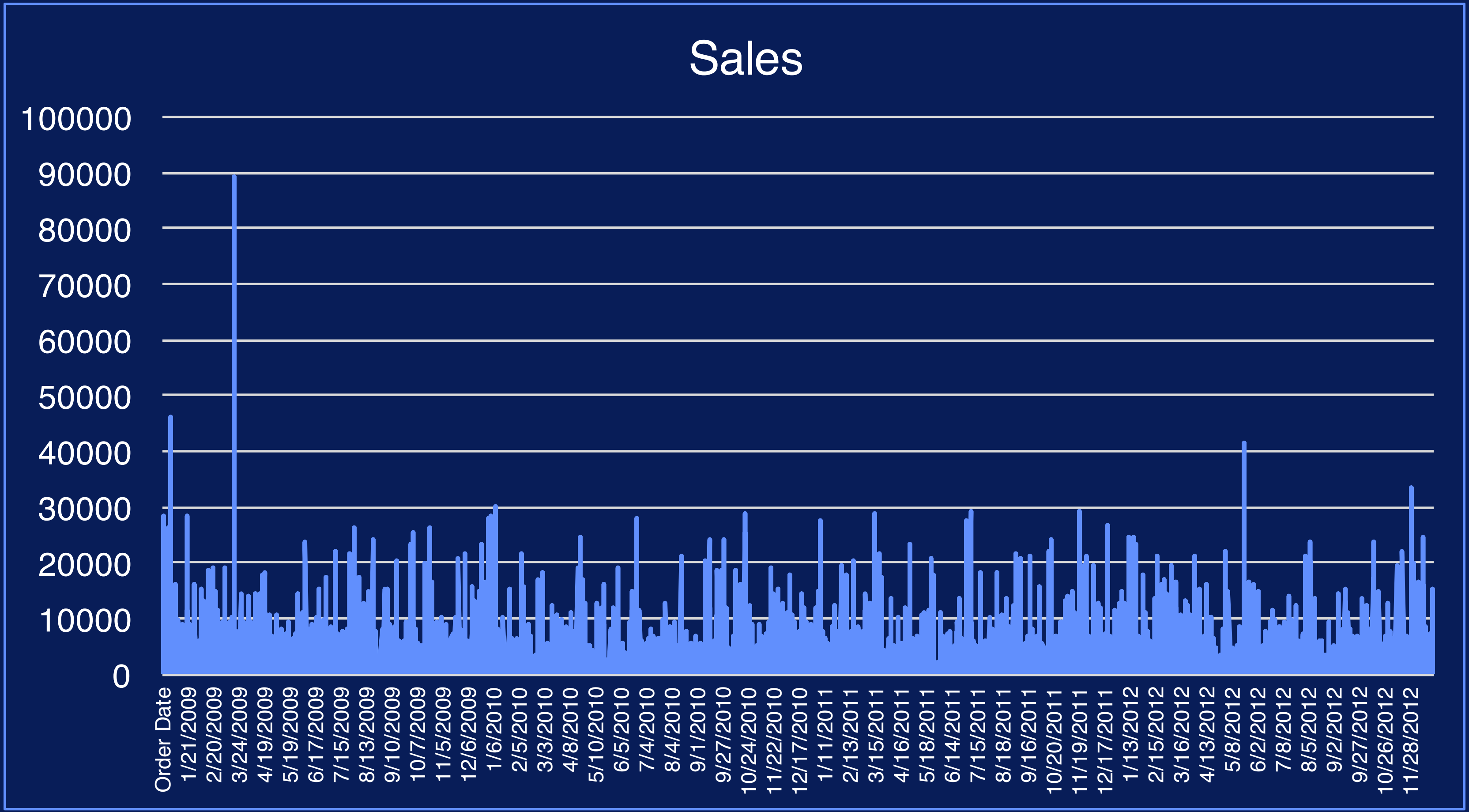
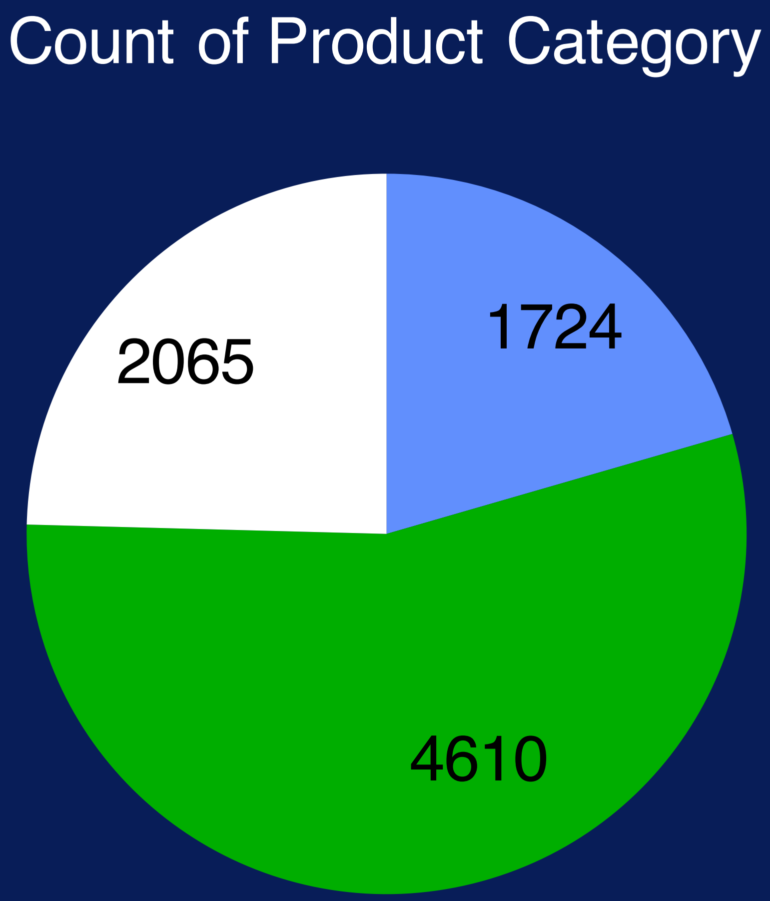
What is a Graph?

Is this a Graph?

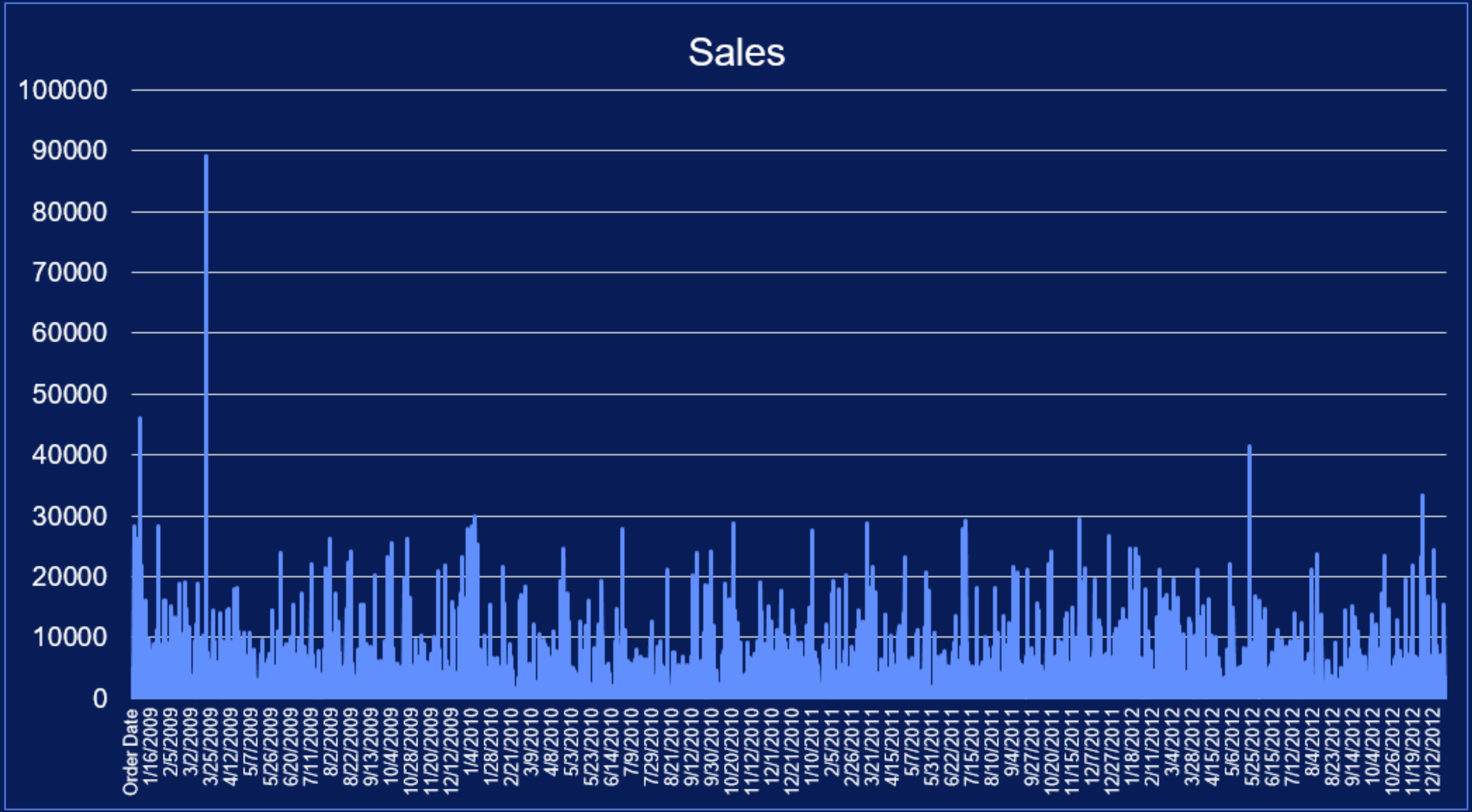


Is this a Graph?



What is NOT a Graph?

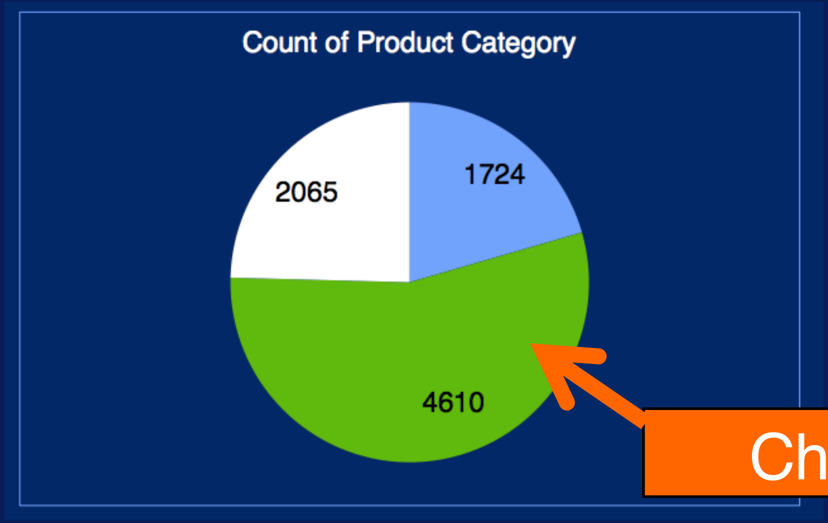
- **This is a chart!**



Graph vs. Graph of a Function

- **Charts represent the *graph of a function***

Function – Mapping from **Category** to **Total**



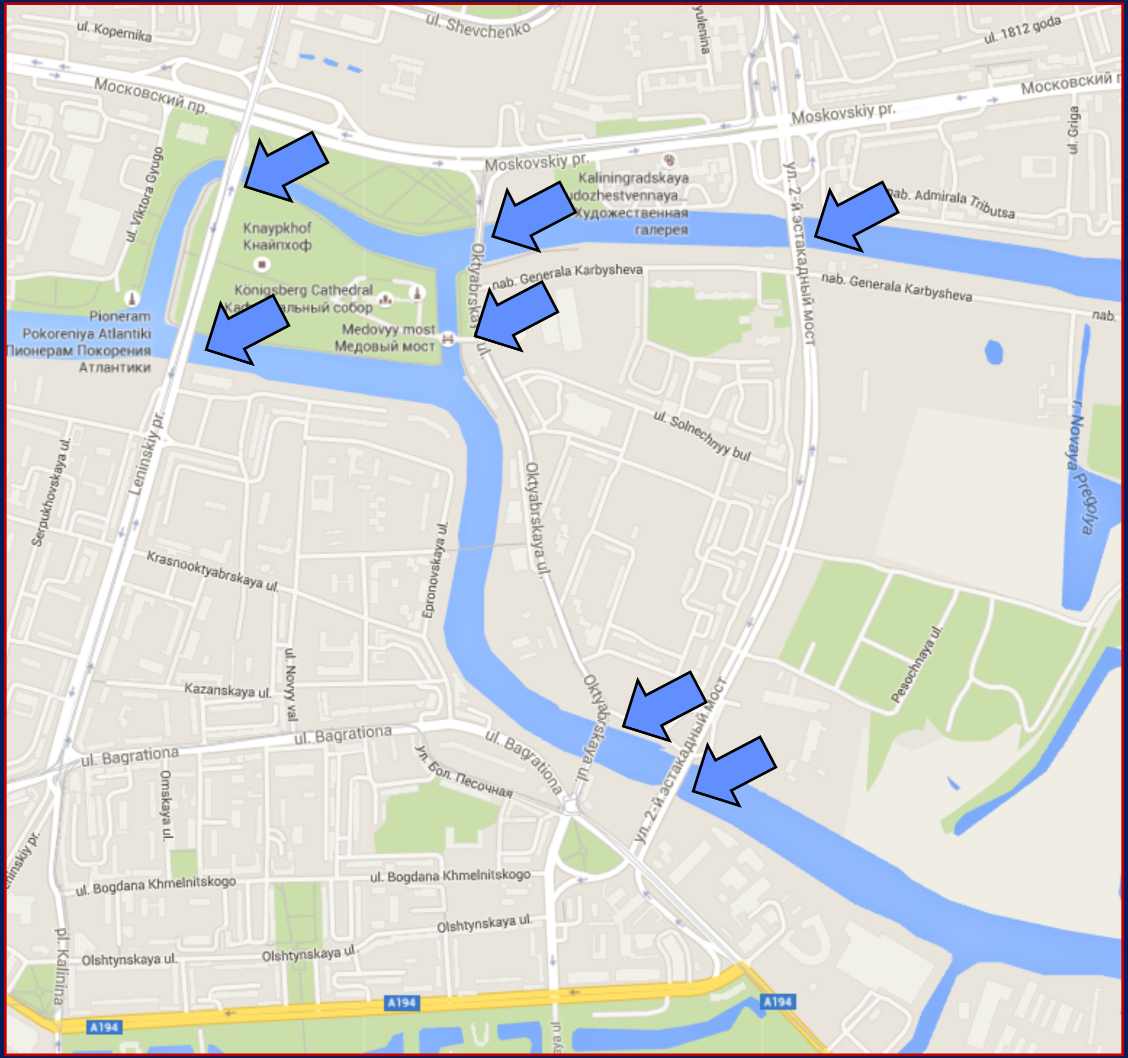
Chart

Product Category →	Total
Furniture	1724
Office Supplies	4610
Technology	2065

Graph Analytics is based on Graph Theory

The city of Königsberg in Prussia (now Kaliningrad, Russia)

Graph Theory Came From An Urban Planning Problem!!



