CLOUD COMPUTING APPLICATIONS

Graph Processing

Roy Campbell & Reza Farivar

Graph Processing

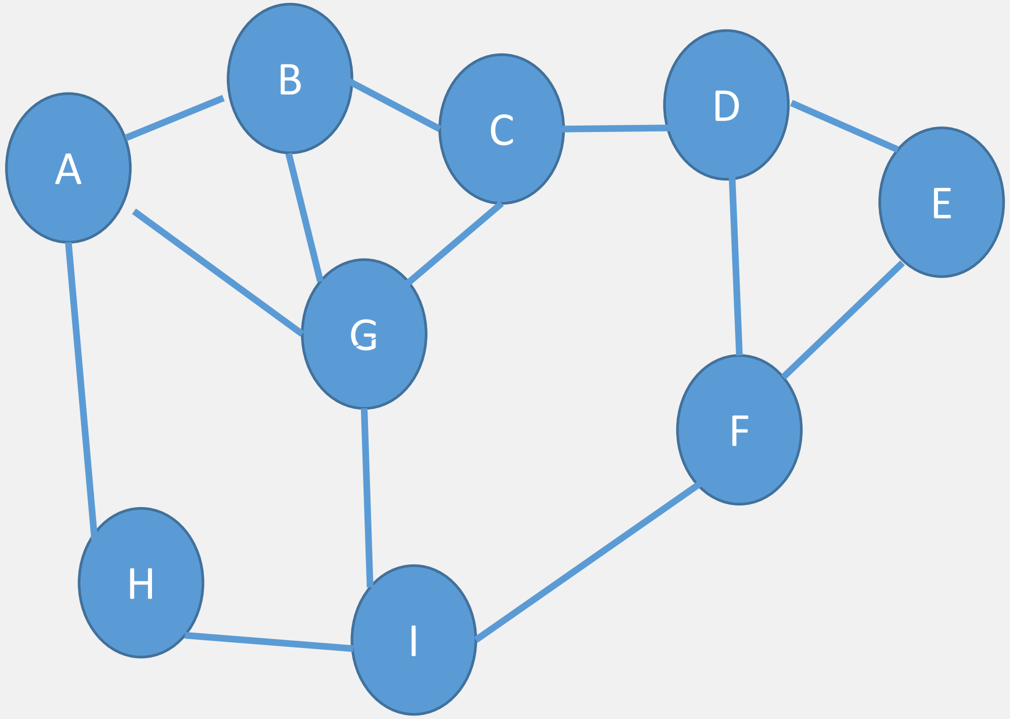
- A graph database is any storage system that provides index-free adjacency. Has pointers to adjacent elements…
- Nodes represent entities (people, businesses, accounts…)
- Properties are pertinent information that relate to nodes
- Edges interconnect nodes to nodes or nodes to properties and they represent the relationship between the two

Graph and Relational Databases

- Graph Database
  - Associative data sets
  - Structure of object-oriented applications
  - Do not require join operators
- Relational Database
  - Perform same operation on large numbers of data elements
  - Use relational model of data
  - Entity type has own table
    - Rows are instances of entity
    - Columns represent values attributed to that instance
  - Rows in one table can be related to rows in another table via unique key per row

