Graphs and the V's of Big Data

Graphs and the V's of Big Data

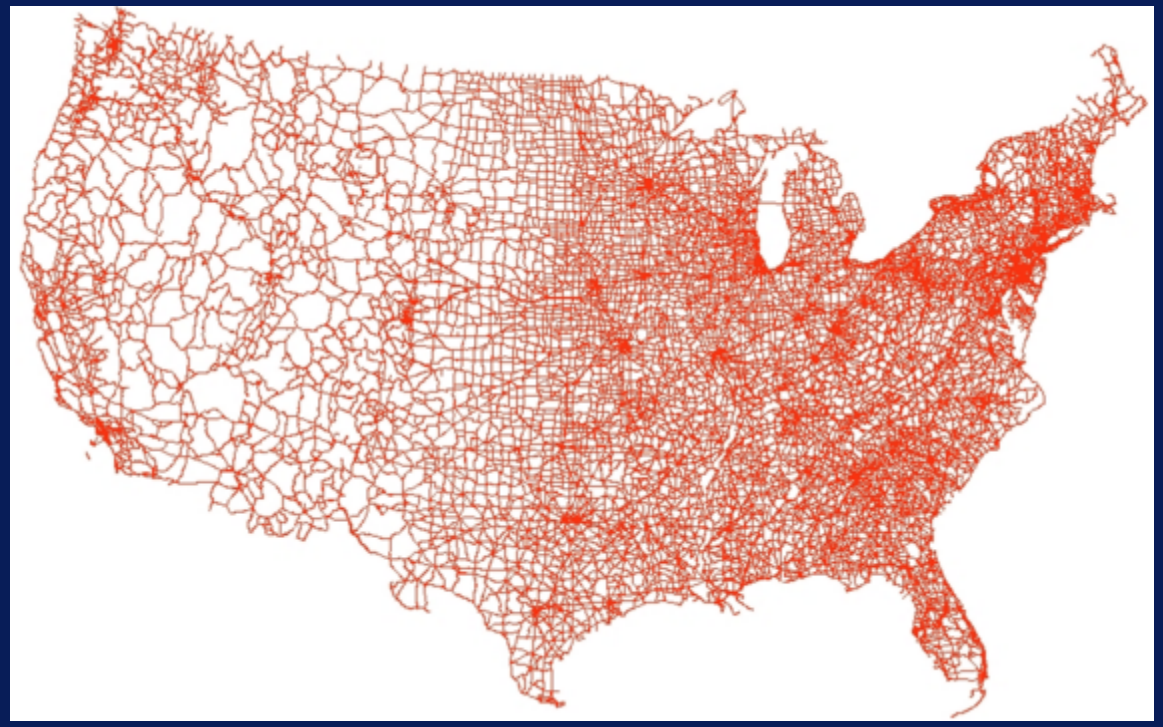
- **Concepts You Will Learn**
  - Algorithmic Complexity
  - Hard decision problems
  - Outcome metrics
  - Streaming edges

Graphs and the V's of Big Data

- Volume
- Velocity
- Variety
- Valence

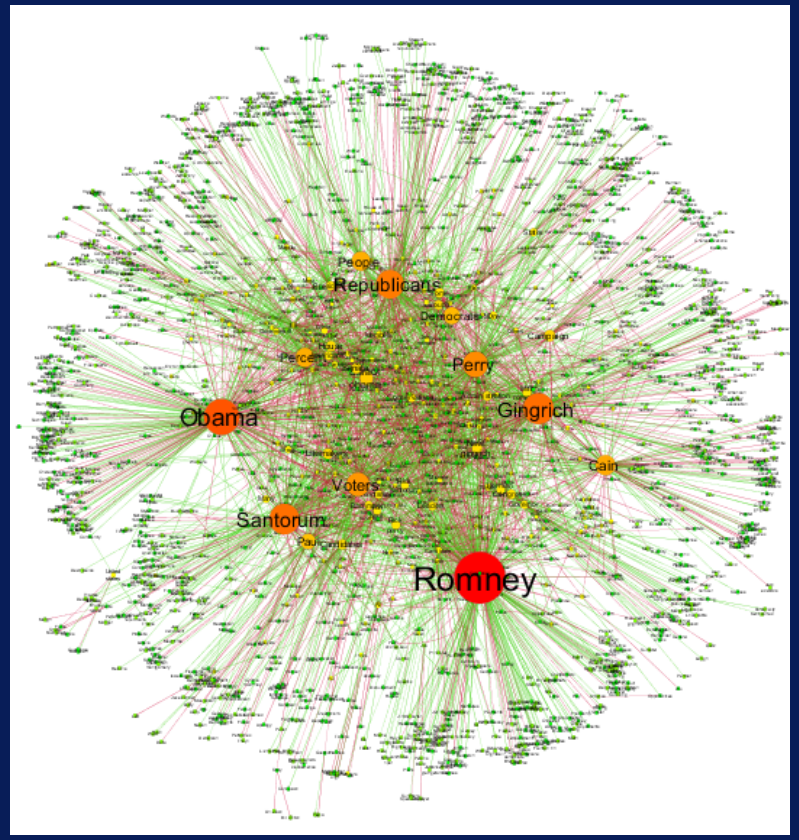
Volume

- **Size**
  - # of nodes and edges



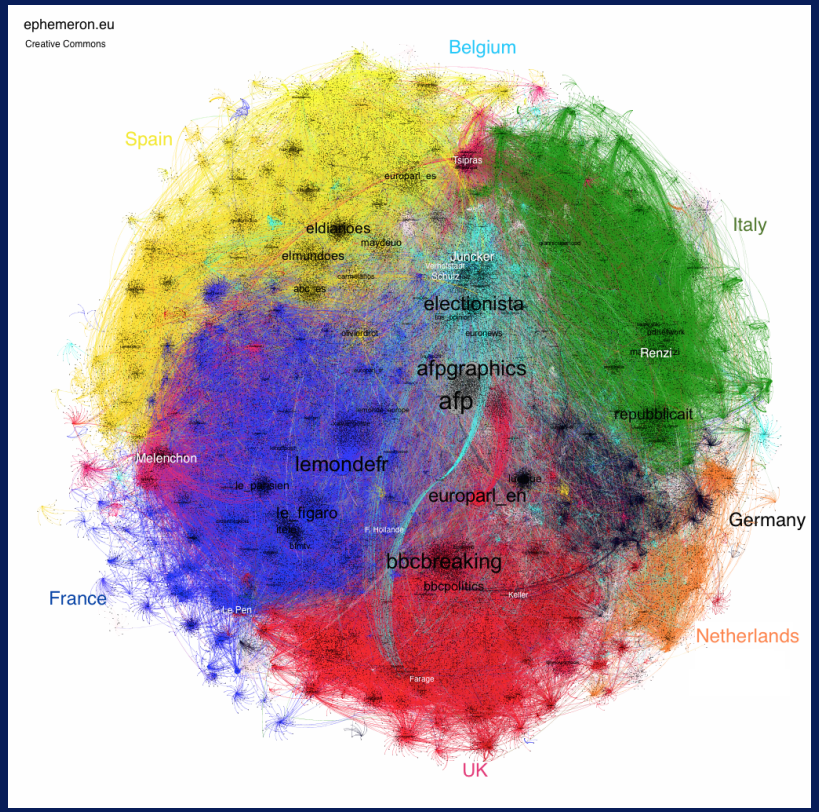
Volume

- **Size**
  - # of nodes and edges



Volume

- **Size**
  - # of nodes and edges



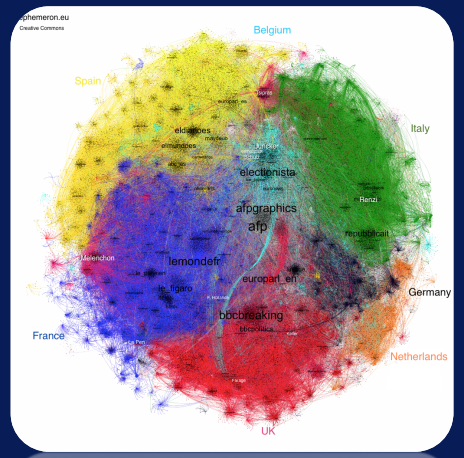
Importance of Volume

- Memory limitations
- Impact on Analytics

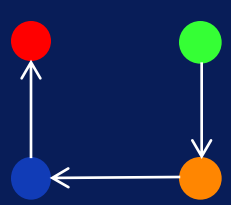
Velocity

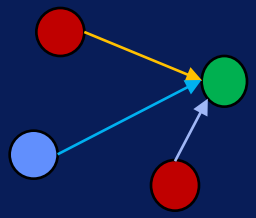
Velocity

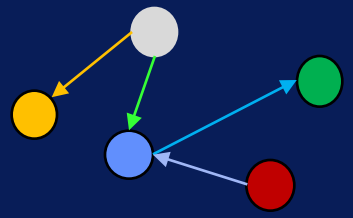

Time


Belgium
Spain
Italy
electionista
afpgraphics
afp
lemondefr
europarl_en
bbcbreaking
Germany
France
Netherlands
UK