Google Map of Kaliningrad, Russia

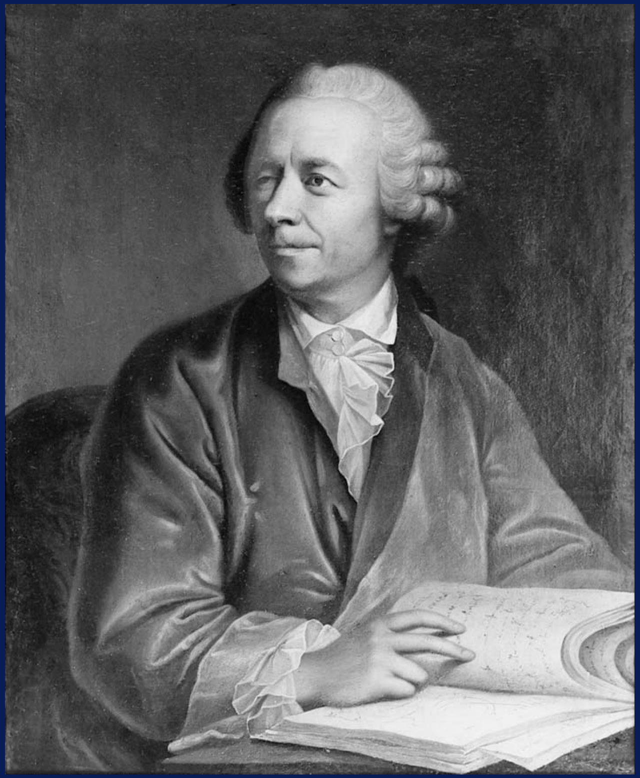
7 Crossing Points on the Pregel River

Devise a walk through the city that would cross each bridge once and only once

Mathematician Euler solved the problem in 1736

Graphs: Defined Differently by Different People

Mathematicians



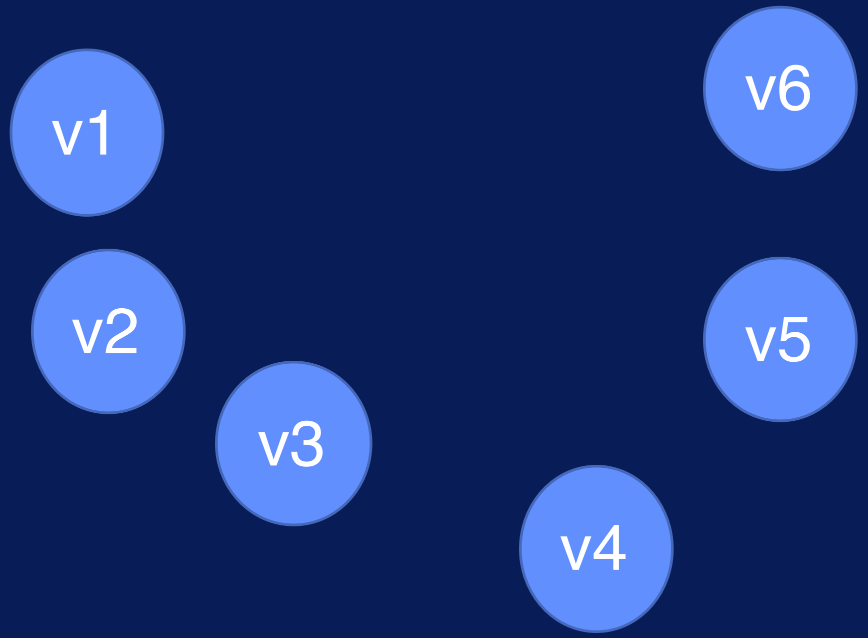
Leonhard Euler

Computer Scientists

Edsger W. Dijkstra

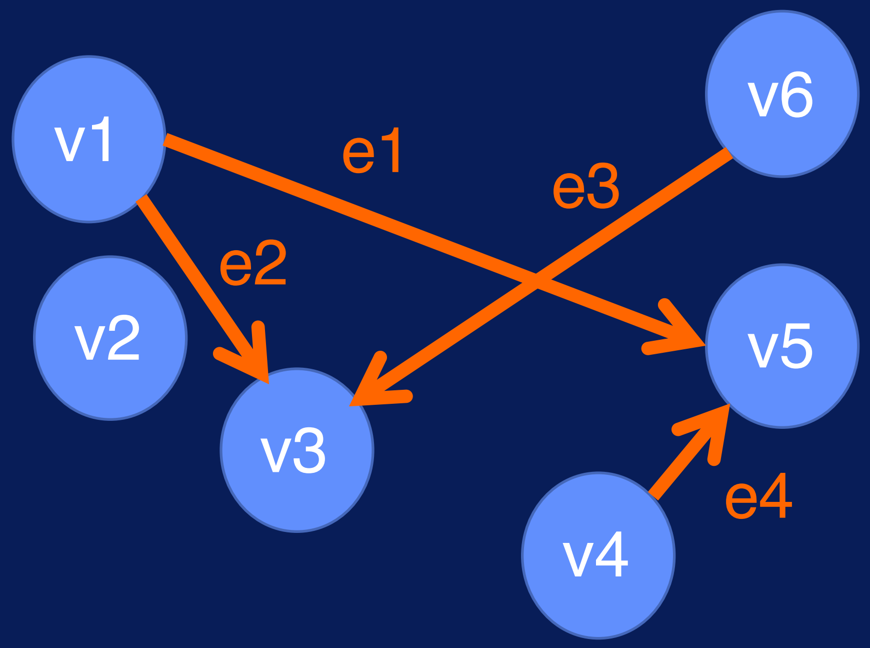
Mathematical Definition:

- V: a set of vertices



Mathematical Definition:

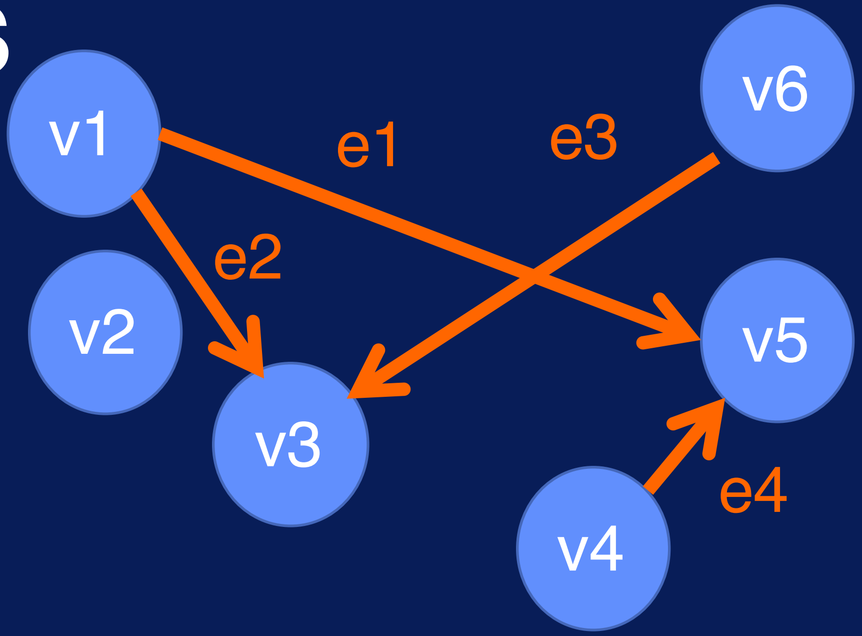
- **V: a set of vertices**
- **E: a set of edges**
  - E = {e1, e2, e3, e4}



Mathematical Definition:

- **V: a set of vertices**
- **E: a set of edges**

E = {e1, e2, e3, e4}



e: an edge designates a pair of vertices

E = {(v1,v5), (v1,v3), (v6,v3), (v4,v5)}

What about the Computer Science definition?

- **An abstract data type**

  1) Has a data structure to represent the mathematical graph

  2) Supports a number of operations (on that graph)

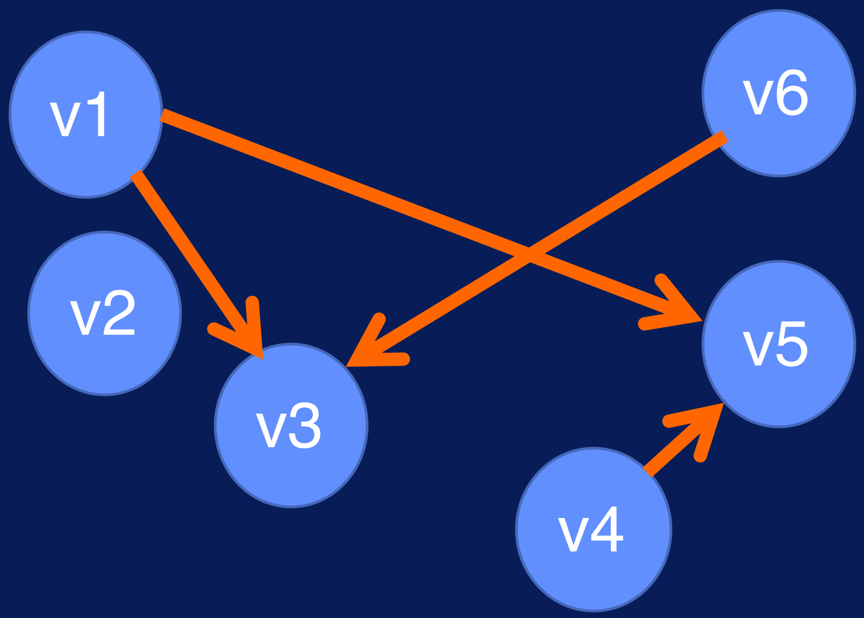
What about the Computer Science definition?

- **An abstract data type**

  1) Has a data structure to represent the mathematical graph

  2) Supports a number of operations (on that graph)
    - Add_edge
    - Add_vertex
    - Get_neighbor (and others)

One Computer Science Representation: Matrix

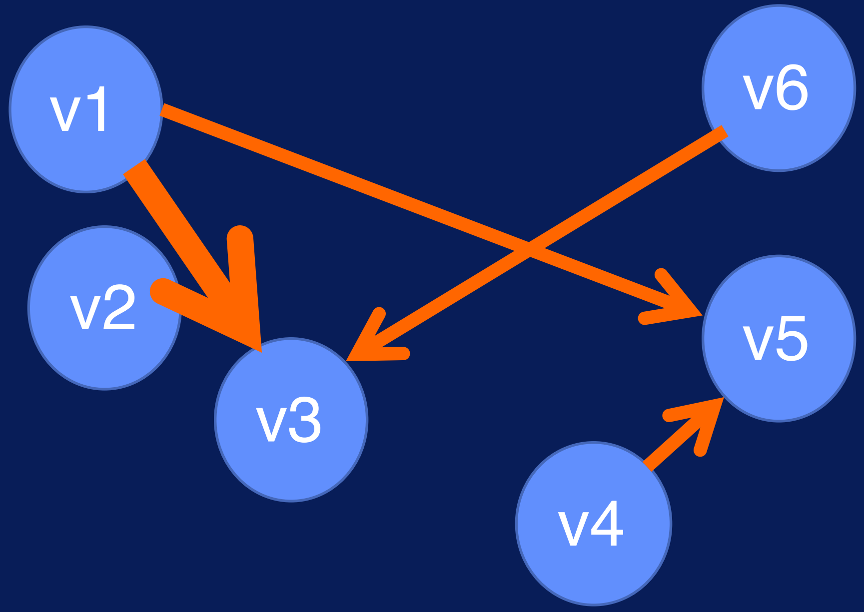


To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3						
v4					1	
v5						
v6			1			

One Computer Science Representation: Matrix



To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3						
v4					1	
v5						
v6			1			

What about the Computer Science definition?

- **An abstract data type**

1) Has a data structure …

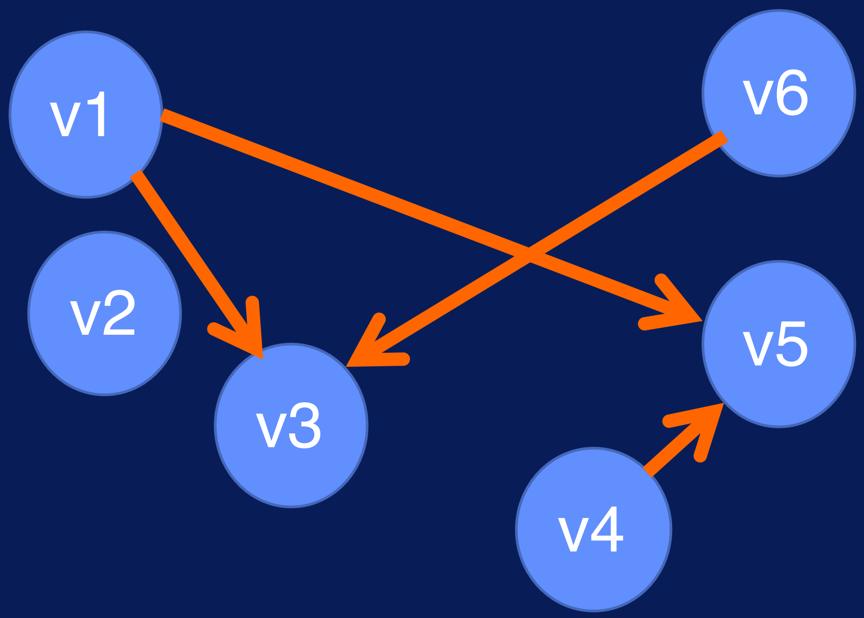
and

2) Supports a number of operations (on that graph)

- Add_edge
- Add_vertex
- Get_neighbor (and others)

Let's look at operations

A Basic Operation:
Add_edge



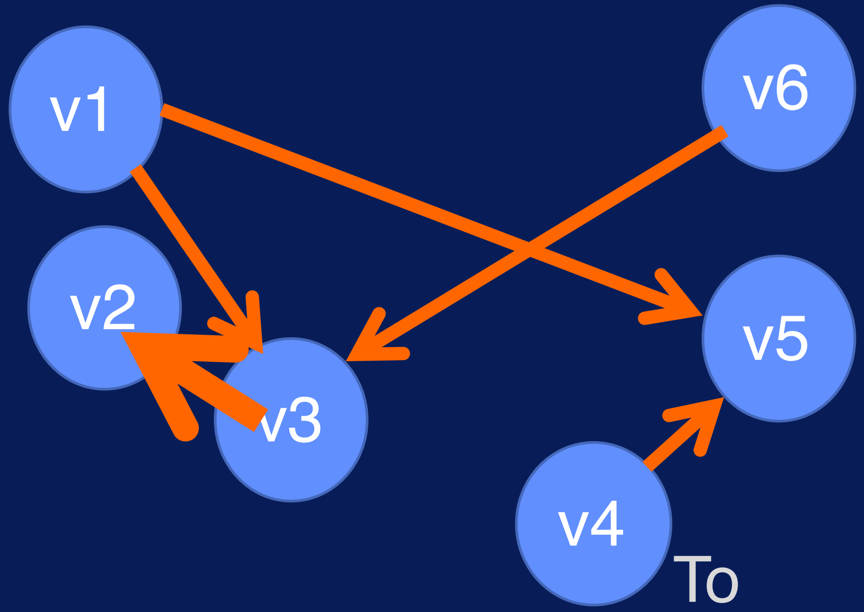
To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3						
v4					1	
v5						
v6			1			

