A Locality-aware Approach to Scalable Parallel Agent-based Models of Spatially Heterogeneous Interactions

Zhaoya Gong[1], Wenwu Tang[1], Jean-Claude Thill[1]

[1] Department of Geography and Earth Sciences, University of North Carolina at Charlotte,
Email: {zgong1, Jean-Claude.Thill, WenwuTang}@uncc.edu

Abstract

Spatially explicit agent-based models have a great potential to mitigate their computational costs by taking advantage of parallel and high-performance computing. However, the spatial dependency and heterogeneity of interactions pose challenges for parallel SE-ABMs to achieve good scalability. This paper applies the principle of data locality to tackle these challenges by extending a theoretical approach to the representation of spatial computational domain. Using a graph-based approach, we minimize the interaction overhead in domain decomposition and maximize the efficiency of allocating computing resources. Our approach is illustrated by simulating agent-level spatial interaction models on different parallel platforms.

Keywords: Locality-aware, parallel agent-based models, spatially heterogeneous interactions**.**

1. Introduction

Spatially explicit agent-based models (SE-ABM) simulate dynamic interactions among spatially situated agents and between these agents and their environments. These decentralized interactions propagate across scales to generate complex regularities that in turn reshape agent behaviors. As a computationally intensive approach, SE-ABM can benefit from various parallel computing resources in the emerging cyberinfrastructure so as to maintain its validity while coping with the challenges of massive geospatial data. However, the spatial dependency and heterogeneity of agent-level interactions pose challenges for parallel SE-ABMs to achieve good scalability (Gong et al, 2013). First, when agents interact interdependently of each other, overhead is introduced by the cost of data access from one agent to the others that are assigned to parallel computing tasks and by the cost of synchronization to maintain their data coherence and integrity. Second, this overhead is unevenly distributed across space because of the heterogeneous patterns of interactions that result from various geographic and social neighborhoods of individual agents. This study aims to tackle these challenges by employing the principle of data locality to extend a theoretical approach to the representation of spatial computational domain (Wang and Armstrong, 2009) that was intended to guide the parallelization of computationally intensive geographical analyses. Our graph-based approach is tailored to minimizing the interaction overhead in domain decomposition and to the efficient allocation of computing resources. The usefulness of this approach is demonstrated by applying it to an ABM of spatial interaction system that simulates information exchange and the diffusion of opinion development among individual decision makers.

2. A locality-aware approach

2.1 The locality principle

The principle of locality is one of the cornerstones of computer science (Denning, 2006). It refers to the idea that system optimality is achieved when the proximity of the clustering of frequently accessed data (also called locality sets) is close to computations. A modern model of locality enables the context awareness of computation through four key elements: *observers*' actions/interactions are monitored in order to identify their *neighborhoods* via *inference*, so that the performance of computational tasks can be *optimized* for observers by being aware of their neighborhoods and adapting to them. This model is well suited to the design of efficient parallel SE-ABMs. In the language of ABM, each agent as an observer has its neighborhood definition that can be inferred from either monitoring its dynamic interaction patterns or static structural declaration. A neighborhood is not necessarily geographical but could be defined through social or other dimensions. Because interactions are most intensive between an agent and its neighborhood, the presence of all neighboring agents in its locality set is imperative to minimize the cost of data access such that the performance of the computational task to process this agent can be optimized.

2.2 Locality-aware computational domain

Wang and Armstrong (2009) formalize a spatial computational domain that comprises layers of two-dimensional computational intensity surfaces mapped from a spatial domain. Each surface is represented by a grid where each cell indicates the computational intensity at location (i,j). This grid representation features a spatially contiguous space. Therefore, neighborhoods defined in this space have to be geographically continuous and cannot accommodate a social structure that has neighboring agents distributed discretely across space. The granularity of the spatial computational domain representation is determined by the choice of cell size, which is coarsely dependent on two trading-off factors, namely the cost of decomposition and the sufficient concurrency, but still contingent on a certain arbitrariness. Most importantly, there is no explicit representation of the interdependency between agents whose granularity is usually much smaller than that for the cell size.

To extend this theoretical approach to parallel SE-ABMs, we employ the principle of locality and use graphs to represent the locality-aware computational domain that comprises, model and platform, two layers (fig. 1). In the model layer, a vertex represents an agent and an edge represents the independency between two agents in terms of data access due to their interaction (refer to table 1 for a comparison with a grid representation). Therefore, the space represented by a graph is topological, which enables discrete neighborhood structures. The granularity is naturally at the agent level, which avoids the uncertainty associated with pre-determined cell size in a grid representation. In the platform layer, vertices represent processing units (e.g., CPUs) of a computing platform, while edges are weighted by communication costs between processing units.