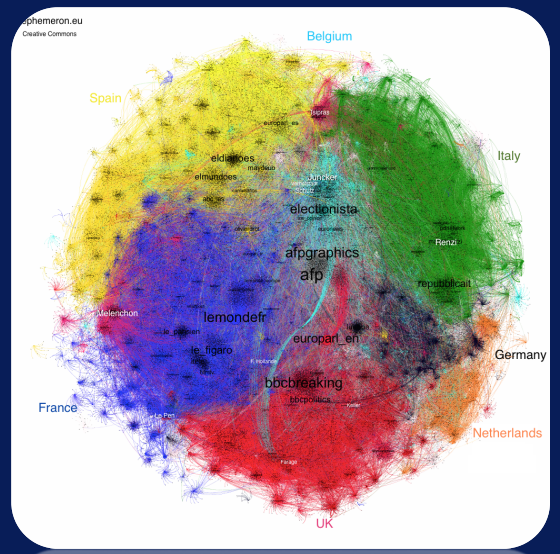
Velocity







Time



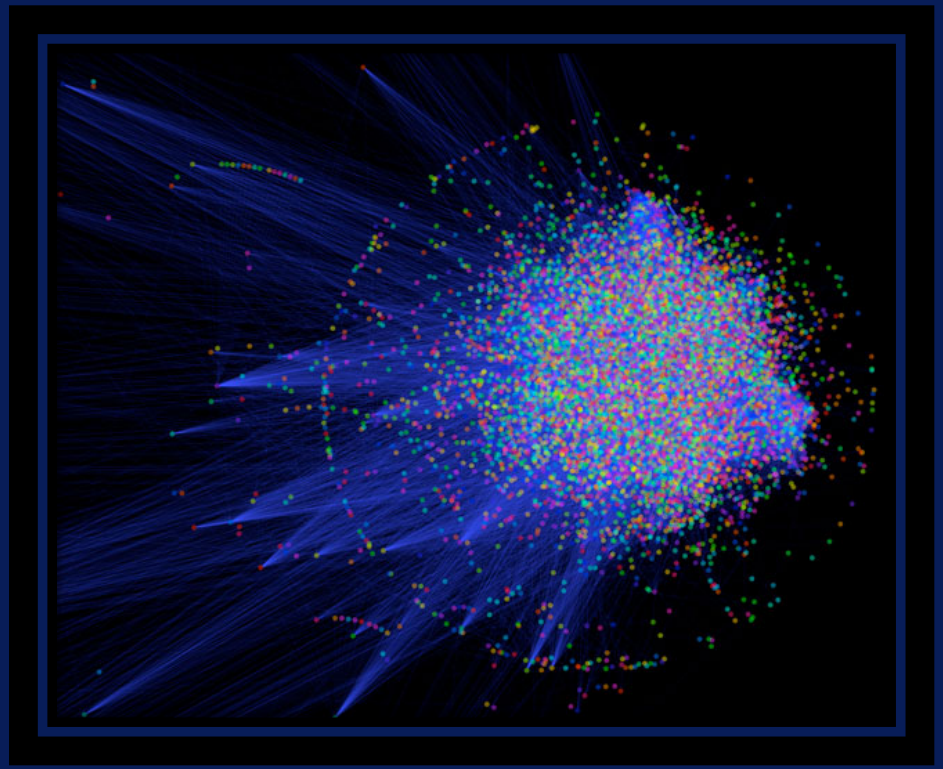
Velocity

- The rate at which a graph increases in size and complexity

Streaming edges into graphs

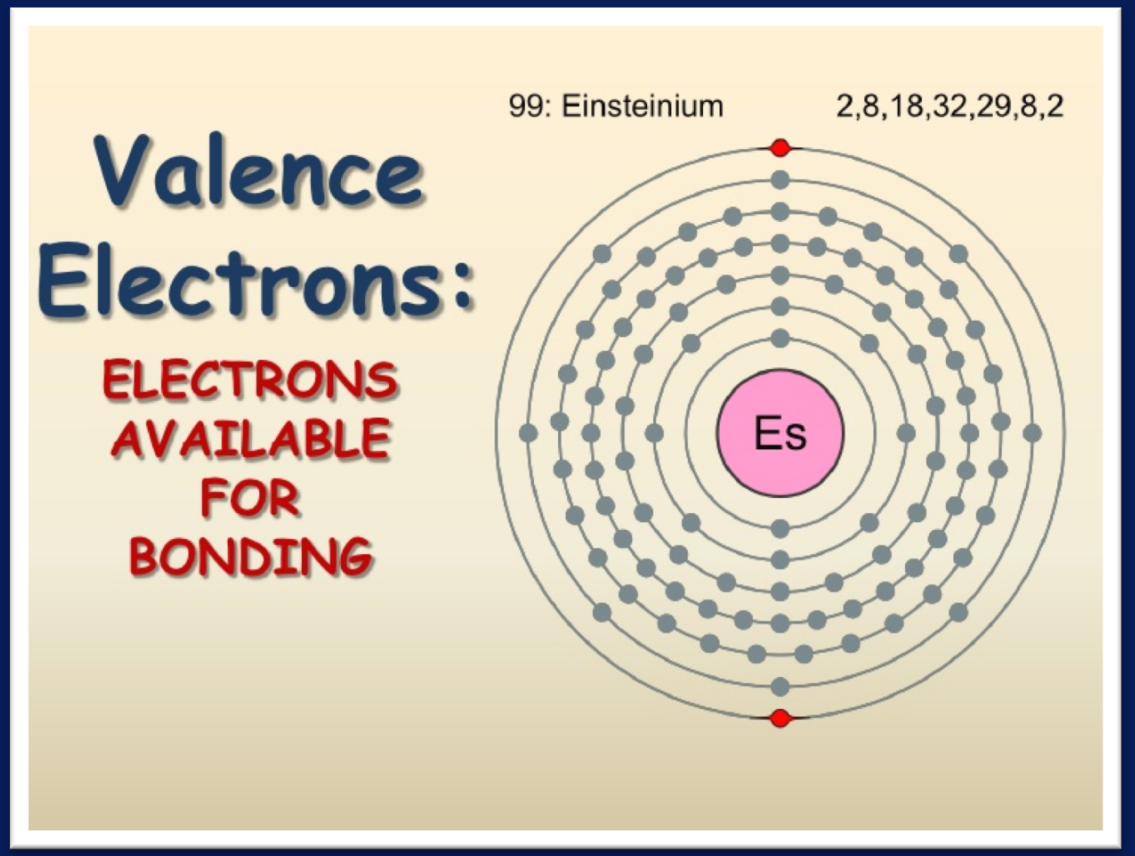
Variety

- Differences in the types and sources of data
- More non-uniform and complex



Valence

- In Chemistry



Valence

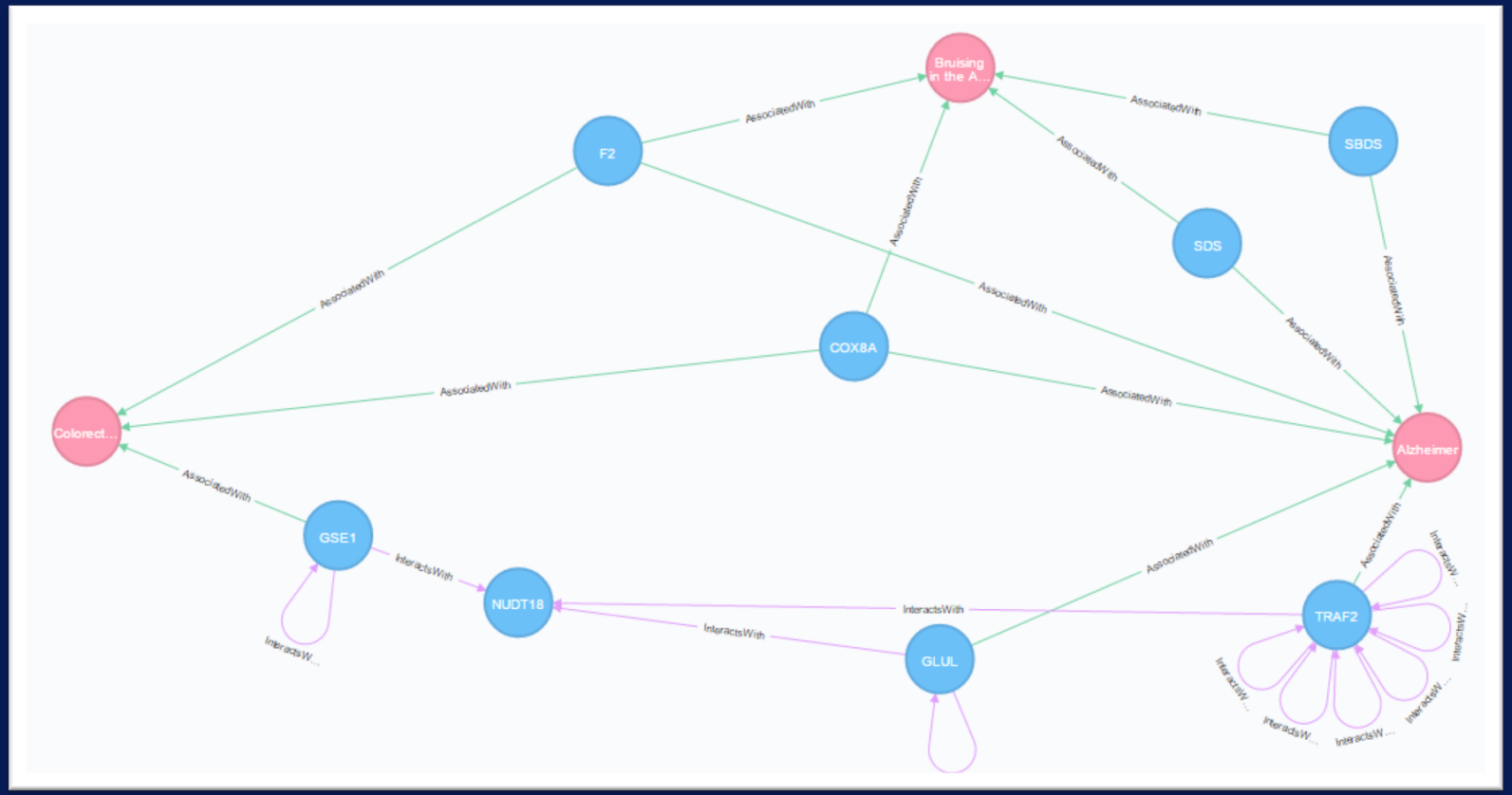