Add_edge (v3,v2)

A Basic Operation:
Add_edge



To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3						
v4					1	
v5						
v6			1			

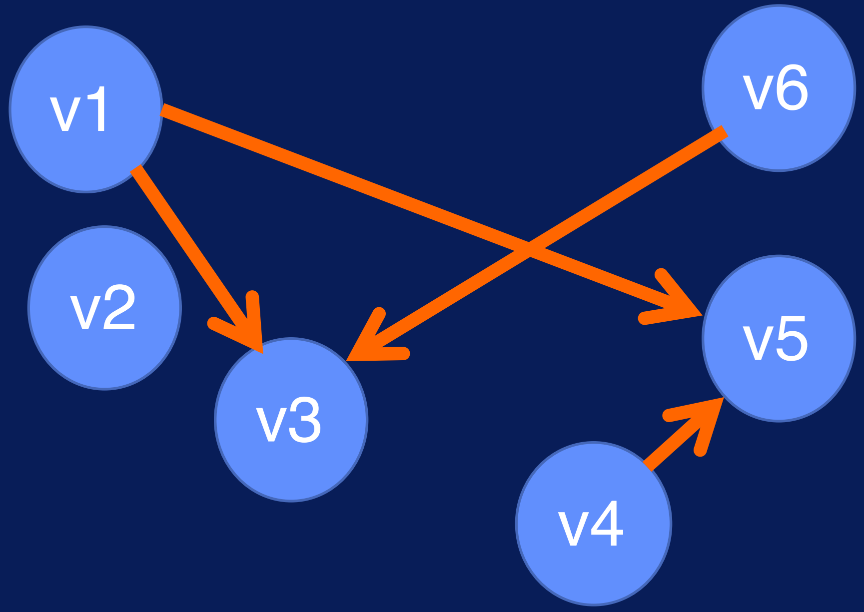
Add_edge (v3,v2)

To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3		1				
v4					1	
v5						
v6			1			

Another Operation:
Get_neighbor



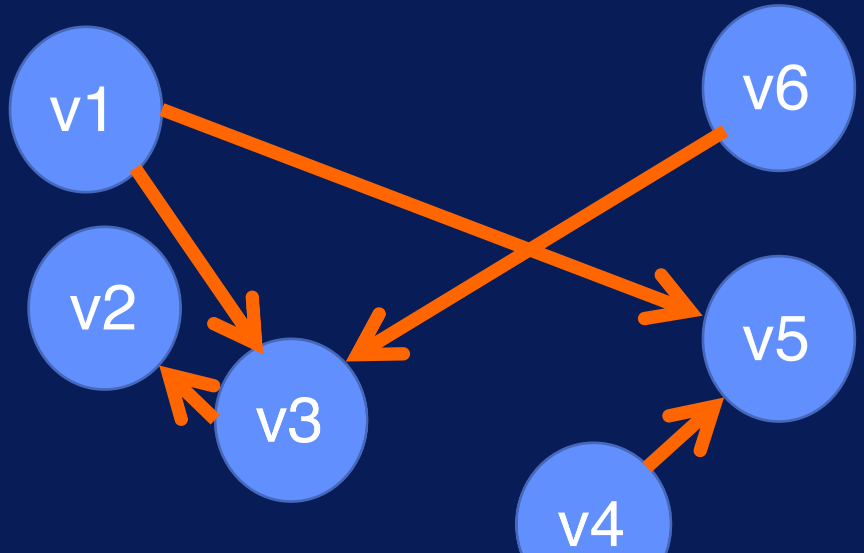
To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3		1				
v4					1	
v5						
v6			1			

Get_neighbor(v3)

Another Operation:
Get_neighbor



To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3		1				
v4					1	
v5						
v6			1			

Search both row and column

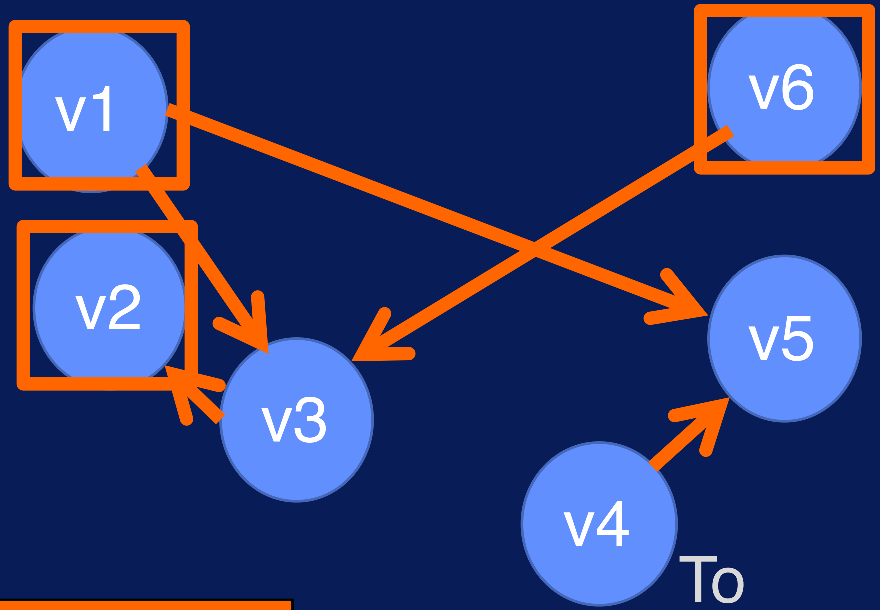
Get_neighbor(v3)

To

From

	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3		1				
v4					1	
v5						
v6			1			

Another Operation:
Get_neighbor



To

From	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3		1				
v4					1	
v5						
v6			1			

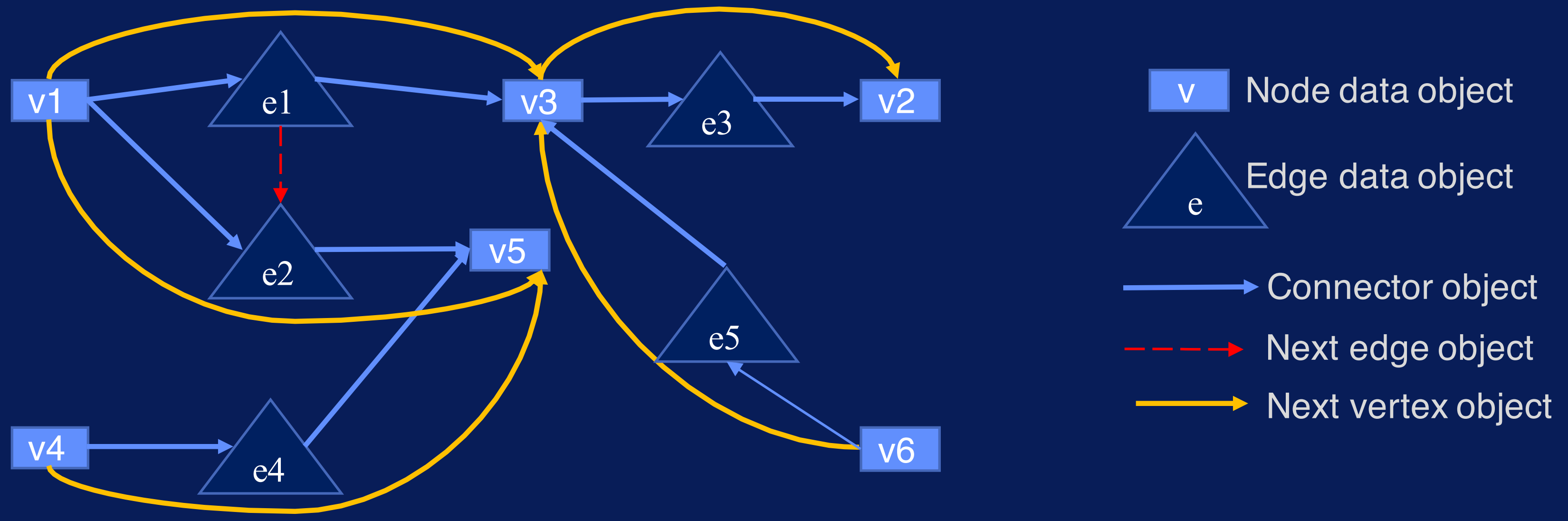
Any entries
are neighbors

Get_neighbor(v3)

To

From	v1	v2	v3	v4	v5	v6
v1			1		1	
v2						
v3		1				
v4					1	
v5						
v6			1			

A Different Representation (just FYI – there are others)



Used internally by some Graph Database Systems

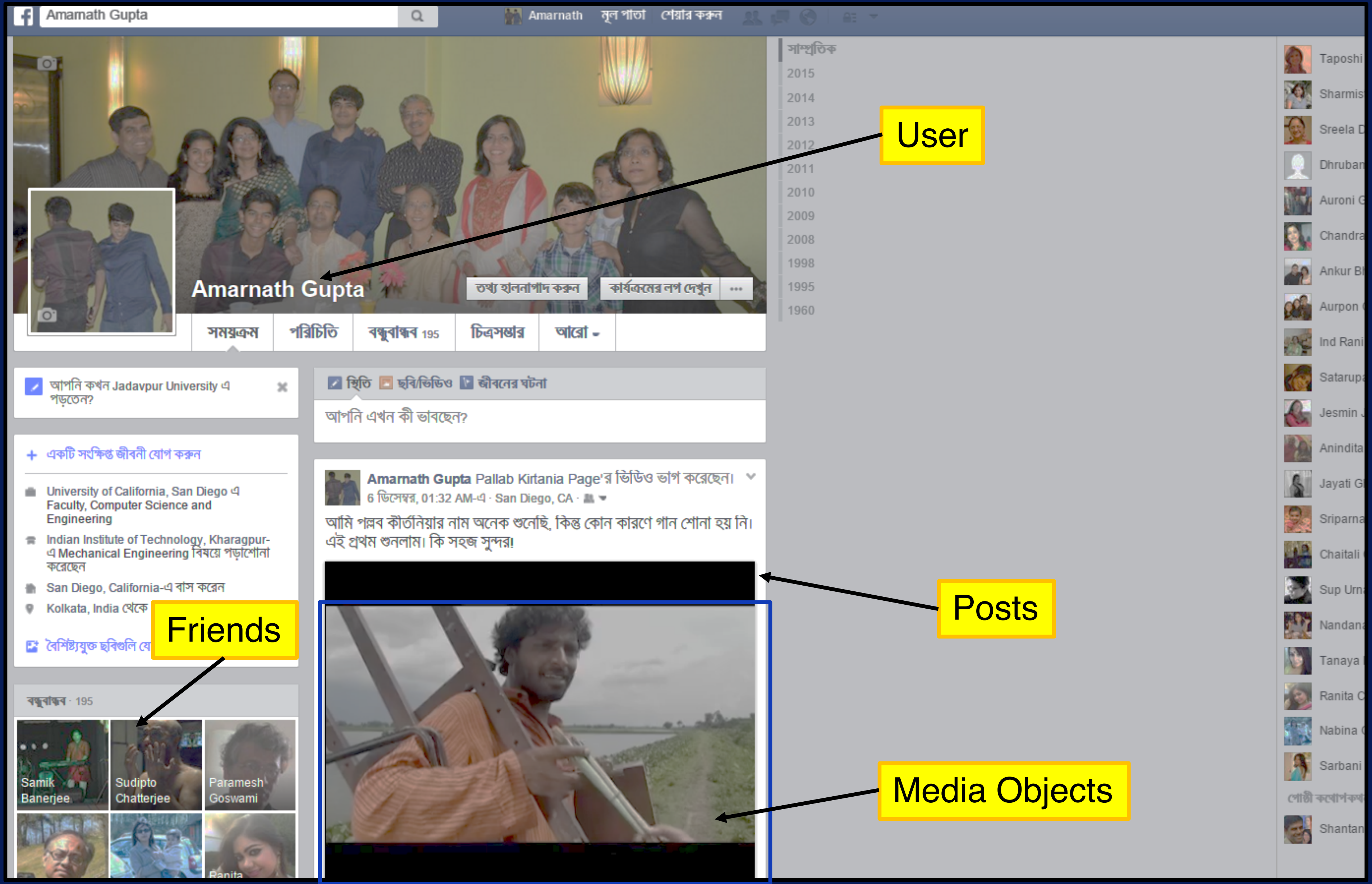
Next: Why we need graphs

Why Graphs?

