/similarity between tweet messages; 2) Apply Affinity Propagation (AP) to identify exemplars of tweet messages by using the distance value derived from the first step; and 3) Apply accumulative exemplars to classify tweet messages using SVM approach. When the volume of datasets is small, the workflow can be completed by using existing software products quickly. When big data is involved, however, the workflow cannot be implemented successfully. In the case of AP calculation, although AP has obvious advantages in comparison to many other approaches for clustering analysis [8-9], it was acknowledged [10] that "Affinity propagation's computational and memory requirements scale linearly with the number of similarities input; for non-sparse problems where all possible similarities are computed, these requirements scale quadratically with the number of data points." It took hours or a day to complete the AP calculation over some sample datasets discussed in [10]. In our pioneering study [1], AP calculation is an embarrassingly parallel problem that needs large memory to process big data, e.g. 10K points need 4GB memory, 20K points need 16GB memory, and 40K points need 64GB memory. Solutions over clusters of GPUs have to be developed to handle big data. In the case of agent based modeling, ABM has been implemented on individual GPUs or multiple GPUs. Data communication may have to be executed multiple times in order to complete the computational processes thus increase the difficulty and challenge in development. Since ABM is a critical component in this research project, we will collaborate with other team members to develop hybrid solutions to complete ABM simulation over big spatial and social media data.

References:

1. Shi, X. 2015. Parallelizing affinity propagation using GPUs for spatial cluster analysis over big geospatial data. Proceedings of Geocomputation 2015.
2. Guan, Q., Shi, X., Huang, M., and Lai, C. 2015. A hybrid parallel cellular automata model for urban growth simulation over GPU/CPU heterogeneous architectures. International Journal of Geographical Information Science, 1-21. Online publication date: May 20, 2015. DOI: 10.1080/13658816.2015.1039538
3. Huang, M., Lai, C., Shi, X., Hao, Z. and You, H. 2015. Study of Parallel Programming Models on Computer Clusters with Intel MIC Coprocessors. International Journal of High Performance Computing Applications, April 2015.
4. Shi, X., Lai, C., Huang, M. and You, H. 2014. Geocomputation over the Emerging Heterogeneous Computing Infrastructure. Transactions in GIS, vol. 18, no. S1, pp. 3-24, Nov. 2014.
5. Shi, X., Huang, M., You, H., Lai, C. and Chen, Z. 2014. Unsupervised image classification over supercomputers Kraken, Keeneland and Beacon. GIScience & Remote Sensing, Volume 51, Issue 3. 2014. pp. 321-338
6. Shi, X. and Ye, F. 2013. Kriging interpolation over heterogeneous computer architectures and systems. GIScience & Remote Sensing. Volume 50, Issue 2, 2013. pp.196-211
7. Ye, F. and Shi, X. 2013. Parallelizing ISODATA Algorithm for Unsupervised Image Classification on GPU. In: Modern Accelerator Technologies for GIScience. pp 145-156. Springer
8. Frey, B.J. and Dueck, D. 2007. Clustering by Passing Messages Between Data Points. Science 315, 972–976, February 2007
9. Frey, B.J. and Dueck, D. 2008. Response to Comment on "Clustering by Passing Messages between Data Points". Science 319, 726 (2008)
10. Dueck, D. 2009. Affinity Propagation: Clustering Data by Passing Messages. Doctoral dissertation, University of Toronto.