- **In Graphs**
  - A measure of **connectedness**

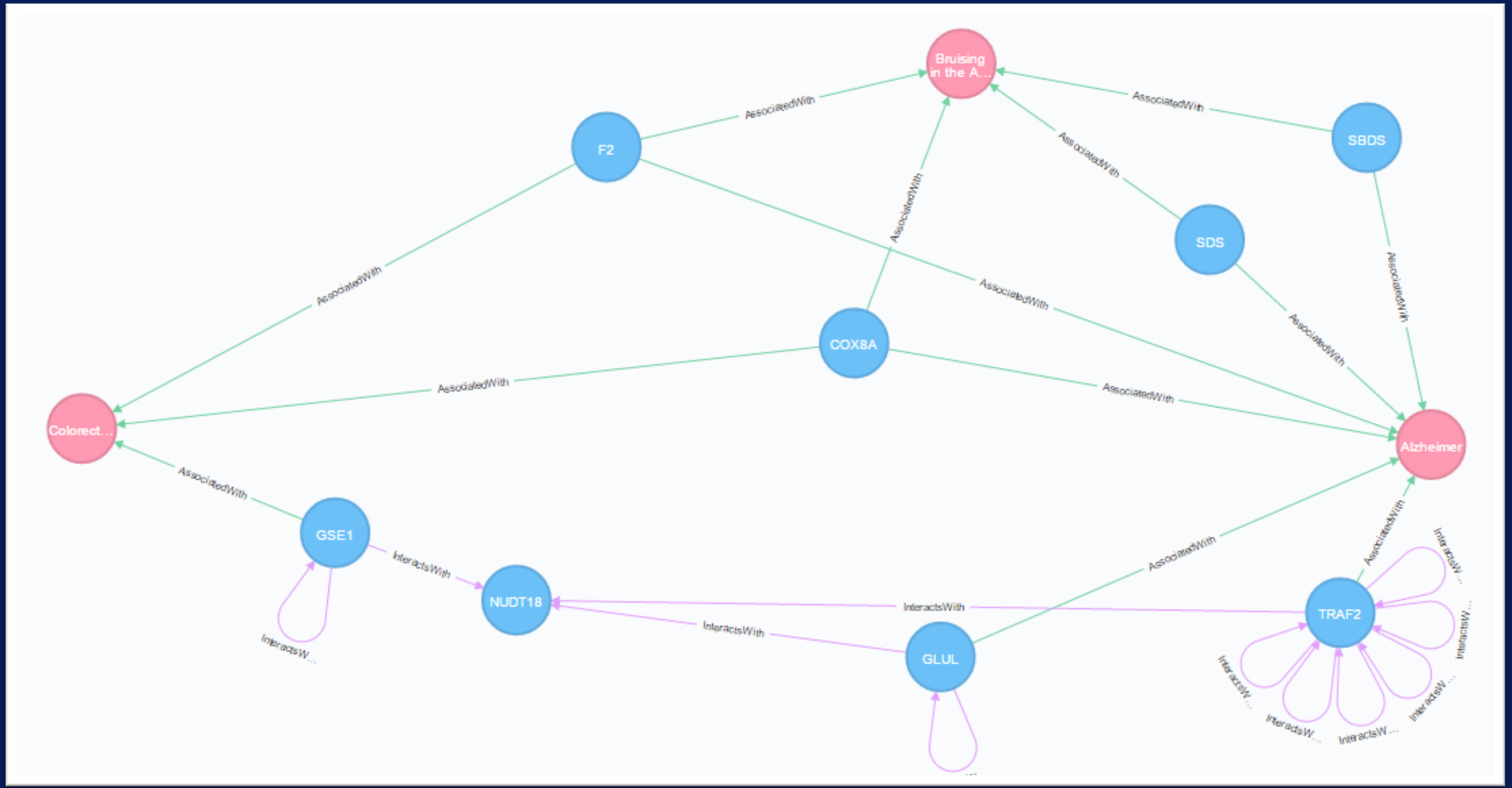
How Volume Impacts Graph Analytics

- Increases Algorithmic complexity
- Data-to-analysis time is too high

How Volume Impacts Graph Analytics

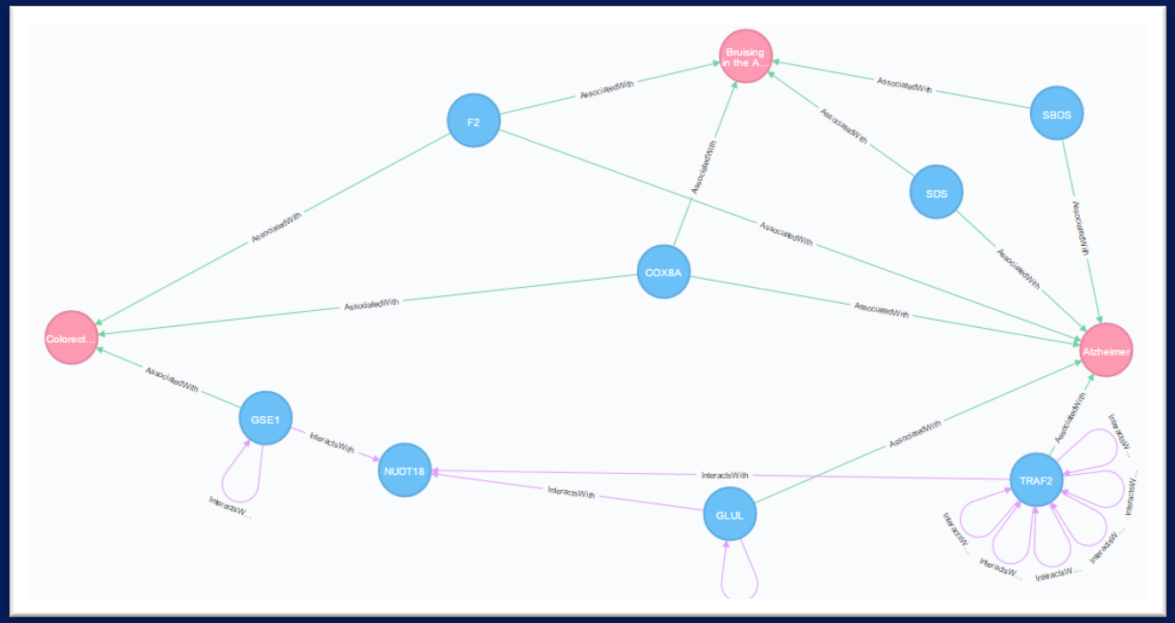
- **An example**
  - Problem: Find a Simple Path between “Alzheimer’s Disease” and “Colorectal Cancer