Case Studies

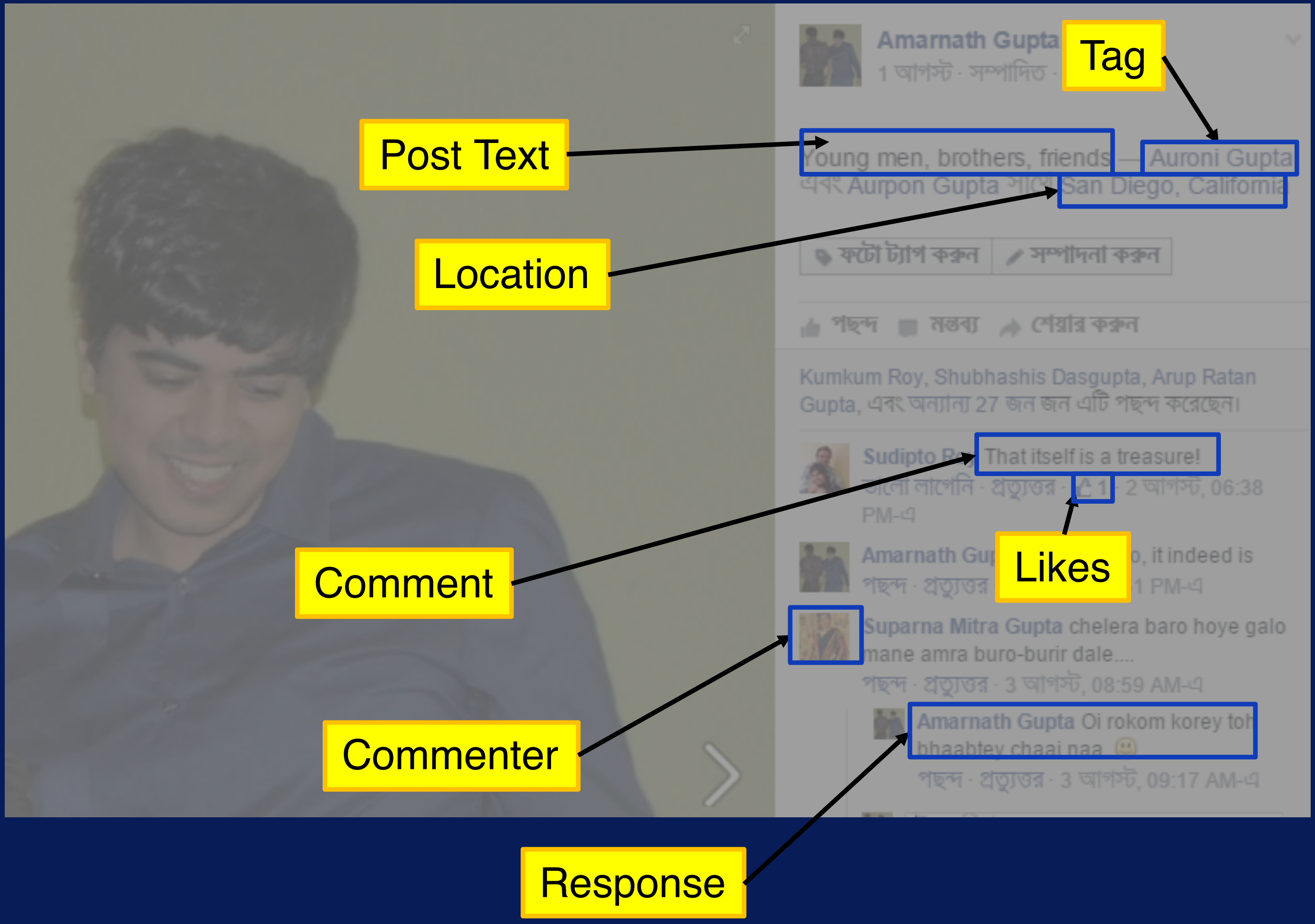
Social Media Produces Graphs

Facebook

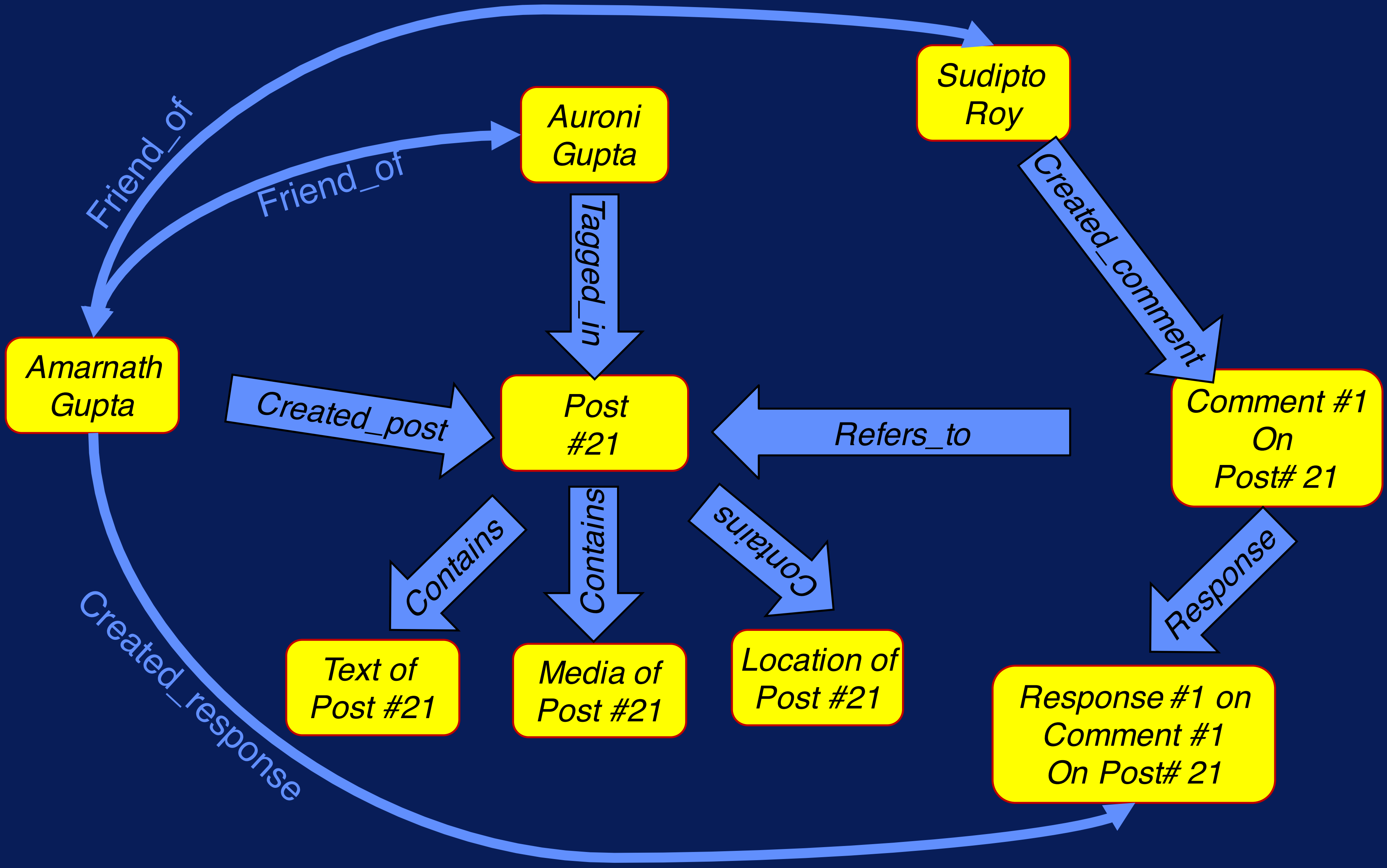


Can you see the graph?

Facebook

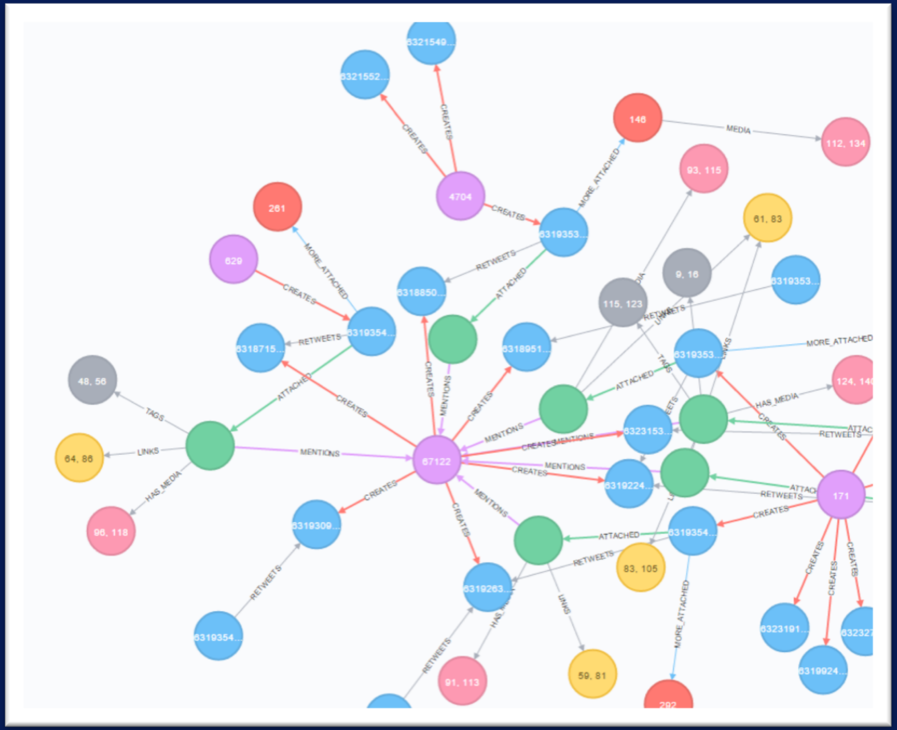


A Fragment of the Graph



4 Use Cases

- **Example 1**
  - Social Media Analytics
- **Example 2**
  - Gene-Phenotype-Disease Networks
- **Example 3**
  - Human Information Network Analytics
- **Example 4**
  - Analysis and Planning for Smart Cities



What is Analytics?

What is Analytics?

- **Discovery and communication of meaningful patterns or interesting insights using**
  - Mathematical properties of data
  - Data computing for accessing and manipulating data
  - Domain knowledge to increase interpretability of data and results of analytics
  - Statistical modeling techniques for drawing inferences or making predictions on data

Some Broad Purposes of "Analytics"

- Uncover characteristics of data set based on its mathematical properties
- Answer specific questions from multiple data sets
- Develop a mathematical model for predicting the behavior of some variables
- Detect emergent phenomena and explain its contributing factors

Graph Analytics

- Analytics where the underlying data is natively structured as or can be modeled as a set of graphs

Graphs and the V's of Big Data

Graphs and the V's of Big Data

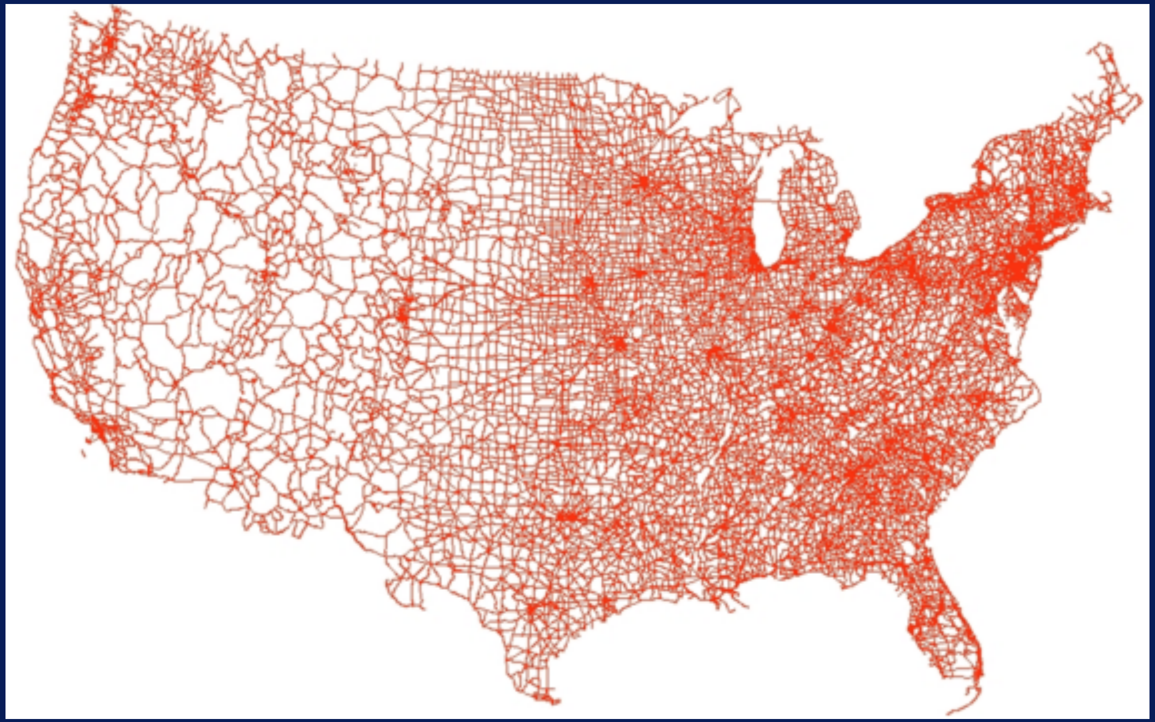
- **Concepts You Will Learn**
  - Algorithmic Complexity
  - Hard decision problems
  - Outcome metrics
  - Streaming edges

