|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | World Population Scale Epidemiological Study | |
| **Vertical (area)** | | Epidemiology, Simulation Social Science, Computational Social Science | |
| **Author/Company/Email** | | Madhav Marathe Stephen Eubank or Chris Barrett/ Virginia Bioinformatics Institute, Virginia Tech, [mmarathe@vbi.vt.edu](mailto:mmarathe@vbi.vt.edu), [seubank@vbi.vt.edu](mailto:seubank@vbi.vt.edu) or cbarrett@vbi.vt.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | Government and non-profit institutions involved in health, public policy, and disaster mitigation. Social Scientist who wants to study the interplay between behavior and contagion. | |
| **Goals** | | (a) Build a synthetic global population. (b) Run simulations over the global population to reason about outbreaks and various intervention strategies. | |
| **Use Case Description** | | Prediction and control of pandemic similar to the 2009 H1N1 influenza. | |
| **Current**  **Solutions** | **Compute(System)** | | Distributed (MPI) based simulation system written in Charm++. Parallelism is achieved by exploiting the disease residence time period. |
| **Storage** | | Network file system. Exploring database driven techniques. |
| **Networking** | | Infiniband. High bandwidth 3D Torus. |
| **Software** | | Charm++, MPI |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Generated from synthetic population generator. Currently centralized. However, could be made distributed as part of post-processing. |
| **Volume (size)** | | 100TB |
| **Velocity**  **(e.g. real time)** | | Interactions with experts and visualization routines generate large amount of real time data. Data feeding into the simulation is small but data generated by simulation is massive. |
| **Variety**  **(multiple datasets, mashup)** | | Variety depends upon the complexity of the model over which the simulation is being performed. Can be very complex if other aspects of the world population such as type of activity, geographical, socio-economic, cultural variations are taken into account. |
| **Variability (rate of change)** | | Depends upon the evolution of the model and corresponding changes in the code. This is complex and time intensive. Hence low rate of change. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Robustness of the simulation is dependent upon the quality of the model. However, robustness of the computation itself, although non-trivial, is tractable. |
| **Visualization** | | Would require very large amount of movement of data to enable visualization. |
| **Data Quality (syntax)** | | Consistent due to generation from a model |
| **Data Types** | | Primarily network data. |
| **Data Analytics** | | Summary of various runs and replicates of a simulation |
| **Big Data Specific Challenges (Gaps)** | | Computation of the simulation is both compute intensive and data intensive. Moreover, due to unstructured and irregular nature of graph processing the problem is not easily decomposable. Therefore it is also bandwidth intensive. Hence, a supercomputer is applicable than cloud type clusters. | |
| **Big Data Specific Challenges in Mobility** | | None | |
| **Security & Privacy**  **Requirements** | | Several issues at the synthetic population-modeling phase (see social contagion model). | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | In general contagion diffusion of various kinds: information, diseases, social unrest can be modeled and computed. All of them are agent-based model that utilize the underlying interaction network to study the evolution of the desired phenomena. | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Social Contagion Modeling | |
| **Vertical (area)** | | Social behavior (including national security, public health, viral marketing, city planning, disaster preparedness) | |
| **Author/Company/Email** | | Madhav Marathe or Chris Kuhlman /Virginia Bioinformatics Institute, Virginia Tech [mmarathe@vbi.vt.edu](mailto:Tech/mmarathe@vbi.vt.edu) or ckuhlman@vbi.vt.edu | |
| **/Actors/Stakeholders and their roles and responsibilities** | |  | |
| **Goals** | | Provide a computing infrastructure that models social contagion processes.  The infrastructure enables different types of human-to-human interactions (e.g., face-to-face versus online media; mother-daughter relationships versus mother-coworker relationships) to be simulated. It takes not only human-to-human interactions into account, but also interactions among people, services (e.g., transportation), and infrastructure (e.g., internet, electric power). | |
| **Use Case Description** | | Social unrest. People take to the streets to voice unhappiness with government leadership. There are citizens that both support and oppose government. Quantify the degrees to which normal business and activities are disrupted owing to fear and anger. Quantify the possibility of peaceful demonstrations, violent protests. Quantify the potential for government responses ranging from appeasement, to allowing protests, to issuing threats against protestors, to actions to thwart protests. To address these issues, must have fine-resolution models and datasets. | |
| **Current**  **Solutions** | **Compute(System)** | | Distributed processing software running on commodity clusters and newer architectures and systems (e.g., clouds). |
| **Storage** | | File servers (including archives), databases. |
| **Networking** | | Ethernet, Infiniband, and similar. |
| **Software** | | Specialized simulators, open source software, and proprietary modeling environments. Databases. |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Many data sources: populations, work locations, travel patterns, utilities (e.g., power grid) and other man-made infrastructures, online (social) media. |
| **Volume (size)** | | Easily 10s of TB per year of new data. |
| **Velocity**  **(e.g. real time)** | | During social unrest events, human interactions and mobility key to understanding system dynamics. Rapid changes in data; e.g., who follows whom in Twitter. |
| **Variety**  **(multiple datasets, mashup)** | | Variety of data seen in wide range of data sources. Temporal data. Data fusion.  Data fusion a big issue. How to combine data from different sources and how to deal with missing or incomplete data? Multiple simultaneous contagion processes. |
| **Variability (rate of change)** | | Because of stochastic nature of events, multiple instances of models and inputs must be run to ranges in outcomes. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Failover of soft realtime analyses. |
| **Visualization** | | Large datasets; time evolution; multiple contagion processes over multiple network representations. Levels of detail (e.g., individual, neighborhood, city, state, country-level). |
| **Data Quality (syntax)** | | Checks for ensuring data consistency, corruption. Preprocessing of raw data for use in models. |
| **Data Types** | | Wide-ranging data, from human characteristics to utilities and transportation systems, and interactions among them. |
| **Data Analytics** | | Models of behavior of humans and hard infrastructures, and their interactions. Visualization of results. |
| **Big Data Specific Challenges (Gaps)** | | How to take into account heterogeneous features of 100s of millions or billions of individuals, models of cultural variations across countries that are assigned to individual agents? How to validate these large models? Different types of models (e.g., multiple contagions): disease, emotions, behaviors. Modeling of different urban infrastructure systems in which humans act. With multiple replicates required to assess stochasticity, large amounts of output data are produced; storage requirements. | |
| **Big Data Specific Challenges in Mobility** | | How and where to perform these computations? Combinations of cloud computing and clusters. How to realize most efficient computations; move data to compute resources? | |
| **Security & Privacy**  **Requirements** | | Two dimensions. First, privacy and anonymity issues for individuals used in modeling (e.g., Twitter and Facebook users). Second, securing data and computing platforms for computation. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Fusion of different data types. Different datasets must be combined depending on the particular problem. How to quickly develop, verify, and validate new models for new applications. What is appropriate level of granularity to capture phenomena of interest while generating results sufficiently quickly; i.e., how to achieve a scalable solution. Data visualization and extraction at different levels of granularity. | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |

**Note: No proprietary or confidential information should be included**