Bruising in the A...
F2
SBDS
SDS
COX8A
Colorect...
Alzheimer
GSE1
NUDT18
GLUL
TRAF2
AssociatedWith
InteractsWith

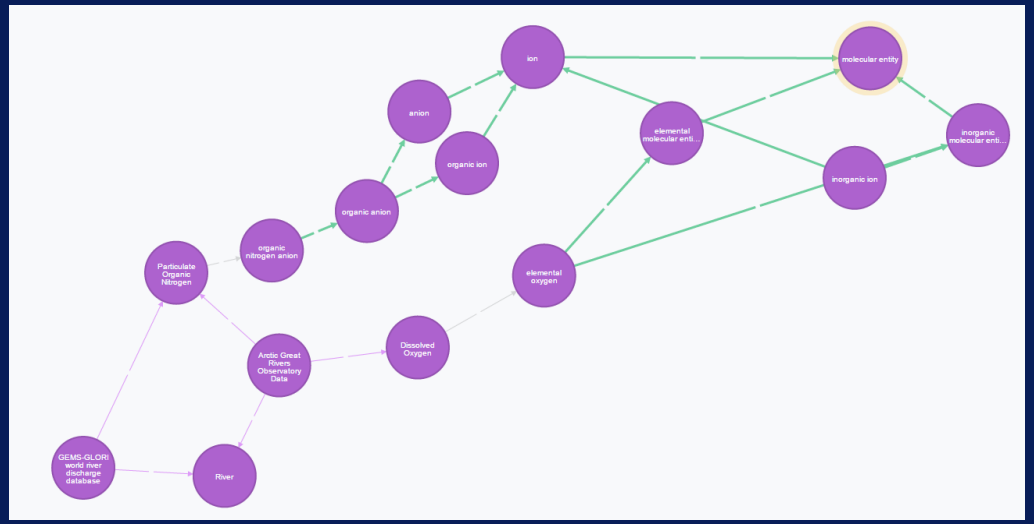
How Volume Impacts Graph Analytics

- Is there a simple path between “Alzheimer’s Disease” and “Colorectal Cancer”?
  - Well-known hard decision problem
- How many simple paths exist between “Alzheimer’s Disease” and “Colorectal Cancer”?
  - Size of result is *exponential in the number of nodes*



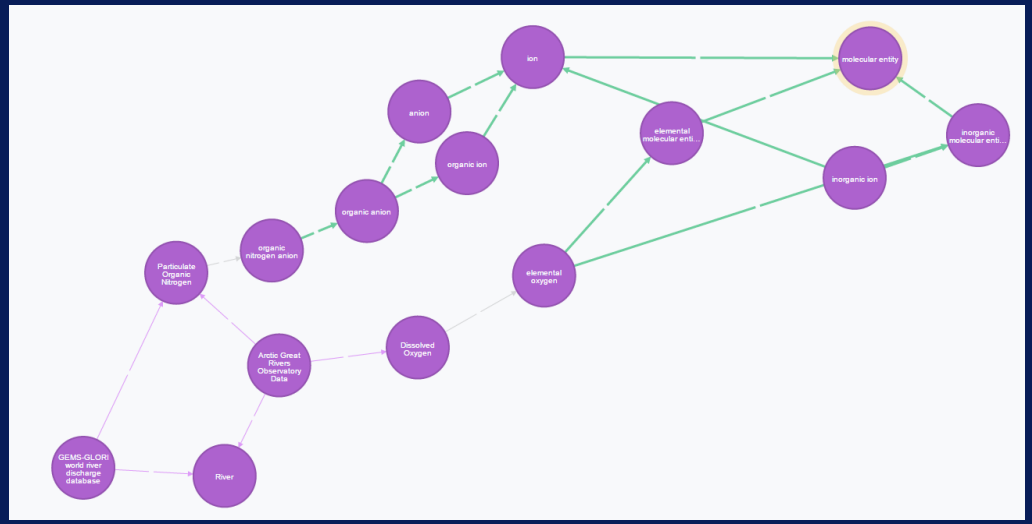
How Velocity Impacts Graph Analytics

- **Social Media**
  - Stream of Updates = Stream of Edges



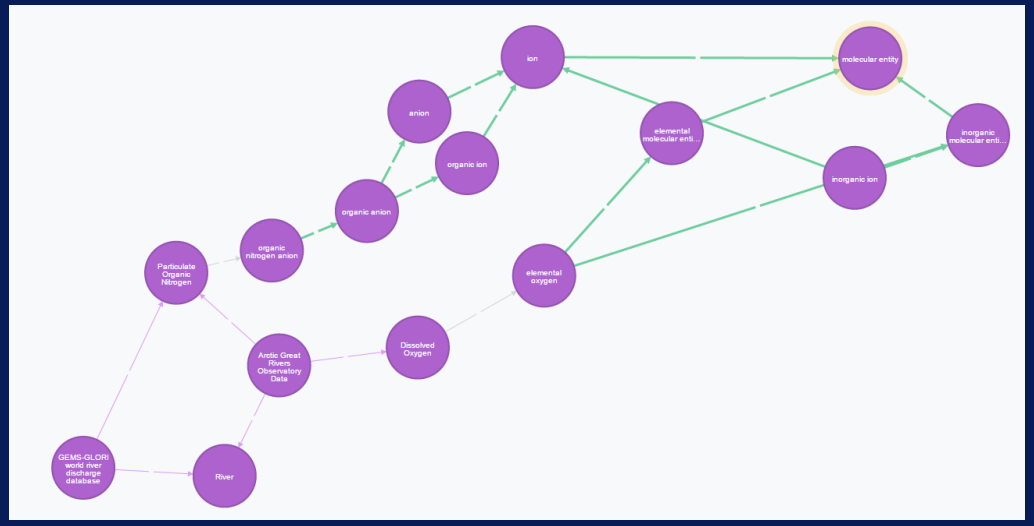
How Velocity Impacts Graph Analytics

- **Computing a Metric**
  - Shortest distance between two nodes
  - Count of strongly connected groups, e.g. a Facebook group