Graphs and the V's of Big Data

- Volume
- Velocity
- Variety
- Valence

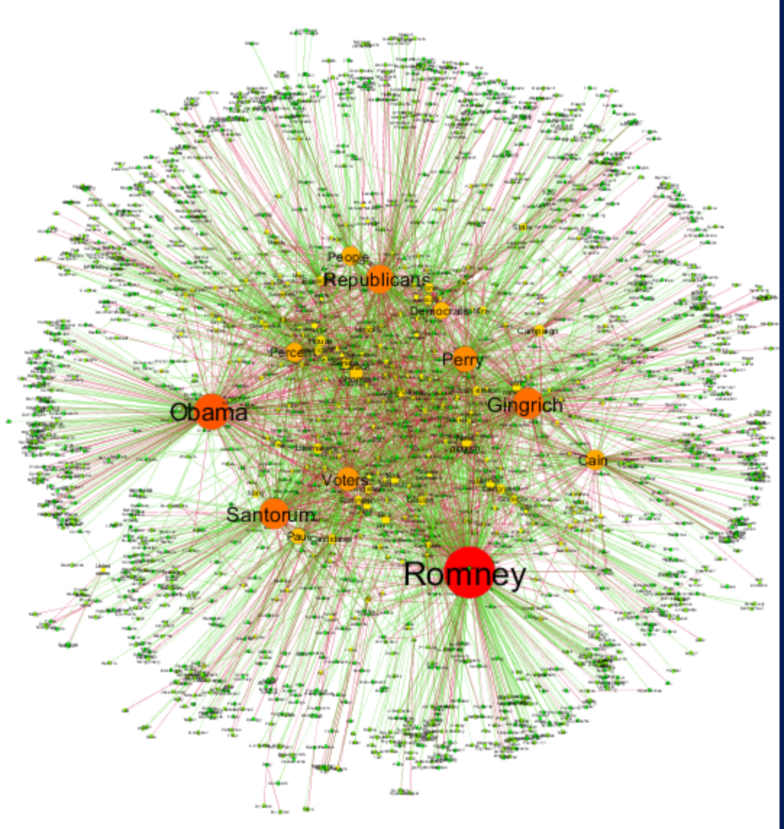
Volume

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



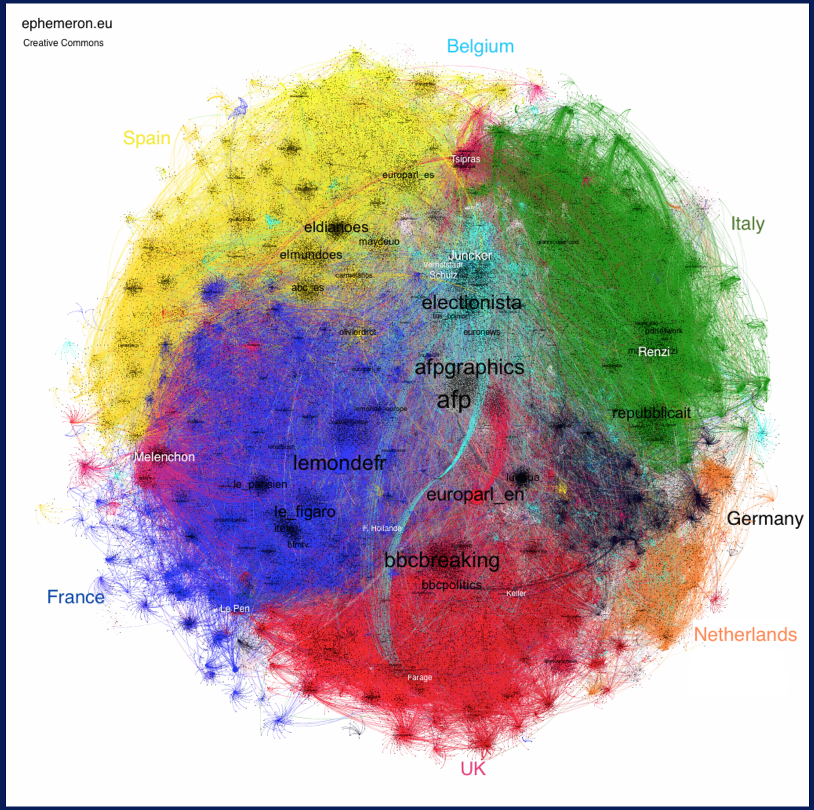
Volume

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



Volume

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

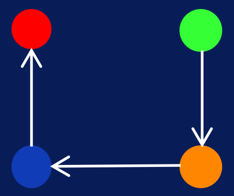


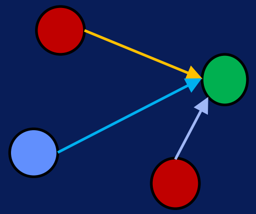
Importance of Volume

- Memory limitations
- Impact on Analytics

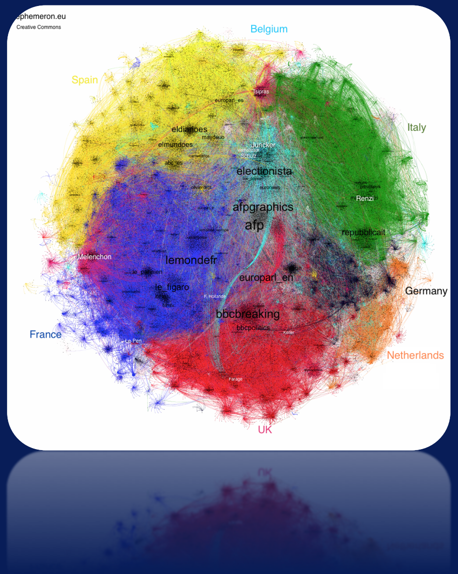
Velocity

Velocity

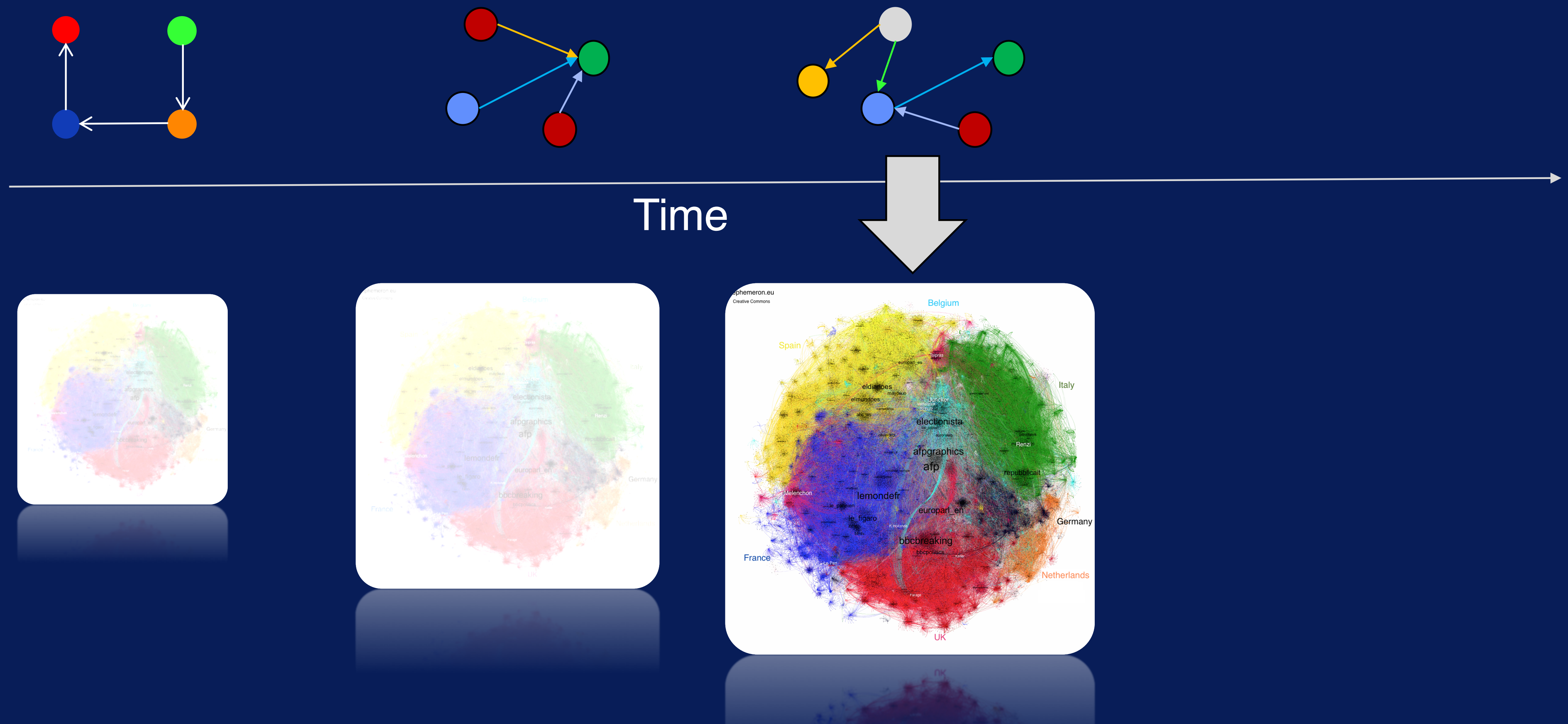




Time



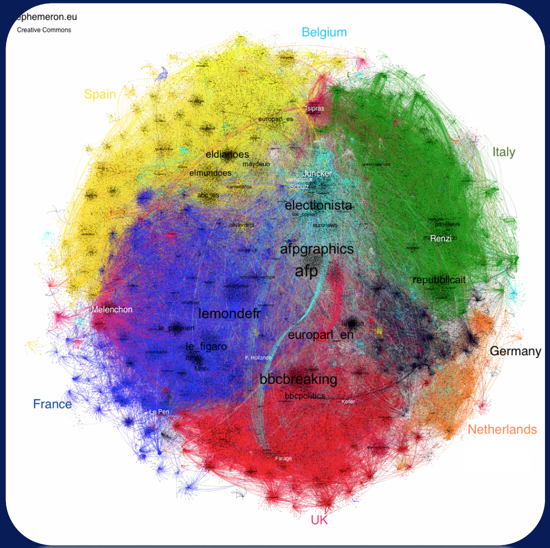
Velocity


Time
ephemeron.eu
Creative Commons
Belgium
Spain
Italy
eldiarioes
elmundoes
electionista
afpgraphics
afp
Renzi
Melenchon
lemondefr
le_figaro
europarl_en
Germany
bbcbreaking
France
Netherlands
UK

Velocity

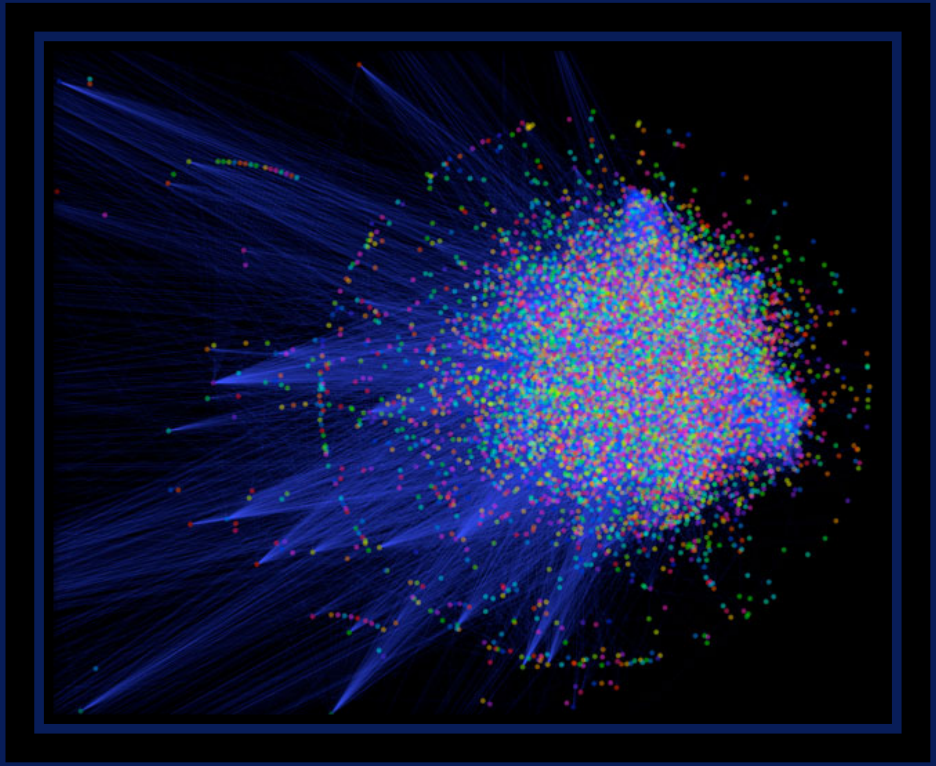
- The rate at which a graph increases in size and complexity

Streaming edges into graphs



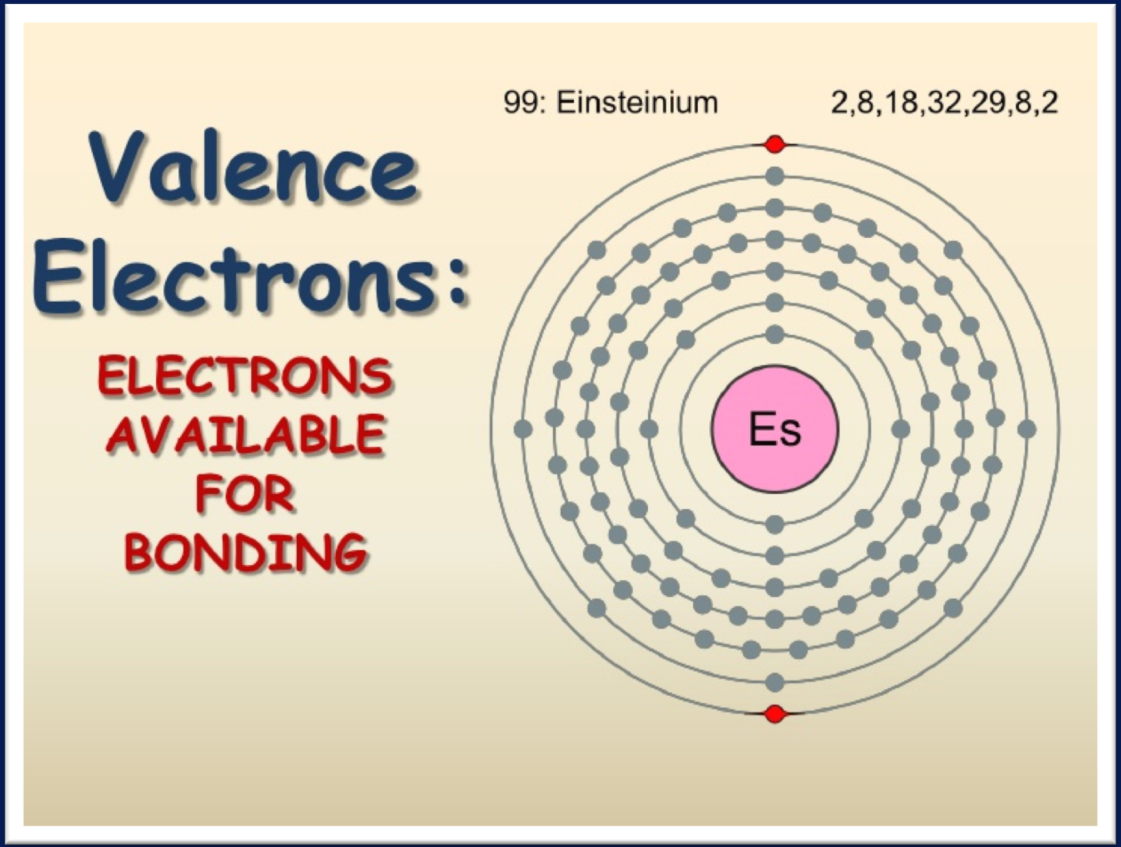
Variety

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



Valence

- In Chemistry



Valence

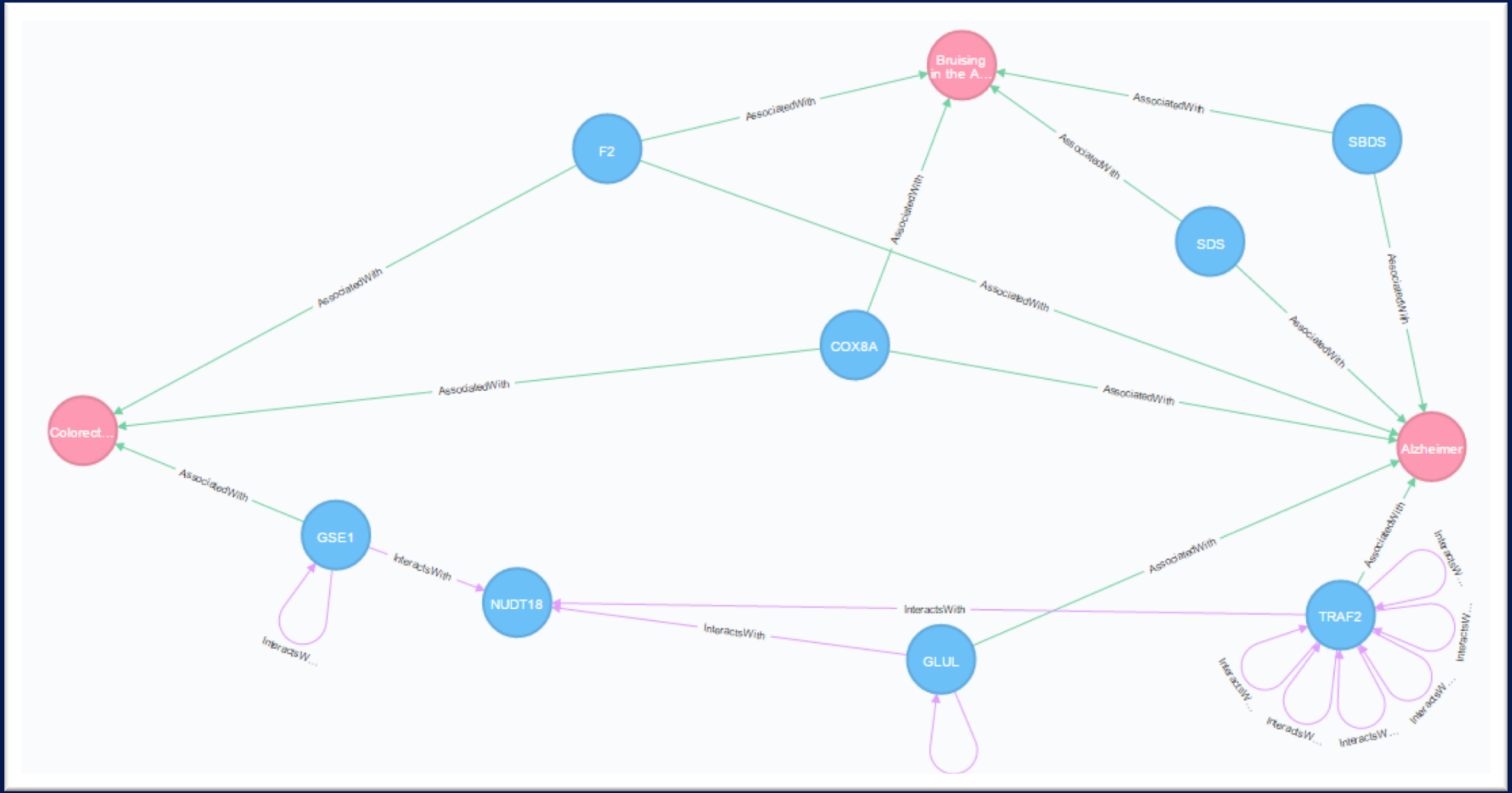
- **In Graphs**
    - A measure of **connectedness**