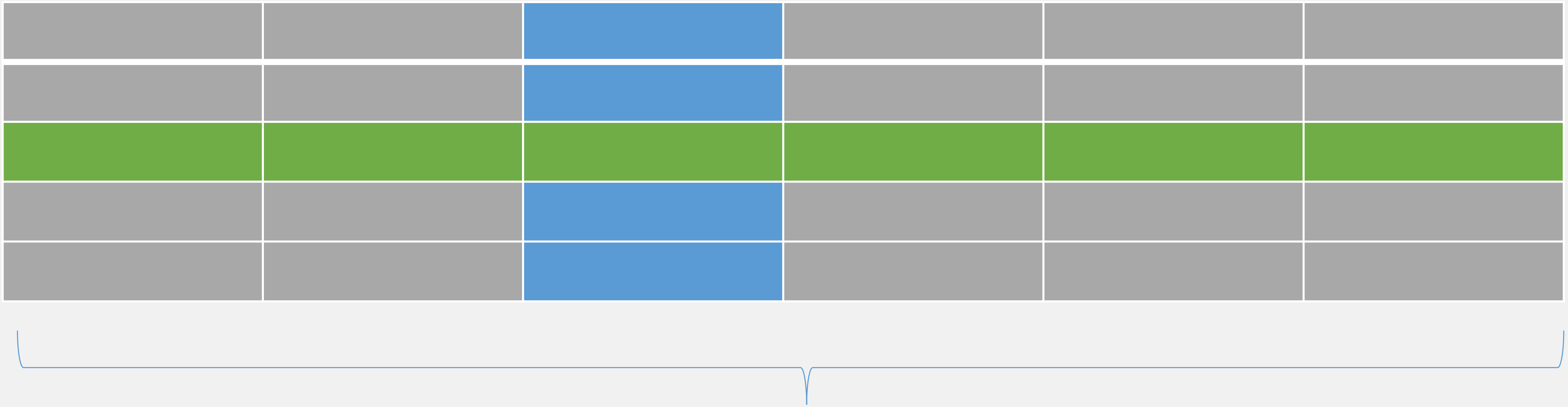
Graph

Vertex	Property/Edge	Vertex
A		B, G, H
B		C, G, A
C		B, D, G
D		C, E, F
E		D, F
F		D, E, I
G		A, B, C, I
H		A, I
I		F, G, H



Or use a sparse matrix (table)