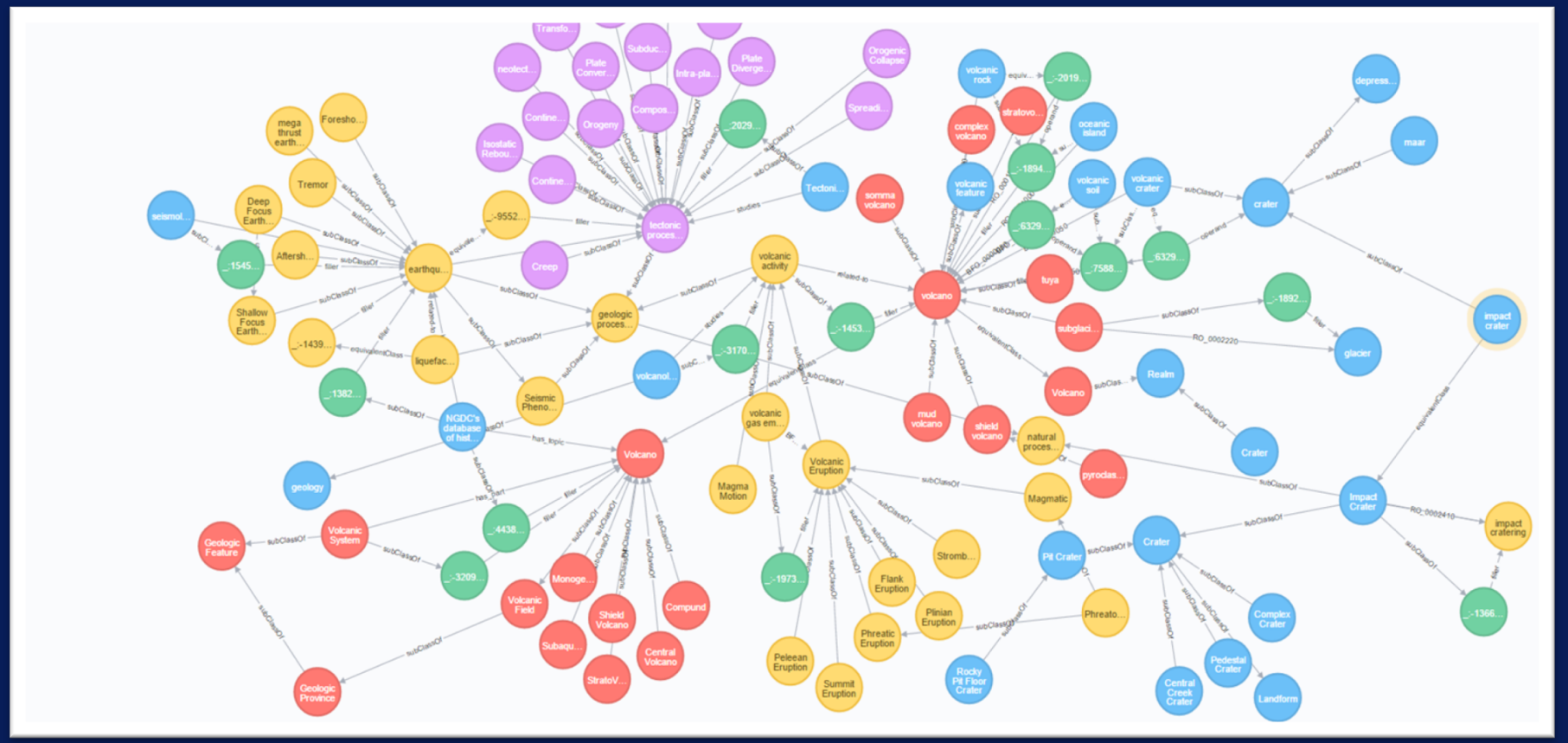
Velocity

- **A continuous stream does not fit in memory**
  - How can these metrics be computed on the edge-stream with limited memory?



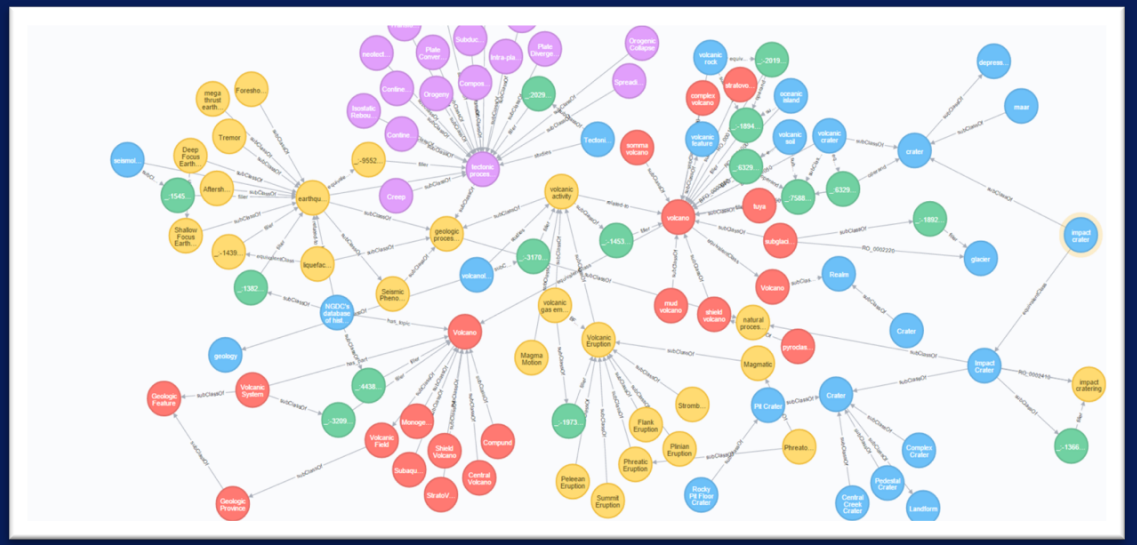
Variety…

…aka Heterogeneity



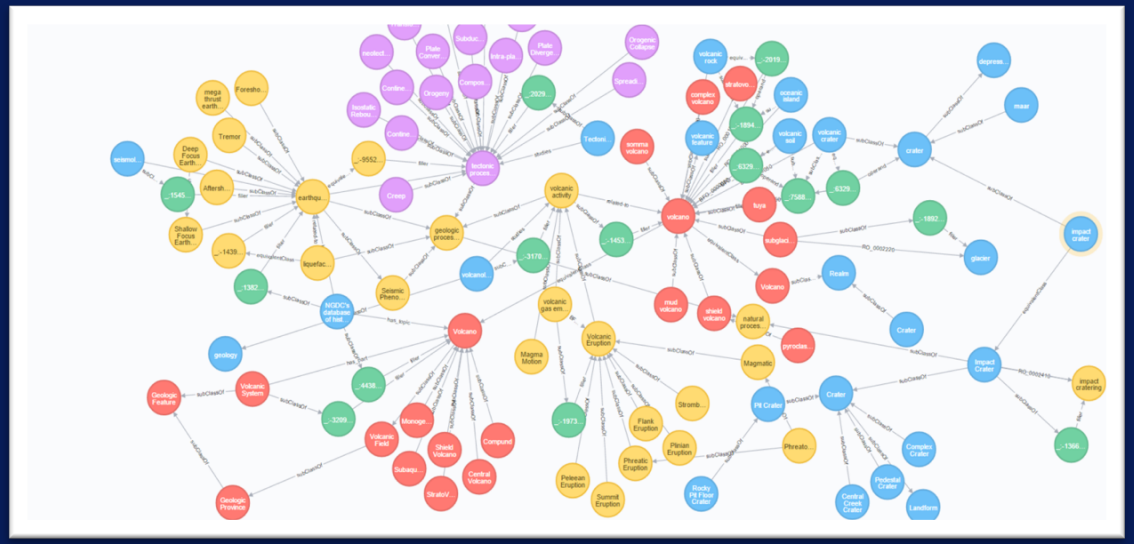
Variety

- Two aspects to consider



Variety

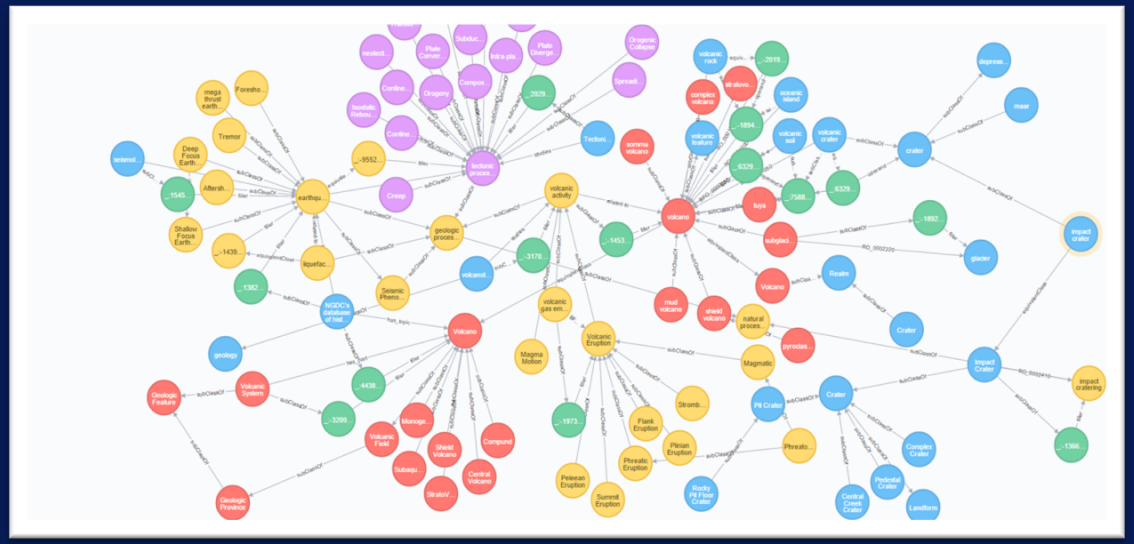
- Graph data is often created through integration