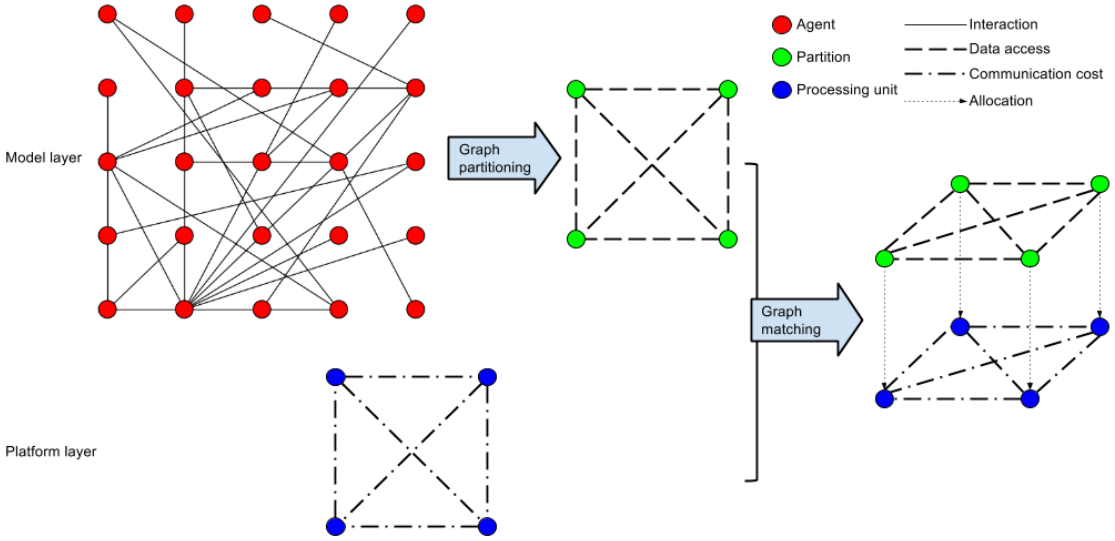

Figure 1. Locality-aware computational domain.

Representation	Grid	Graph
Space	Spatially contiguous	Topological
Neighborhood	Continuous	Discrete
Granularity	Cell size	Agent level
Interdependency	N/A	Data access

Table 1. A comparison of grid and graph representations.

Transformations are performed for the model layer to estimate the computational intensity of a SE-ABM in order to guide its domain decomposition to produce equal partitions of the model graph with inter-partition data access minimized. Two types of transformations can be differentiated: 1) Operation-centric transformations estimate the algorithmic time complexity of the operations that process agent behaviors; 2) Data-centric transformations can be further distinguished by two specific functions. The data access function (analog to the distance function in the principle of locality) estimates the intensity to access and transfer data. The memory function estimates the algorithmic space complexity for the memory requirement. The estimated computational intensity values are embedded in the model layer as follows: a vertex has a vector of values attached to it, indicating the estimated operation and memory intensity while an edge carries a weight indicating the data access intensity between two agents.

Partitioned model layer becomes a generalized graph with partitions as vertices and data access intensity between partitions as weighted edges. To allocate model partitions to the processing units with an awareness of data locality, the problem can be formalized as matching two graphs. Given model graph $G_M = (V_M, E_M)$ and platform graph $G_P = (V_P, E_P)$, with $|V_M| = |V_P|$, the problem is to find a one-to-one mapping $f: V_M \rightarrow V_P$ such that an objective is optimized, e.g., minimizing total data access cost $\sum_{E_M \rightarrow E_P} E_M^W \, E_P^W$.

3. Experimentations on shared-memory platforms

In order to validate our expended theoretical approach and to evaluate its effectiveness to alleviate the two stated challenges for a scalable parallel SE-ABM, we design two sets of experiments and conduct them on three different shared-memory platforms.

The first set of experiments focuses on a homogenous neighborhood configuration for every agent. A distance decay function with stochasticity is used to create this neighborhood. Various interaction ranges are examined across experiments to investigate how different levels of data access cost are handled by our approach. The second set of experiments emphasizes the heterogeneous neighborhood configurations for agents, which renders the spatial heterogeneity of interaction patterns. A scale-free network is constructed among agents to form their neighborhoods and generate the heterogeneous patterns of interaction. Different topologies of scale-free networks are tested to examine the efficiency of our approach to guide domain decomposition and resource allocation in terms of maximizing data locality and minimizing data access cost.

A shared-memory paradigm comprises multiple or many processors/cores connected to a common memory space. Symmetric multiprocessing (SMP) is an architecture that provides every processing unit with identical access speed to the memory. On a CPU-based multi-core platform, because all cores share a fixed bandwidth to access the memory, the scalability of SMP becomes a bottleneck as cores grows. The architecture of Non-Uniform Memory Access (NUMA) resolves this issue by using distributed memory banks attached to CPU cores, which enables all memory banks to be shared, while a core accesses its local bank faster than non-local banks. In contrast to multi-core CPU processors, Intel Many Integrated Core architecture is a many-core coprocessor platform that can run hundreds of threads in parallel. It aims at achieving great throughput performance with attached high-bandwidth memory, while adopting a SMP design with cache hierarchies similar to many-core GPUs. Data locality can be exploited by maximizing its cache usage.

4. Concluding discussions

This study aims to address challenges to scalable parallel SE-ABMs: the spatial dependency and heterogeneity of agent interactions. To this end, we formalize locality-aware computational domain with a graph-based approach. Two sets of experimentation are conducted on three different shared-memory platforms in order to illustrate the applicability of our approach. Scalable performance is expected by evaluating its effectiveness for resolving the identified challenges in experiments. Although it has only been tested for shared-memory systems, our approach intends to be generic and can potentially be applied to other parallel paradigms such as message-passing.

5. References

Denning P J, 2006, The locality principle. In: Barria J (eds), *Communication Networks and Computer Systems*. Imperial College Press, London, 43-67.

Gong Z, Tang W, Bennett D A and Thill J-C, 2013, Parallel agent-based simulation of individual-level spatial interactions within a multicore computing environment. *International Journal of Geographical Information Science*, 27(6):1152-1170.

Wang S and Armstrong M P, 2009, A theoretical approach to the use of cyberinfrastructure in geographical analysis, *International Journal of Geographical Information Science*, 23(2):169-193.