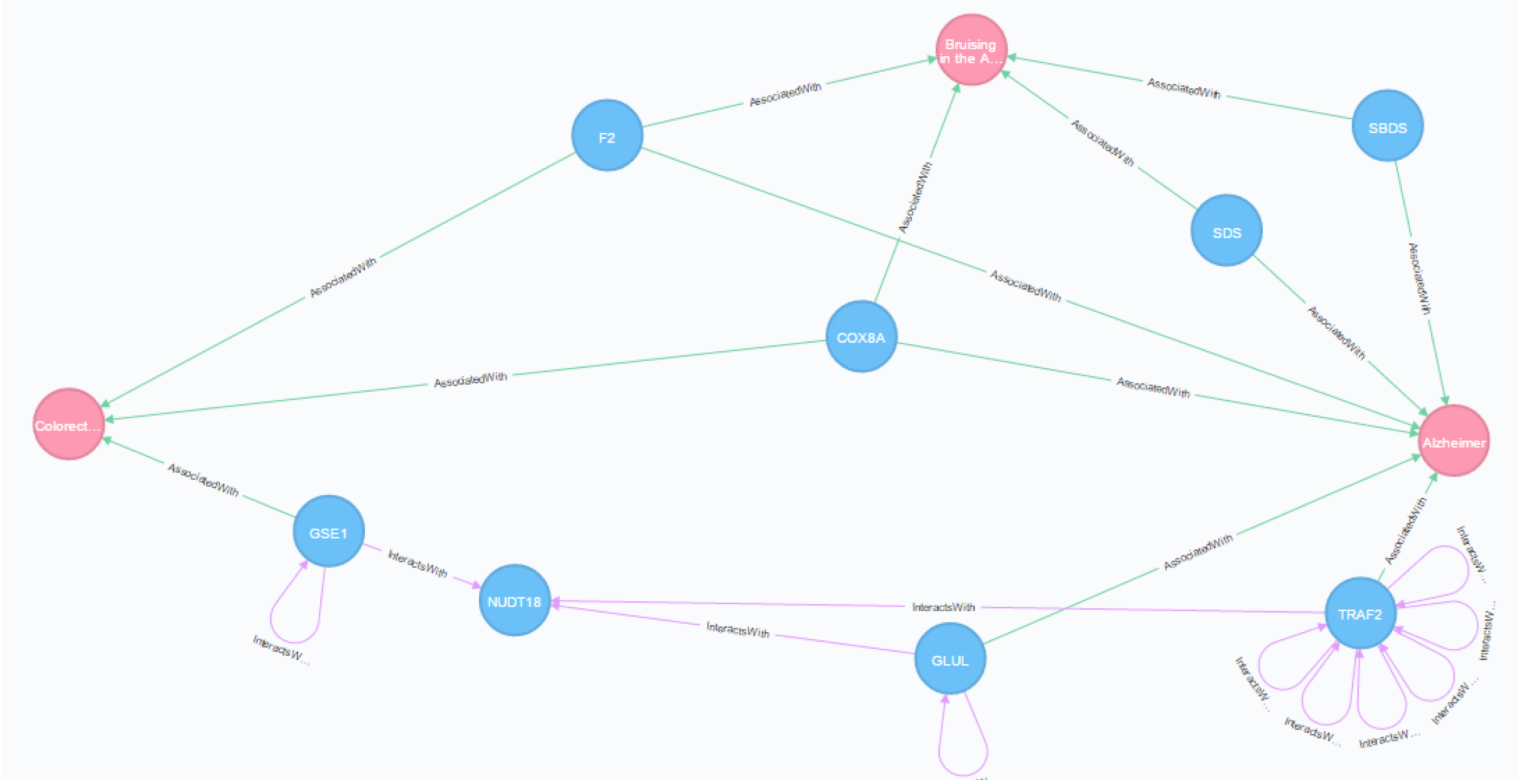
How Volume Impacts Graph Analytics

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

How Volume Impacts Graph Analytics

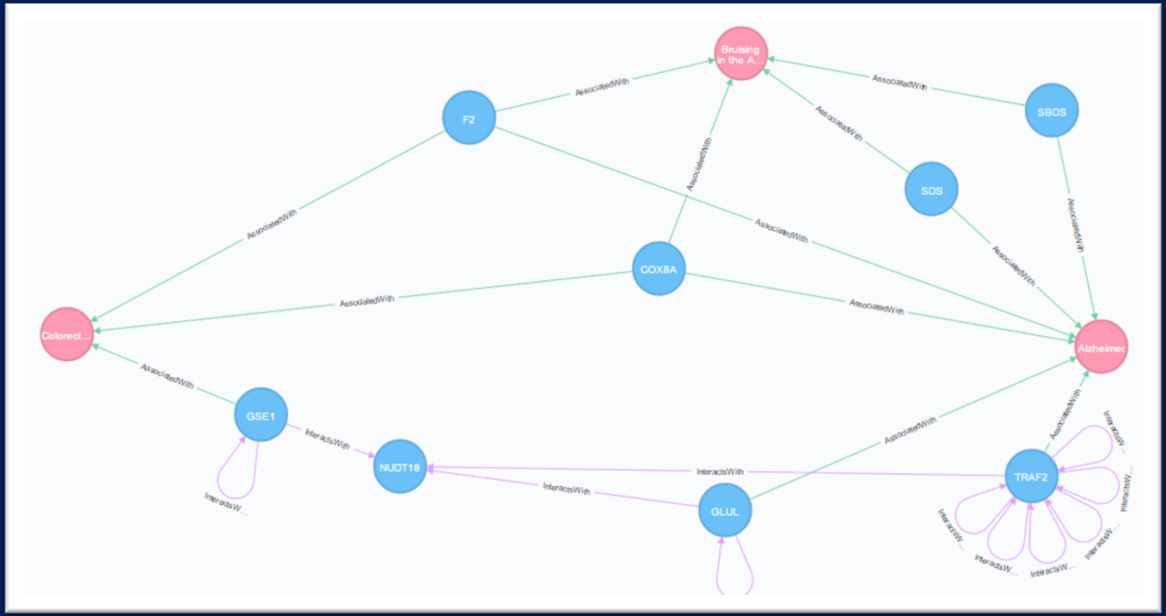
- **An example**
  - Problem: Find a Simple Path between “Alzheimer’s Disease” and “Colorectal Cancer