Variety

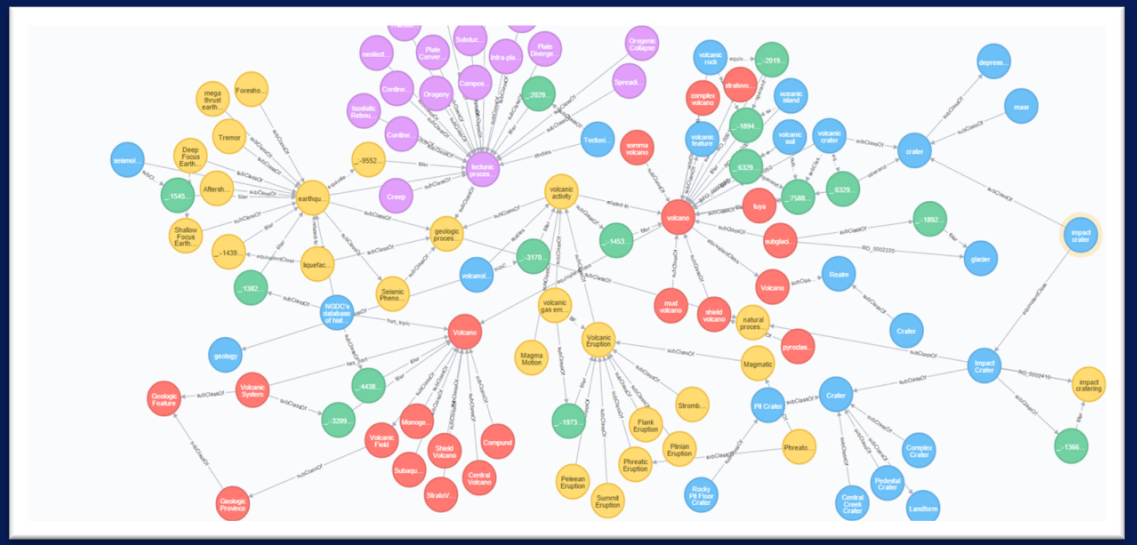
- Graph data is often created through integration