Relational Database



Attribute or Column

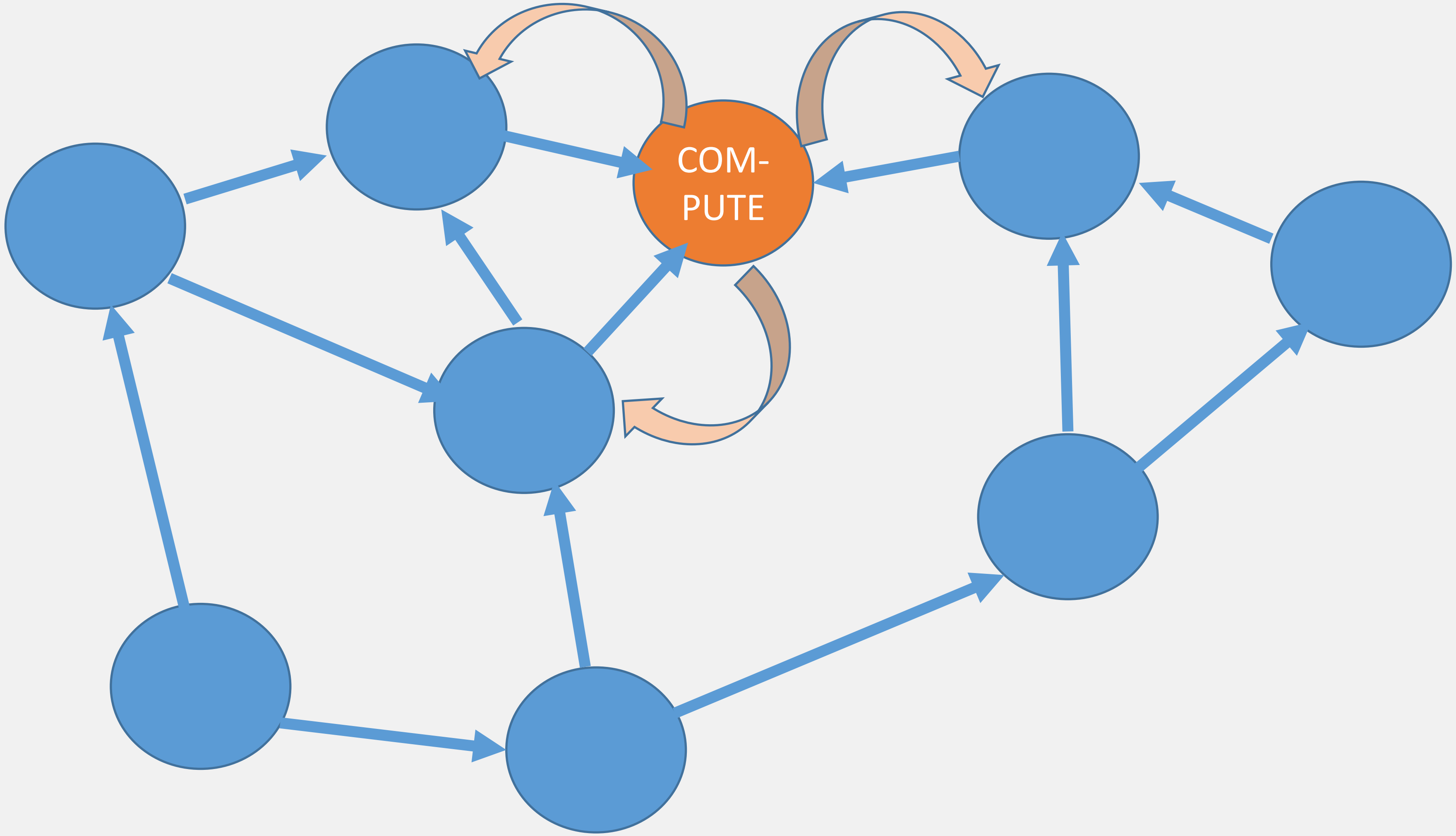
Tuple or row

Table is relation

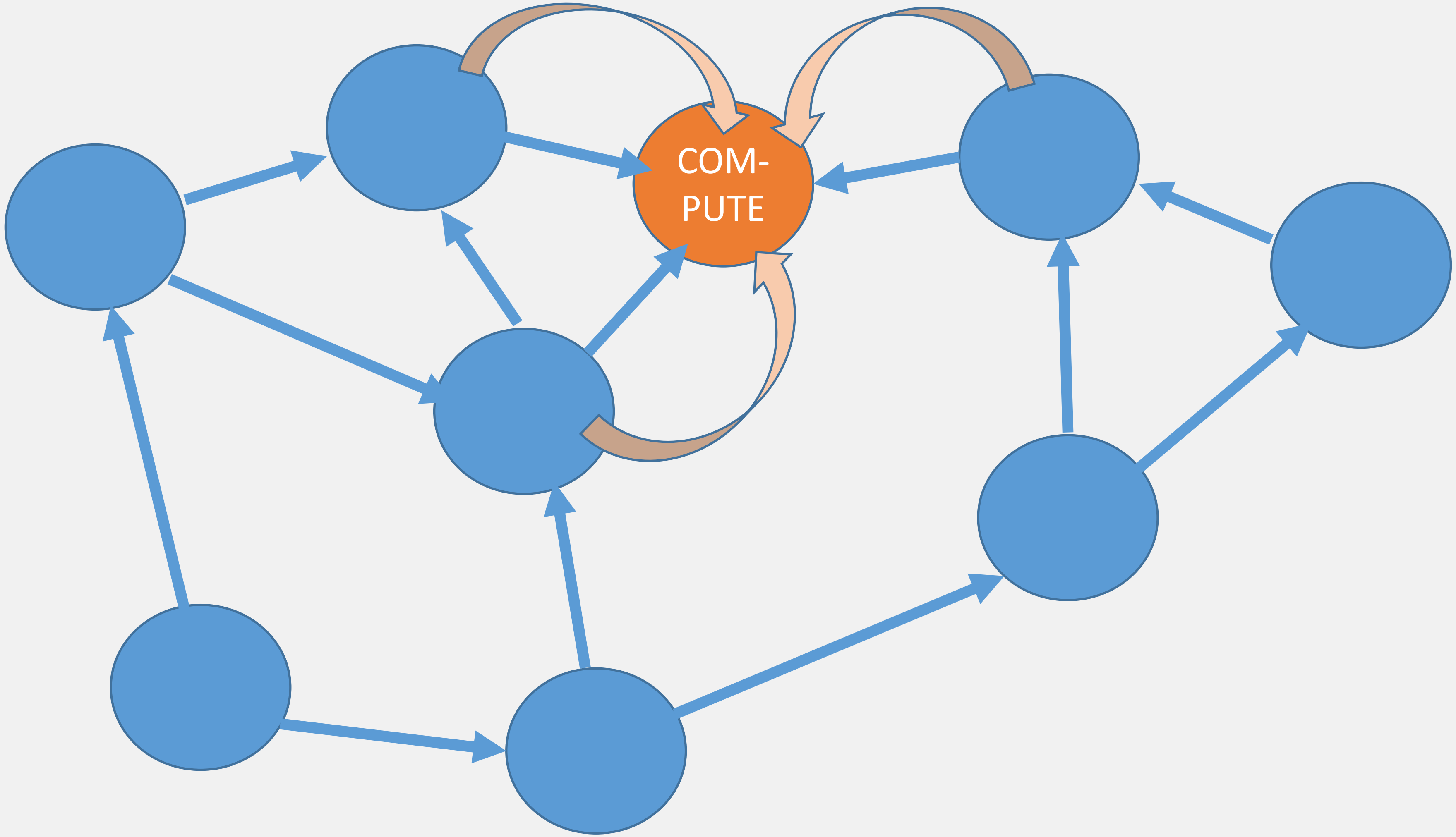
Graph Computing

- Think like a vertex
- Two basic operations:
  - Fusion: aggregate information from neighbors to a set of entities
  - Diffusion: propagate information from a vertex to neighbors

Diffusion



Fusion

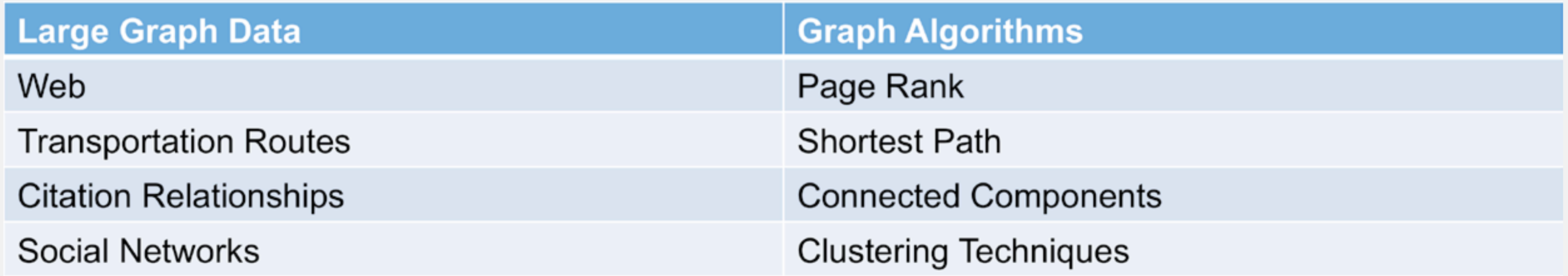


Graph Problem Example

Return all sets of vertices (triad) with edges (A,B) (A,C) (C,B) from a directed graph

Graph Processing

Graph computations involve local data (small part of graph surrounding a vertex), and the connectivity between vertices is sparse. The data may not all fit into one node. This makes it difficult to fit always into the map/reduce model.



Large Graph Data	Graph Algorithms
Web	Page Rank
Transportation Routes	Shortest Path
Citation Relationships	Connected Components
Social Networks	Clustering Techniques

