**Use Cases from NBD(NIST Big Data) Requirements WG**

<http://bigdatawg.nist.gov/home.php>

**Contents**

1. Blank Template
2. UAVSAR Data Processing, Data Product Delivery, and Data Services (Scientific Research: Earth Science) Andrea Donnellan and Jay Parker, NASA JPL
3. MERRA Analytic Services MERRA/AS (Scientific Research: Earth Science) John L. Schnase & Daniel Q. Duffy / NASA Goddard Space Flight Center
4. DataNet Federation Consortium DFC (Scientific Research: Collaboration Environments) Reagan Moore, University of North Carolina at Chapel Hill
5. Semantic Graph-search on Scientific Chemical and Text-based Data (Management of Information from Research Articles) Talapady Bhat, NIST
6. Atmospheric Turbulence - Event Discovery and Predictive Analytics (Scientific Research: Earth Science) Michael Seablom, NASA HQ
7. Pathology Imaging/digital pathology (Healthcare) Fusheng Wang, Emory University
8. Genomic Measurements (Healthcare) Justin Zook, NIST
9. Cargo Shipping (Industry) William Miller, MaCT USA
10. Radar Data Analysis for CReSIS (Scientific Research: Polar Science and Remote Sensing of Ice Sheets) Geoffrey Fox, Indiana University
11. Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle (Scientific Research: Physics) Geoffrey Fox, Indiana University
12. Netflix Movie Service (Commercial Cloud Consumer Services) Geoffrey Fox, Indiana University
13. Web Search (Commercial Cloud Consumer Services) Geoffrey Fox, Indiana University

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | CINET: Cyberinfrastructure for Network (Graph) Science and Analytics | |
| **Vertical (area)** | | Network Science | |
| **Author/Company/Email** | | Team lead by Virginia Tech and comprising of researchers from Indiana University, University at Albany, North Carolina AT, Jackson State University, University at Houston Downtown, Argonne National Laboratory  Point of Contact: Madhav Marathe or Keith Bisset, Network Dynamics and Simulation Science Laboratory, Virginia Bio-informatics Institute Virginia Tech, mmarathe@vbi.vt.edu / kbisset@vbi.vt.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | Researchers, practitioners, educators and students interested in the study of networks. | |
| **Goals** | | CINET cyberinfrastructure middleware to support network science. This middleware will give researchers, practitioners, teachers and students access to a computational and analytic environment for research, education and training. The user interface provides lists of available networks and network analysis modules (implemented algorithms for network analysis). A user, who can be a researcher in network science area, can select one or more networks and analysis them with the available network analysis tools and modules. A user can also generate random networks following various random graph models. Teachers and students can use CINET for classroom use to demonstrate various graph theoretic properties and behaviors of various algorithms. A user is also able to add a network or network analysis module to the system. This feature of CINET allows it to grow easily and remain up-to-date with the latest algorithms.  The goal is to provide a common web-based platform for accessing various (i) network and graph analysis tools such as SNAP, NetworkX, Galib, etc. (ii) real-world and synthetic networks, (iii) computing resources and (iv) data management systems to the end-user in a seamless manner. | |
| **Use Case Description** | | Users can run one or more structural or dynamic analysis on a set of selected networks. The domain specific language allows users to develop flexible high level workflows to define more complex network analysis. | |
| **Current**  **Solutions** | **Compute(System)** | | A high performance computing cluster (DELL C6100), named Shadowfax, of 60 compute nodes and 12 processors (Intel Xeon X5670 2.93GHz) per compute node with a total of 720 processors and 4GB main memory per processor.  Shared memory systems ; EC2 based clouds are also used  Some of the codes and networks can utilize single node systems and thus are being currently mapped to Open Science Grid |
| **Storage** | | 628 TB GPFS |
| **Networking** | | Internet, infiniband. A loose collection of supercomputing resources. |
| **Software** | | Graph libraries: Galib, NetworkX.  Distributed Workflow Management: Simfrastructure, databases, semantic web tools |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | A single network remains in a single disk file accessible by multiple processors. However, during the execution of a parallel algorithm, the network can be partitioned and the partitions are loaded in the main memory of multiple processors. |
| **Volume (size)** | | Can be hundreds of GB for a single network. |
| **Velocity**  **(e.g. real time)** | | Two types of changes: (i) the networks are very dynamic and (ii) as the repository grows, we expect atleast a rapid growth to lead to over 1000-5000 networks and methods in about a year |
| **Variety**  **(multiple datasets, mashup)** | | Data sets are varied: (i) directed as well as undirected networks, (ii) static and dynamic networks, (iii) labeled, (iv) can have dynamics over these networks, |
| **Variability (rate of change)** | | The rate of graph-based data is growing at increasing rate. Moreover, increasingly other life sciences domains are using graph-based techniques to address problems. Hence, we expect the data and the computation to grow at a significant pace. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Challenging due to asynchronous distributed computation. Current systems are designed for real time synchronous response. |
| **Visualization** | | As the input graph size grows the visualization system on client side is stressed heavily both in terms of data and compute. |
| **Data Quality (syntax)** | |  |
| **Data Types** | |  |
| **Data Analytics** | |  |
| **Big Data Specific Challenges (Gaps)** | | Parallel algorithms are necessary to analyze massive networks. Unlike many structured data, network data is difficult to partition. The main difficulty in partitioning a network is that different algorithms require different partitioning schemes for efficient operation. Moreover, most of the network measures are global in nature and require either i) huge duplicate data in the partitions or ii) very large communication overhead resulted from the required movement of data. These issues become significant challenges for big networks.  Computing dynamics over networks is harder since the network structure often interacts with the dynamical process being studied.  CINET enables large class of operations across wide variety, both in terms of structure and size, of graphs. Unlike other compute + data intensive systems, such as parallel databases or CFD, performance on graph computation is sensitive to underlying architecture. Hence, a unique challenge in CINET is manage the mapping between workload (graph type + operation) to a machine whose architecture and runtime is conducive to the system.  Data manipulation and bookkeeping of the derived for users is another big challenge since unlike enterprise data there is no well defined and effective models and tools for management of various graph data in a unified fashion. | |
| **Big Data Specific Challenges in Mobility** | |  | |
| **Security & Privacy**  **Requirements** | |  | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | HPC as a service. As data volume grows increasingly large number of applications such as biological sciences need to use HPC systems. CINET can be used to deliver the compute resource necessary for such domains. | |
| **More Information (URLs)** | | http://cinet.vbi.vt.edu/cinet\_new/ | |
| **Note:** <additional comments> | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **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**