Relational



Variety

- Graph data is often created through integration

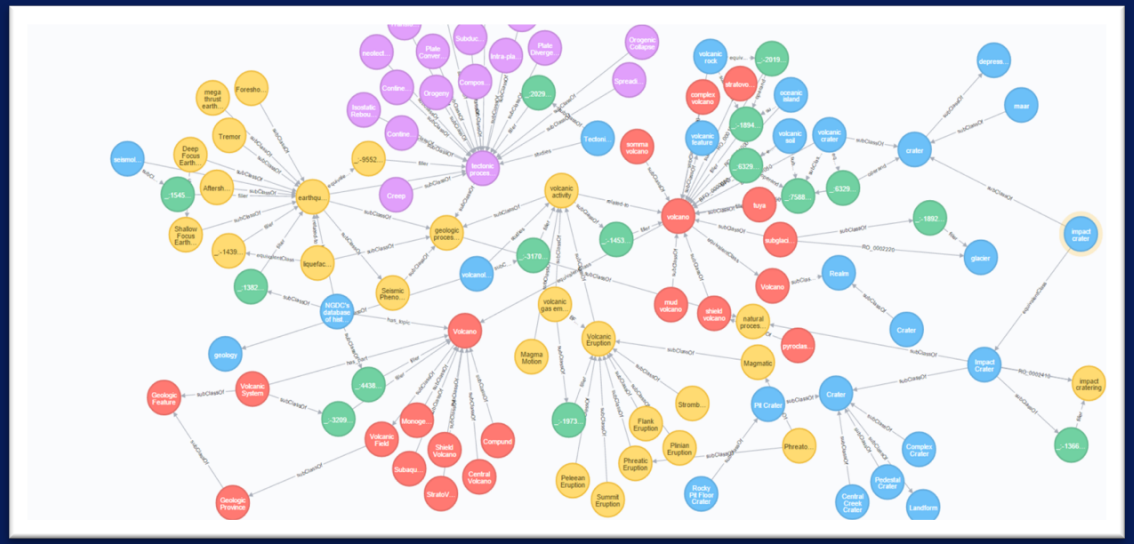


Relational

XML/JSON

Variety

- Graph data is often created through integration



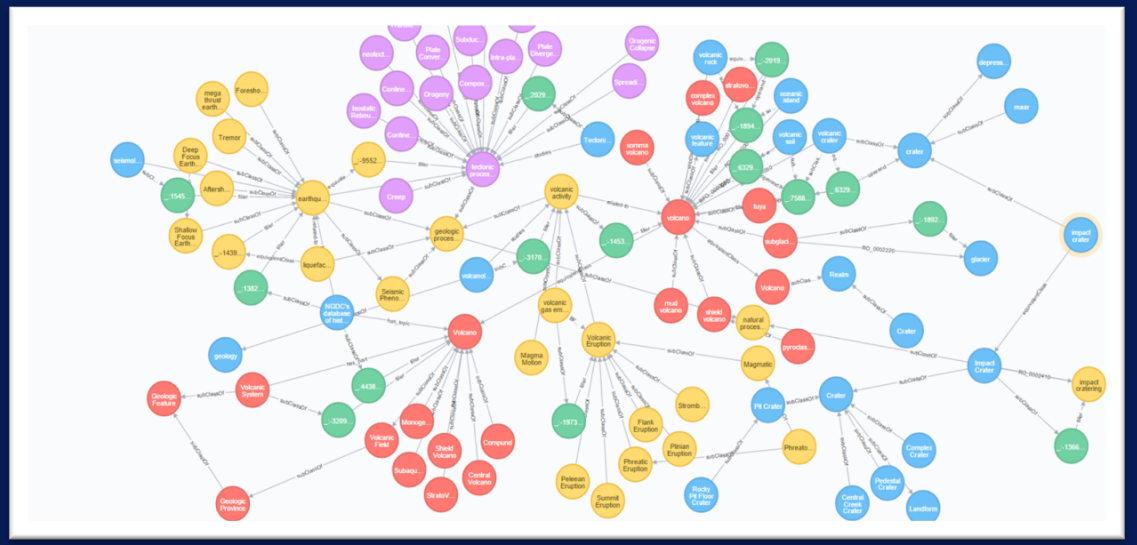
Relational

XML/JSON

Graph-structured

Variety

- Graph data is often created through integration



Relational

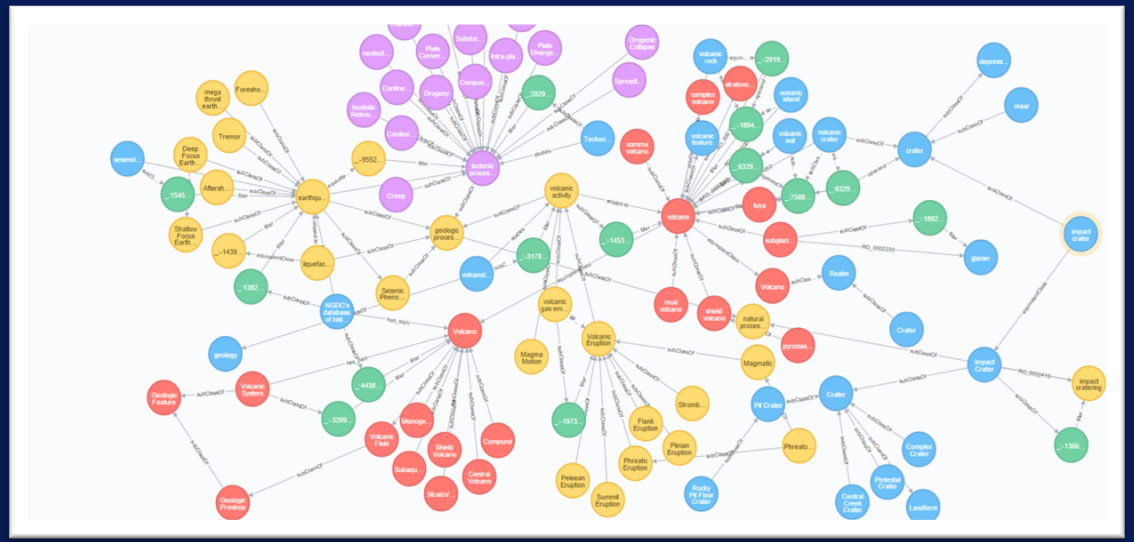
XML/JSON

Graph-structured

Document

Variety

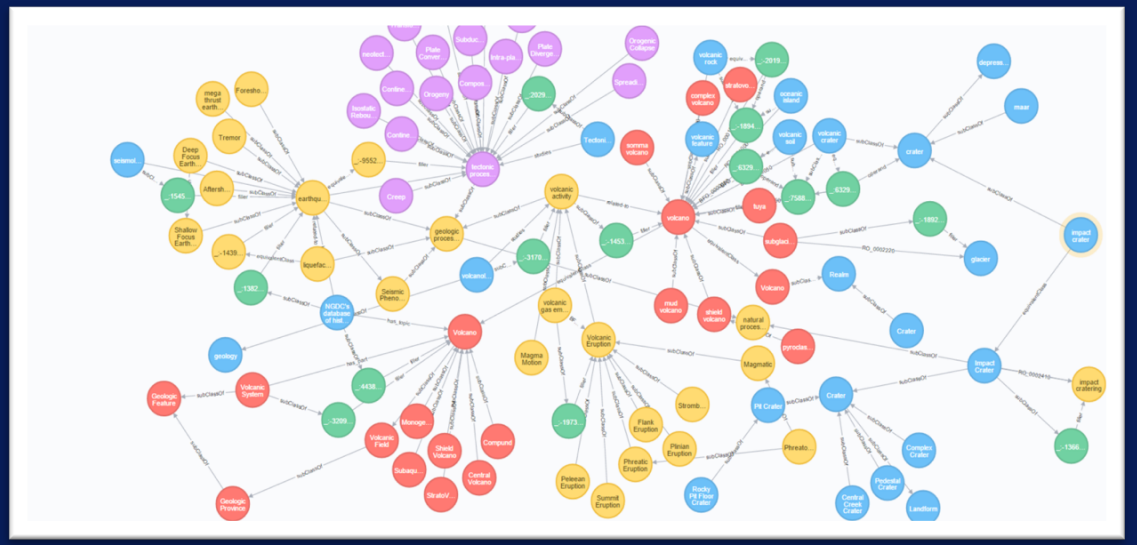
- Different kinds of graphs may have very different meanings



Variety

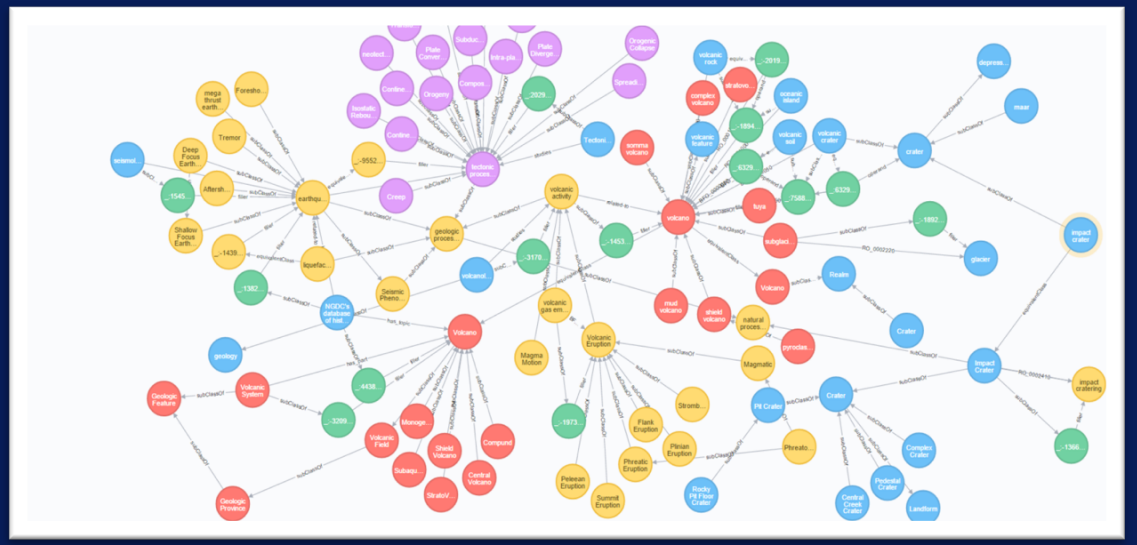
- Different kinds of graphs may have very different meanings