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

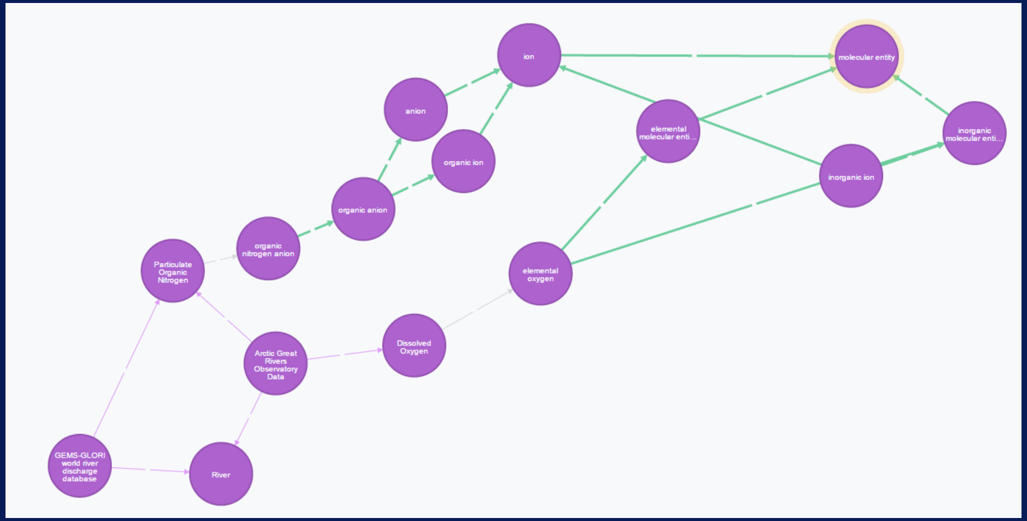
How Volume Impacts Graph Analytics

- 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*



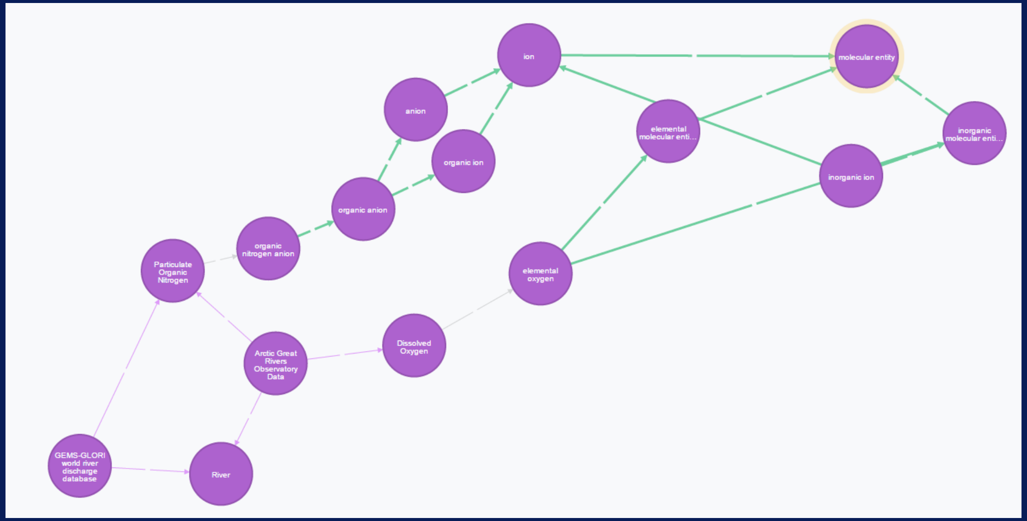
How Velocity Impacts Graph Analytics

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



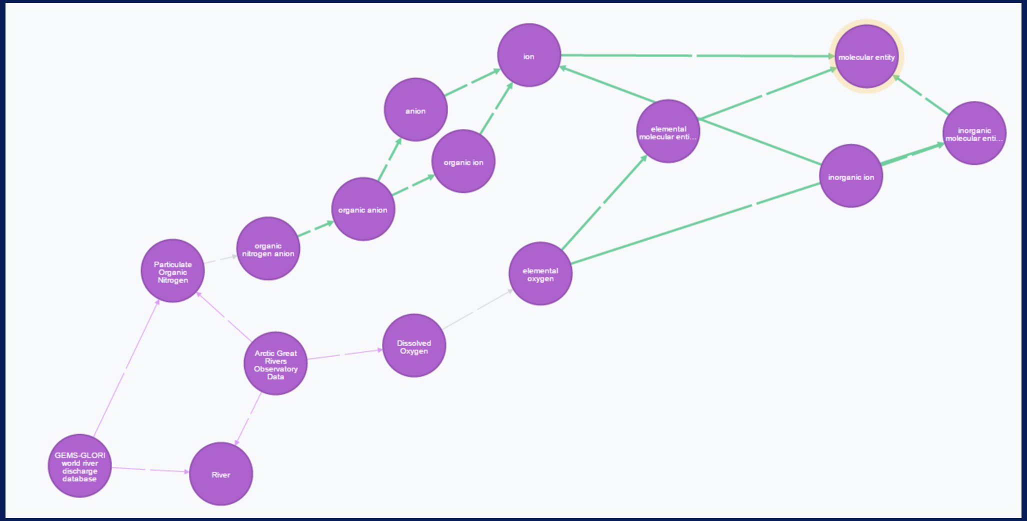
How Velocity Impacts Graph Analytics

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



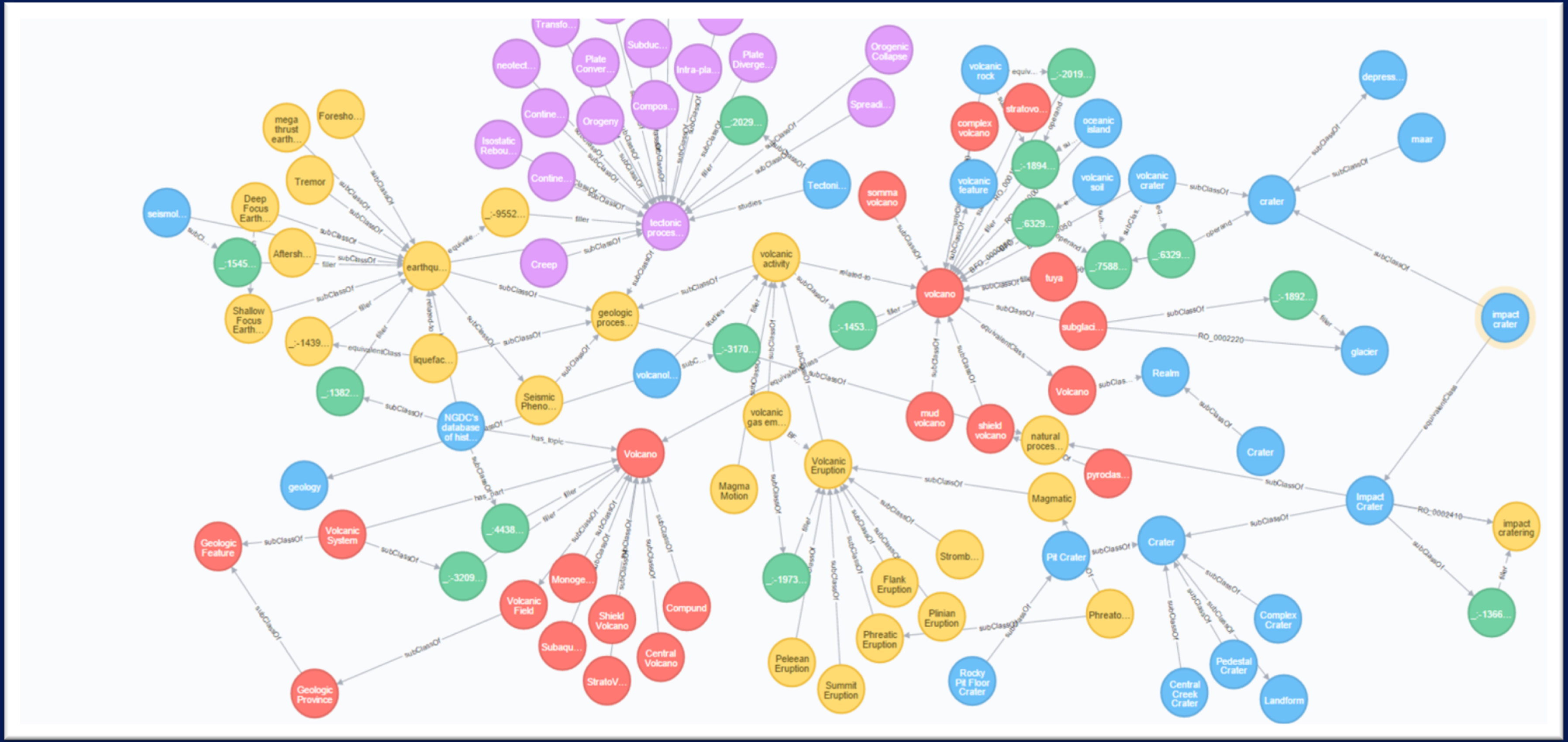
Velocity

- **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

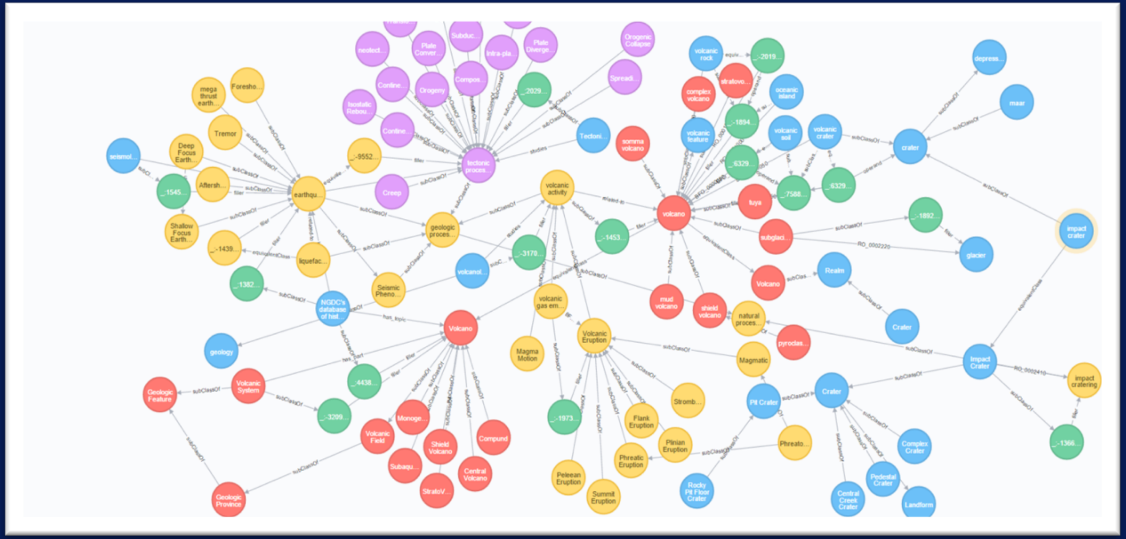
- Different kinds of graphs may have very different meanings

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

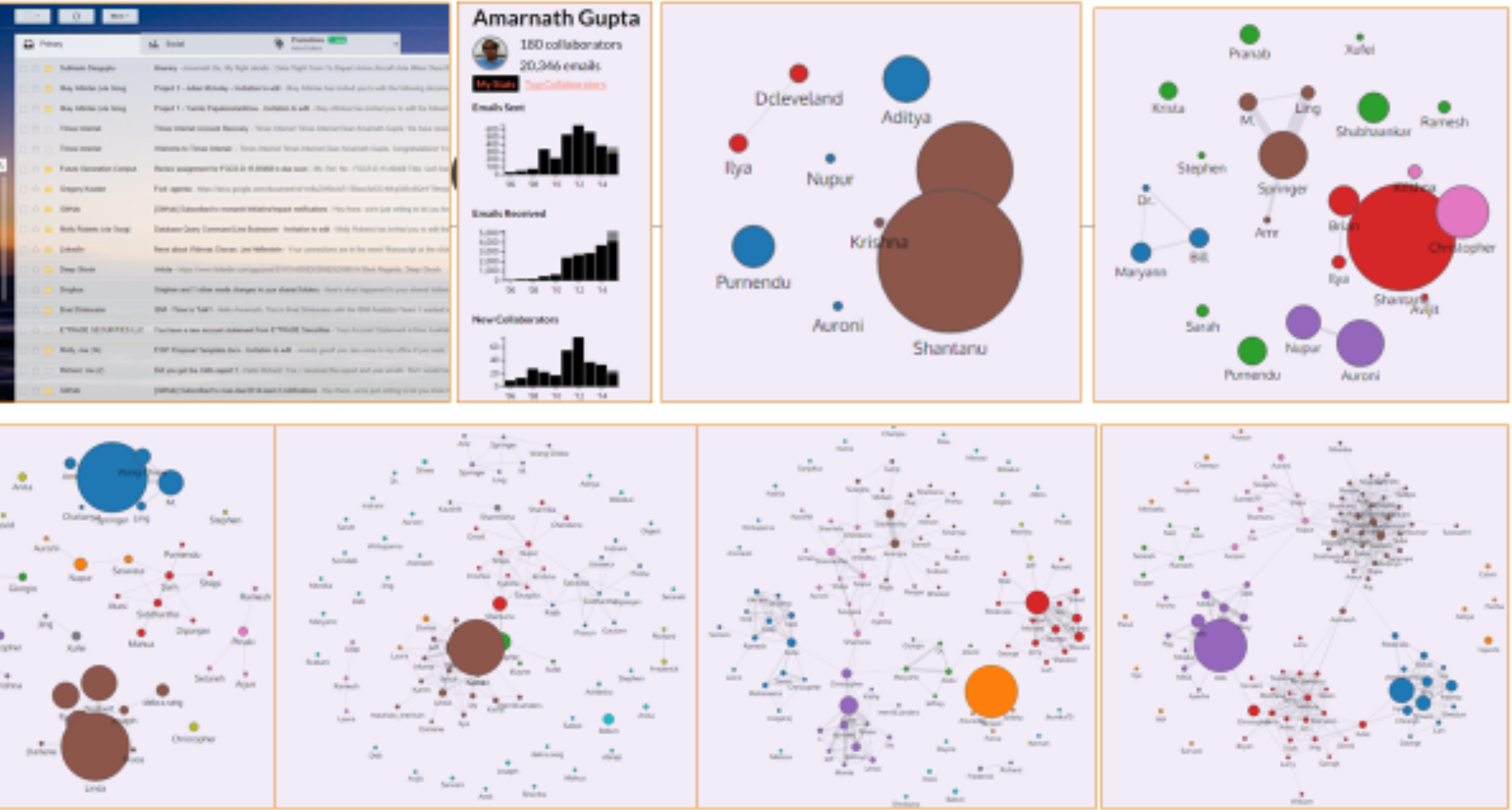
Ontologies

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

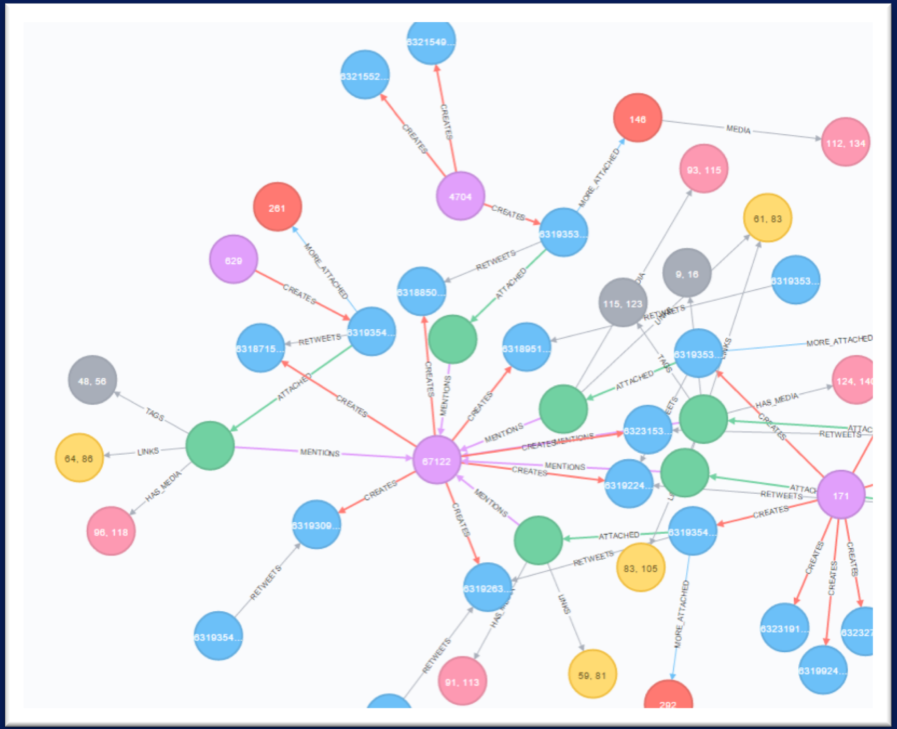

Amarnath Gupta
180 collaborators
20,346 emails
My Stats
Emails Sent
Emails Received
New Collaborators
Dcleveland
Aditya
Ilya
Nupur
Krishna
Purnendu
Auroni
Shantanu
Pranab
Xufei
Krista
M.
Ling
Shubhaankar
Ramesh
Stephen
Springer
Dr.
Amr
Brian
Christopher
Bill
Maryann
Ilya
Shantanu
Avijit
Sarah
Nupur
Purnendu
Auroni

Why Graphs?

Example 1: Social Media

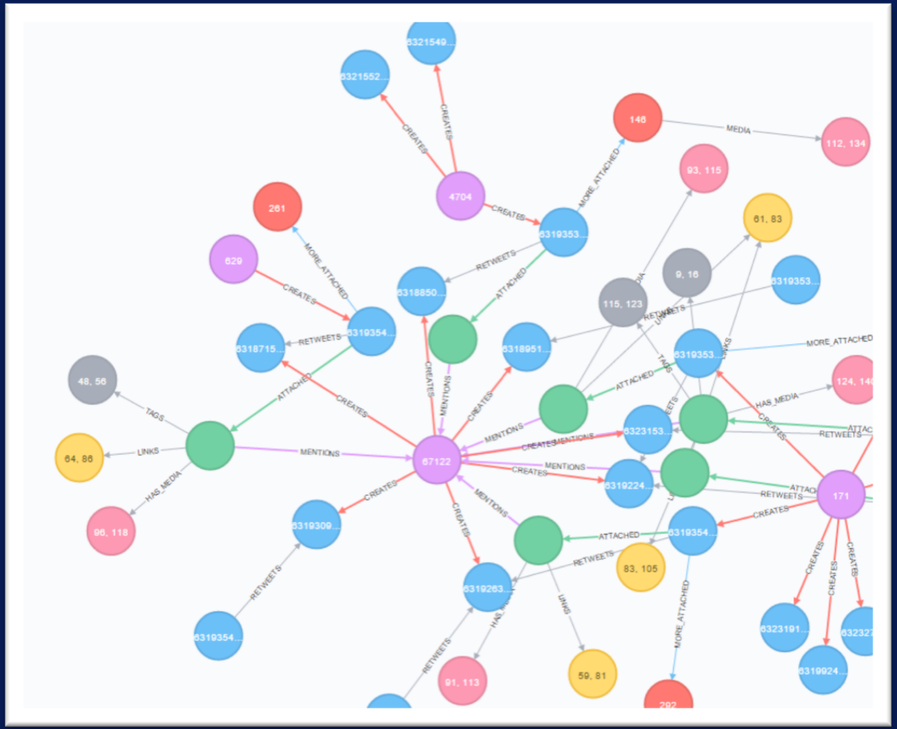
Tweets are Graphs

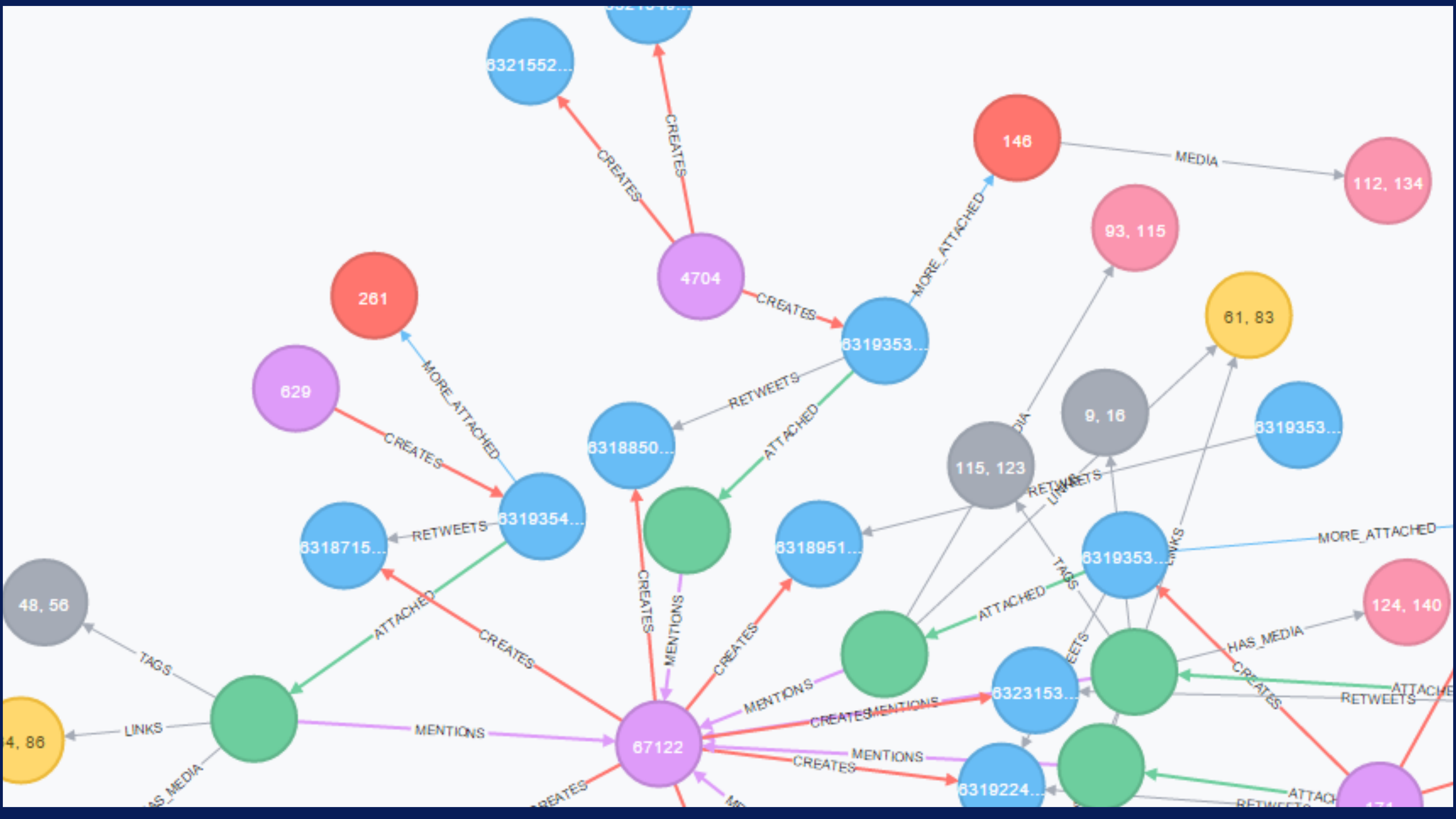
- **Many kinds of nodes**
  - Users
  - Tweets
  - URLs
  - Hashtags
  - Media



