The *Ga* Astrometric Database

Big Data Processing

Vermont Technical College

Like Gaia, Only Not Really

- Gaia is a European Space Agency Mission
  - Measures accurate positions and radial velocities of 1 billion stars
  - Requires multiple measurements of each star over a five year period
  - Will allow us to make a 3D map of this region of the galaxy
    - … with accurate stellar distances and velocities in real space
  - See: http://sci.esa.int/gaia/
- Gaia Data Processing and Analysis Consortium (DPAC)
  - See: http://www.cosmos.esa.int/web/gaia/dpac

Gaia Data Volume

- From the DPAC web site: “A primary motivation behind the Data Processing and Analysis Consortium (DPAC) is the unprecedented amount of data Gaia generates: surveying **1 billion stars**, **70 times each over five years amounts to an average of 70 million objects observed each day**! This translates into **40 Gigabytes of information per day**, or **73 Terabytes over the full, nominal life of the mission**. Taking into account the *additional data products* that are created from the basic observations leads to a **total volume of about 1 Petabyte (1 million Gigabytes) for the complete dataset**.”

What is Ga?

- Ga is an imaginary astrometrics mission producing imaginary data
  - Conceptually similar to Gaia, but far more limited
  - Much simpler
- What is the point of Ga?
  - *Educational!*
  - We can generate as much (or as little) imaginary data as we require
  - We can generate imaginary data with whatever properties we want
  - We can avoid complicating factors, adding them later if desired

How Many Stars?

- Assume…
  - Average stellar separation is 4 ly. (The distance between the sun and the α-Centuri system)
  - One star in a box 4 ly on a side => density = 0.016 stars per cubic ly.
  - Number of stars in a sphere with radius r light years => $0.016 * (4/3)\pi r^3$

Radius (light years)	Number of Stars
10	67
100	67,000
1,000	67,000,000
10,000	67,000,000,000

Realistic?

- In the real galaxy…
  - The stellar density is not uniform
    - Stars primarily in a disk of spiral arms
    - Greater density in the core; less in the arms; even less between the arms
  - The galaxy is about 100,000 ly across; contains ~250 billion stars
    - Makes you realize how little Gaia is actually going to observe!
  - The numbers on the previous slide are in the right general range
- The Ga imaginary data should be at about this scale

Ga Data Set

- Let Ga observe all stars out to about 1000 light years
  - Assume: 50 million stars
- Let Ga observe all these stars once per day
  - That’s 50 million observations per day
  - Gaia is doing 70 million observations per day, so this isn’t a crazy amount
  - Total = 365*50 million = 17.8 billion observations per year

Ga Observation

- Each observation is a line of text containing:
  - The JD number (12 characters, e.g.: 2417815.4650)
  - The star's ID number (8 digits)
  - The precise position of the star on the sky using ecliptic longitude and latitude (in degrees to 10 µas resolution, e.g.: +020.123456789,-20.123456789)
    - That's approximately the resolution Gaia has
- Total of 51 characters per observation
- Data Set = 51 * 17.8 billion = **908 GiB**
  - We will tune the size of the data set by adjusting the number of stars (and hence the average stellar density)

Imaginary Data Generator

- The Imaginary Data Generator (IDG) creates “fake” Ga data
  - A program that takes the number of stars as a parameter
  - Generates observations for the year 2020
  - Outputs the data to a file
- Current version of IDG is very simple minded
  - Future versions could embed various “interesting” effects in the data that later analysis can find.
- IDG written in Scala and available on the class web site
  - You can create imaginary data at a small scale for local testing
  - We can create imaginary data at a large scale for performance testing