Social Networks



Variety

- Different kinds of graphs may have very different meanings

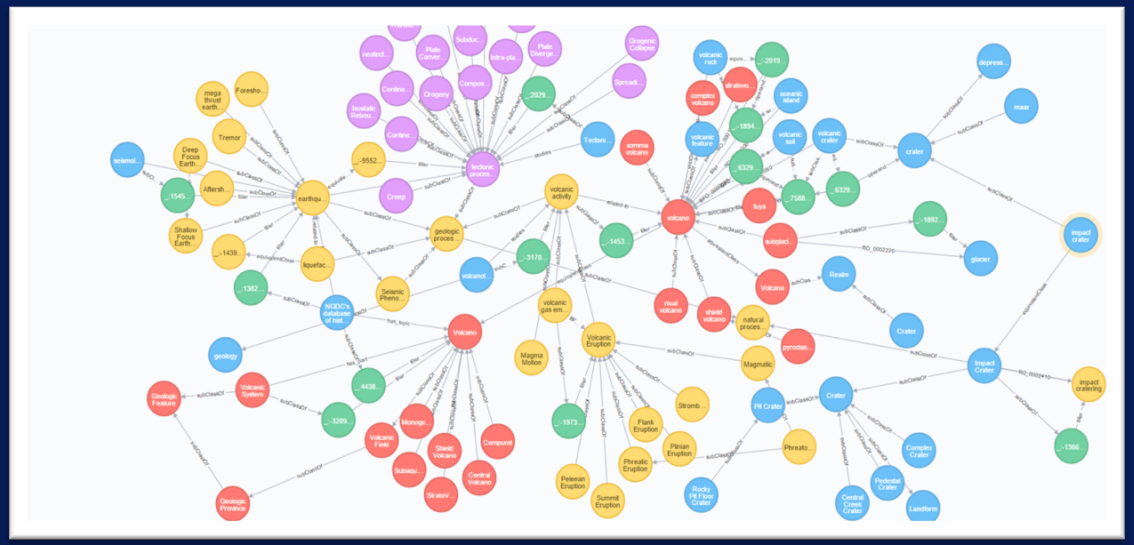


Social Networks

Citation Networks

Variety

- Different kinds of graphs may have very different meanings



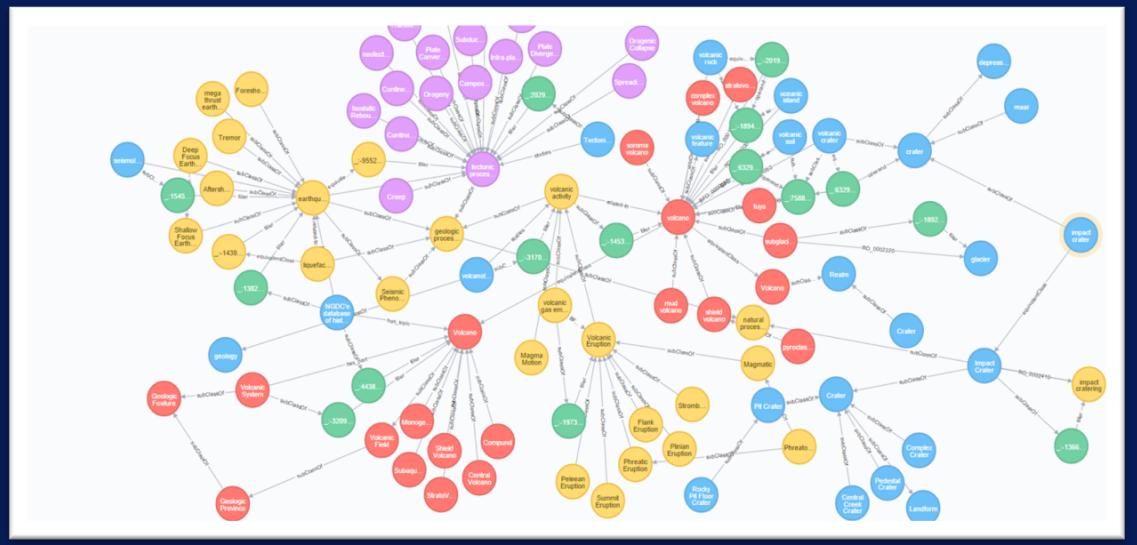
Social Networks

Citation Networks

Interaction Networks

Variety

- Different kinds of graphs may have very different meanings



Social Networks

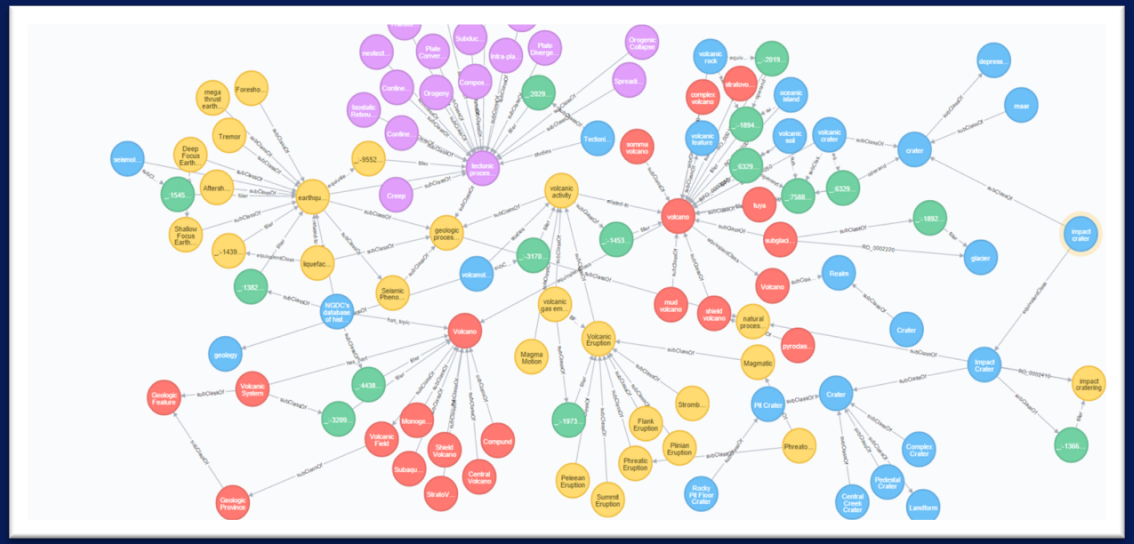
Citation Networks

Interaction Networks

Semantic Web/
Linked Data

Variety

- Different kinds of graphs may have very different meanings



Social Networks

Citation Networks

Interaction Networks

Semantic Web/
Linked Data

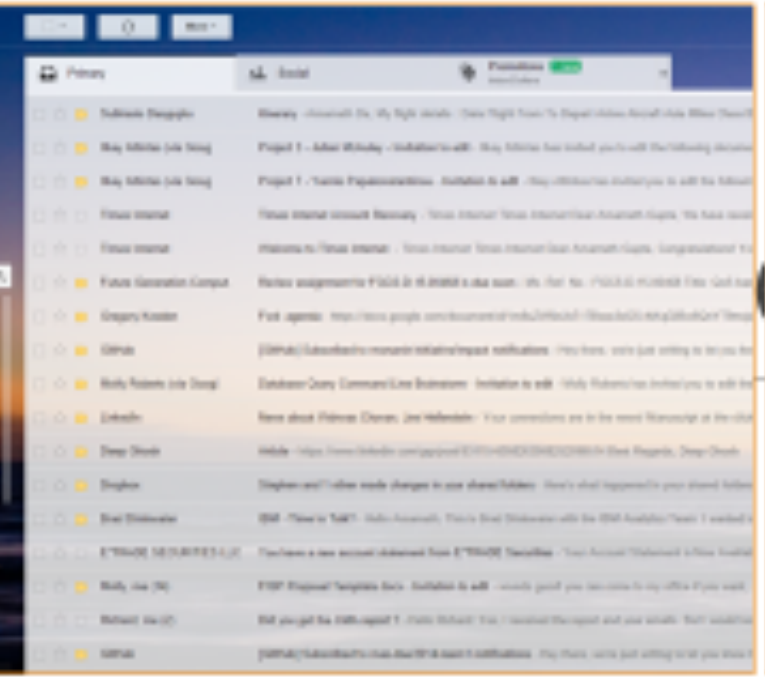
Ontologies

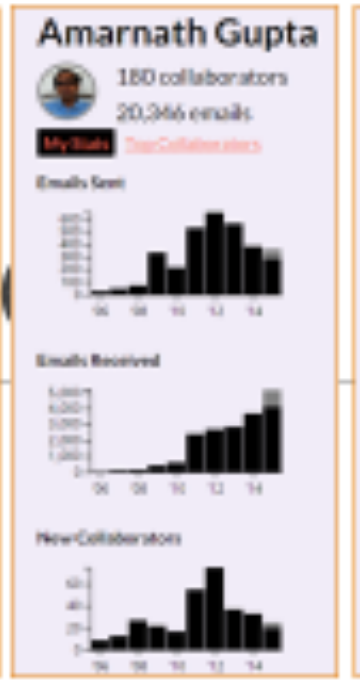
Valence

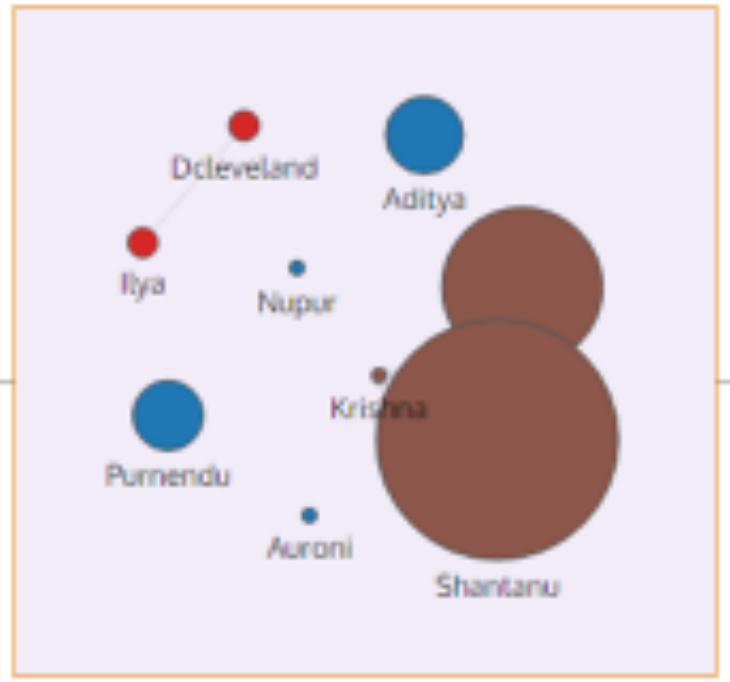
- **Degree of connectedness or interdependence**
  - Higher valence implies that the data elements are strongly related
  - This relatedness is significantly exploited in analytics

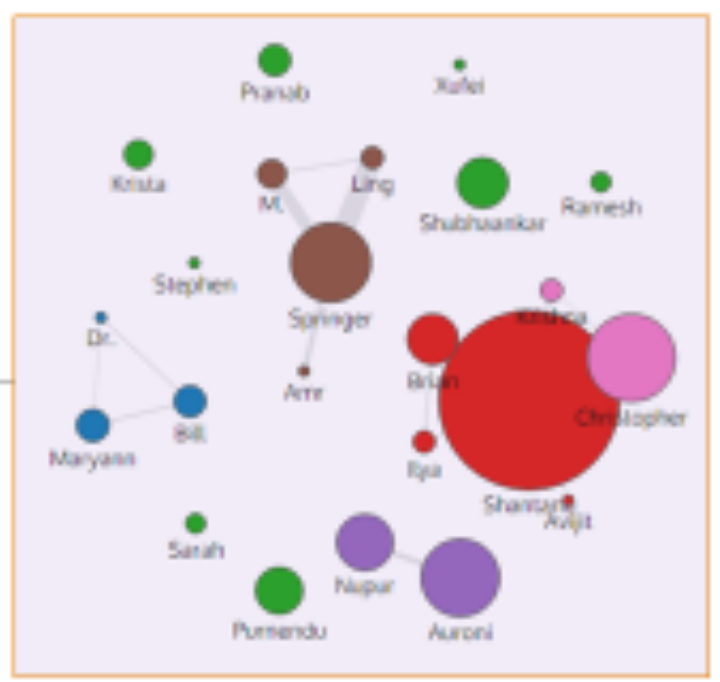
Valence

- **In many cases, valence increases with time**
  - Parts of the graph become denser
  - Average distance between arbitrary node pairs decreases




Amarnath Gupta
180 collaborators
20,366 emails
My Stats
Top Collaborators
Emails Sent
Emails Received
New Collaborators


Dcleveland
Aditya
Ilya
Nupur
Krishna
Purnendu
Auroni
Shantanu


Pranab
Rafel
Krista
M.
Ling
Shubhasankar
Ramesh
Stephen
Springer
Dr.
Anne
Maryann
Bill
Ilya
Christopher
Sarah
Nupur
Purnendu
Auroni