Scale of Graphs in Current MLDM Literature

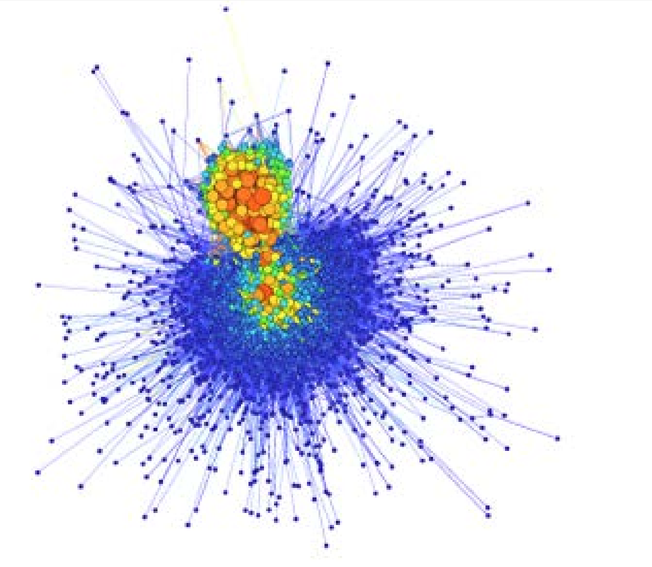
	n (vertices in millions)	m (edges in millions)	size
DBLP	0.3	1	10 MB
AS-Skitter	1.7	11	142 MB
LJ	4.8	69	337.2 MB
USRD	24	58	586.7 MB
BTC	165	773	5.3 GB
WebUK	106	1877	8.6 GB
Twitter	42	1470	24 GB
YahooWeb 2002	1413	6636	120 GB

Graph scale: on order of billions of edges, tens of gigabytes

Scale of Real-World Graphs

Social scale …

- 1 billion vertices, 100 billion edges
- 2.92 TB adjacency list

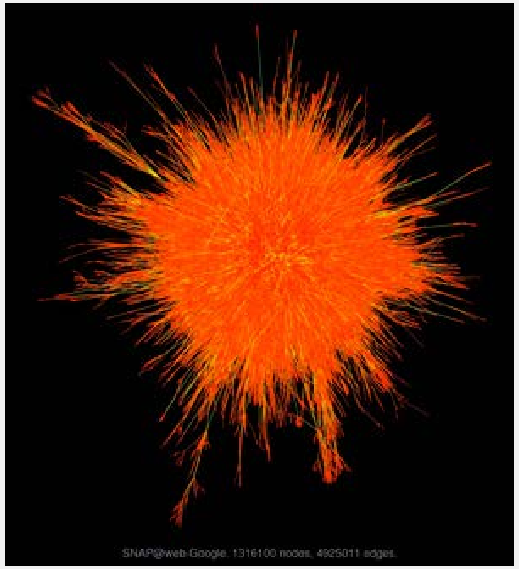


Twitter graph from Gephi data set (http://www.gephi.org)

Scale of Real-World Graphs

Web scale …

- 50 billion vertices, 1 trillion edges
- 29.5 TB adjacency list

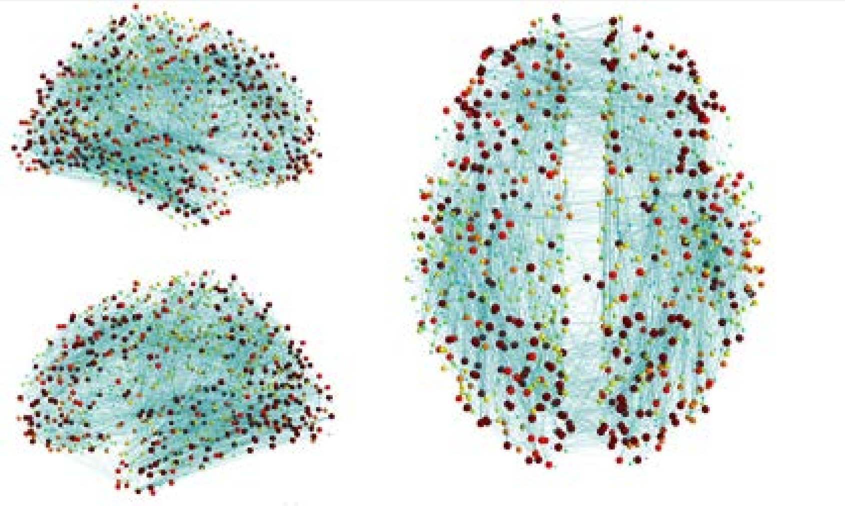


Web graph from the SNAP database (http://snap.stanford.edu/data)

Scale of Real-World Graphs

Brain scale …

- 100 billion vertices, 100 trillion edges
- 2.84 PB adjacency list



Human connectome. Gerhard et al., Frontiers in Neuroinformatics 5(3), 2011