Tweets capture Human Interactions

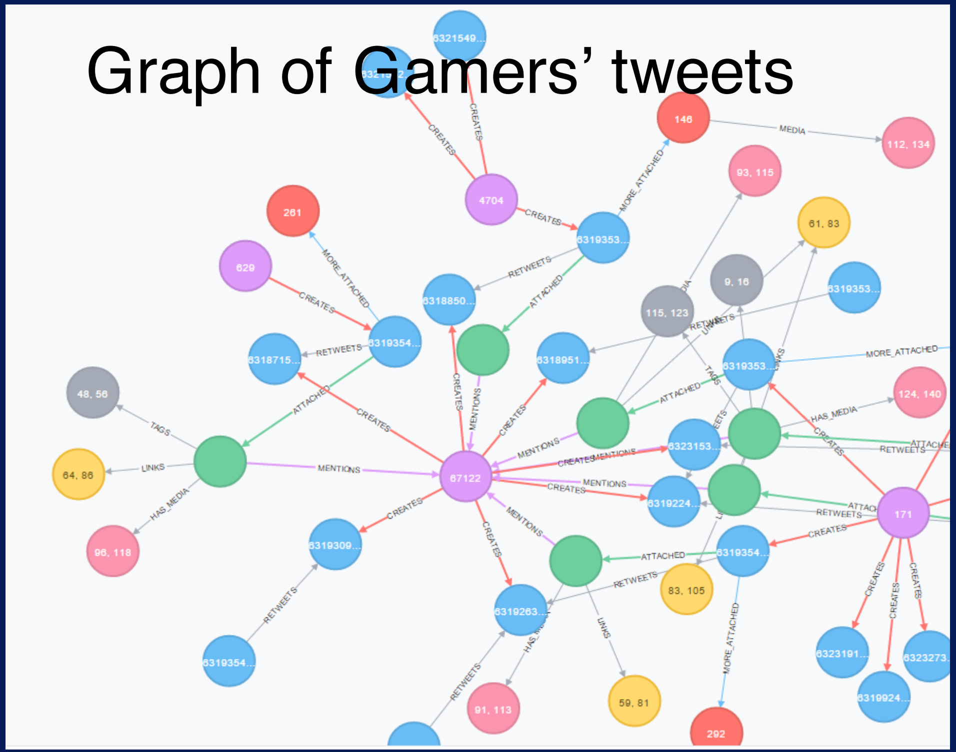
- **Many kinds of edges**
  - User creates Tweets
  - A tweet is in response to another
  - A tweet retweets another
  - User mentions User
  - Tweet contains Hashtag
  - User follows User




6321552...
CREATES
CREATES
146
MEDIA
112, 134
93, 115
4704
CREATES
MORE_ATTACHED
261
61, 83
6319353...
629
MORE_ATTACHED
CREATES
RETWEETS
ATTACHED
6318850...
9, 16
6319353...
115, 123
6319354...
RETWEETS
6318715...
6318951...
6319353...
MORE_ATTACHED
48, 56
ATTACHED
CREATES
CREATES
MENTIONS
CREATES
TAGS
ATTACHED
124, 140
TAGS
HAS_MEDIA
MENTIONS
LINKS
MENTIONS
6323153...
RETWEETS
67122
MENTIONS
CREATES
6319224...

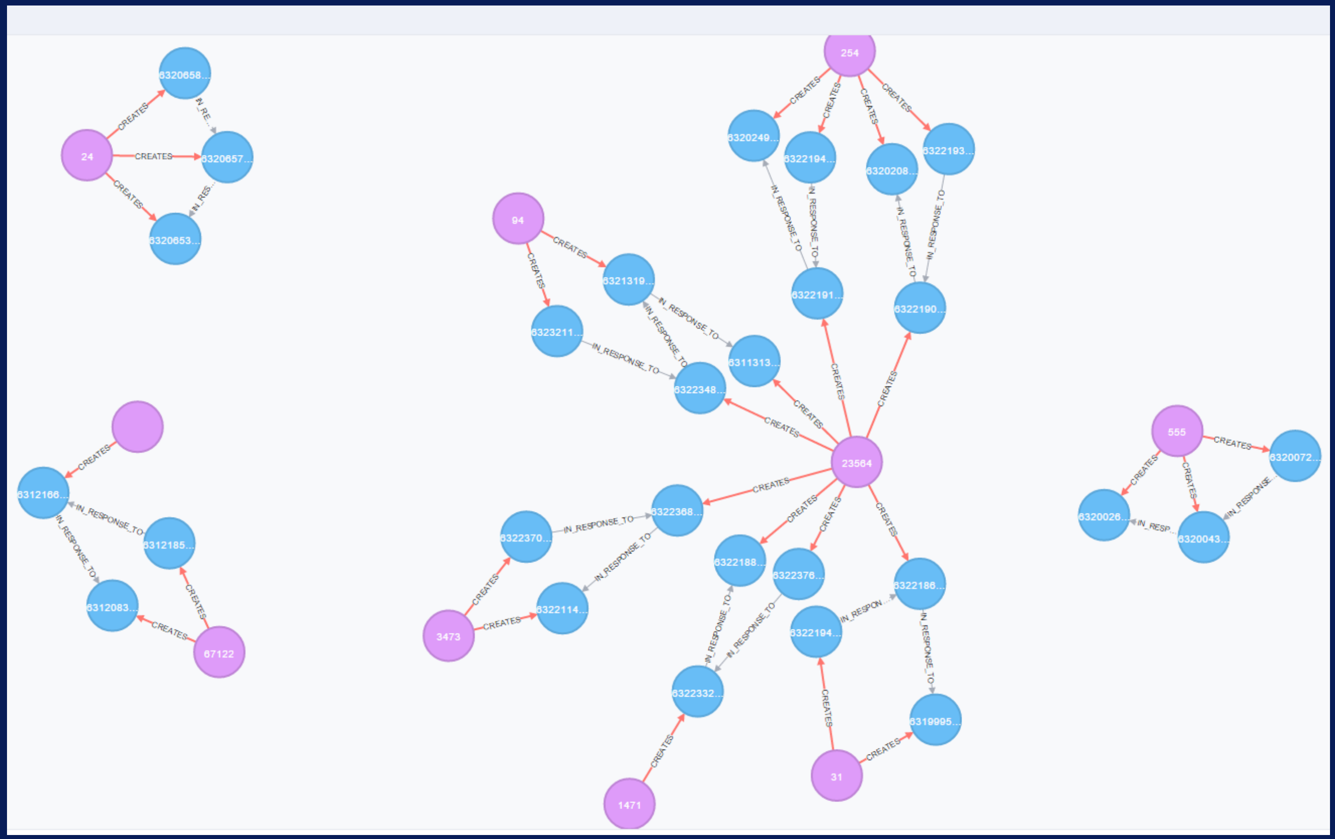
Example 1: Social Media Analytics

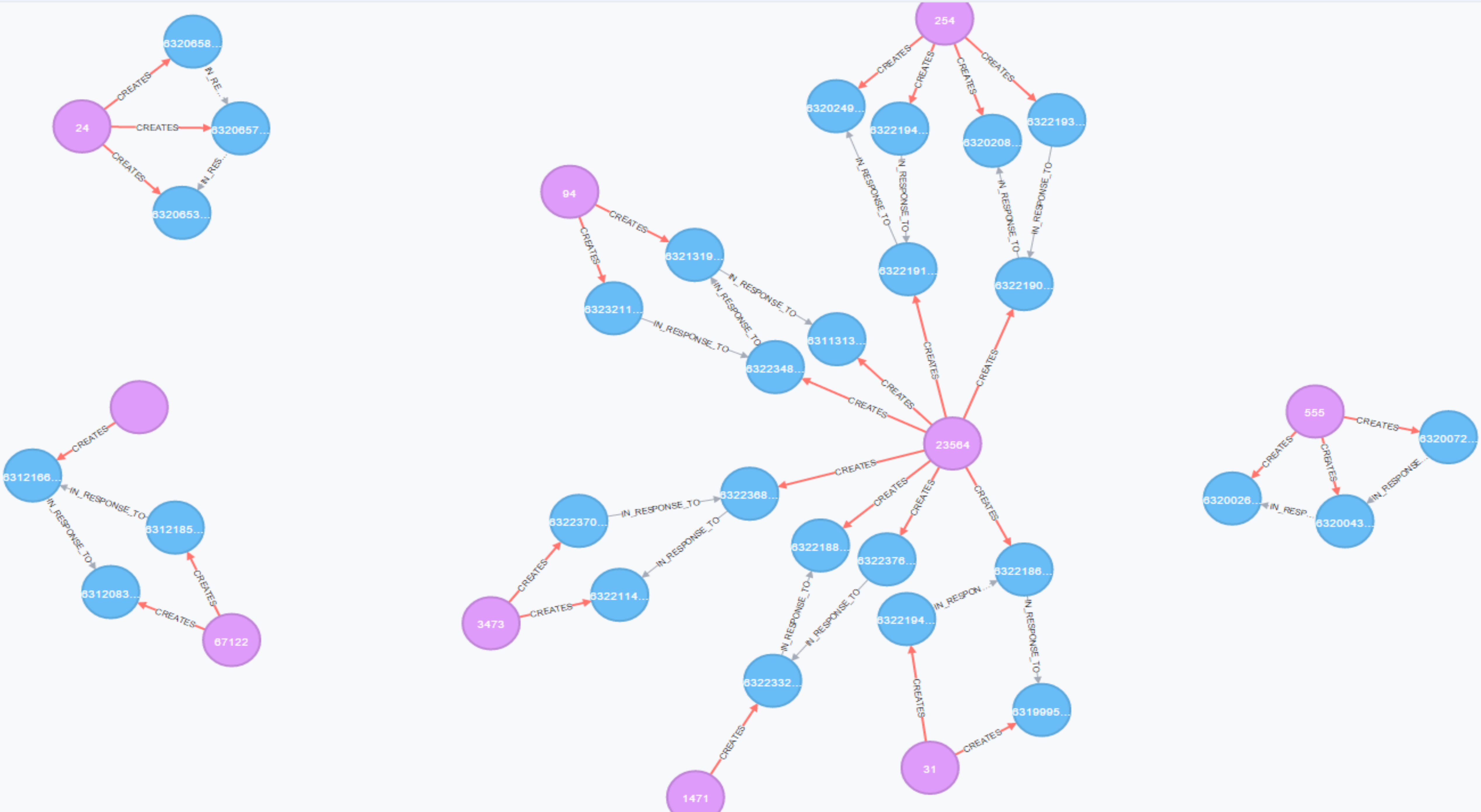
- **Behavioral Psychology**
  - A branch of psychology that studies people's behavior
- **Data Science questions of a behavioral psychologist**
  - Do players of violent online games show more confrontational behavior in their tweet conversations?
  - Can we tell how "addicted" a user is to a game?



Why Graphs?

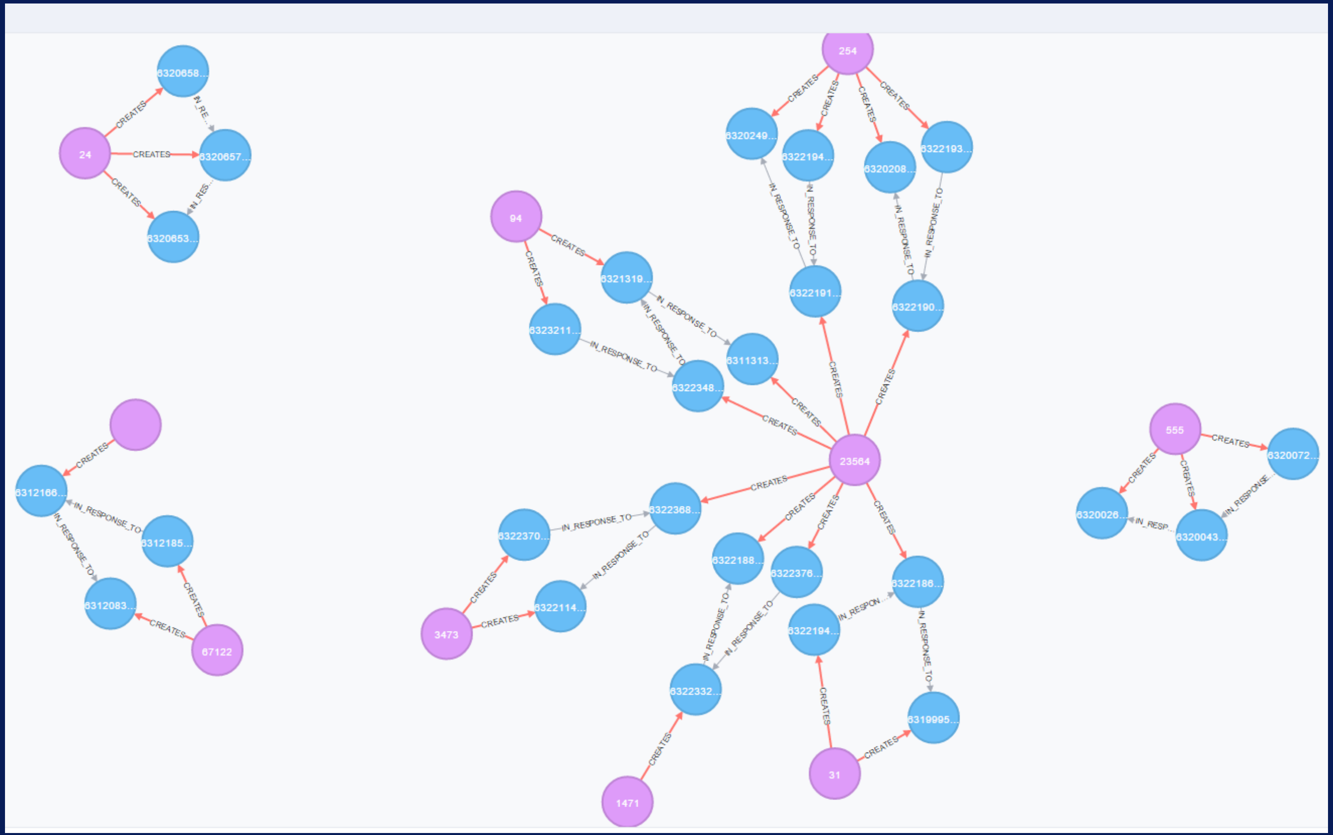
- **Extracting Conversation Threads**




24
6320658...
6320657...
6320653...
254
6320249...
6322194...
6320208...
6322193...
6322191...
6322190...
94
6321319...
6323211...
6311313...
6322348...
23564
6312166...
6312185...
6312083...
67122
555
6320072...
6320026...
6320043...
3473
6322370...
6322368...
6322114...
6322188...
6322376...
6322186...
6322194...
6322332...
6319995...
1471
31
CREATES
IN_RESPONSE_TO
IN_RE...
IN_RES...
IN_RESP...
IN_RESPON...
IN_RESPONSE...

Why Graphs?

- Extracting Conversation Threads
- Finding Interacting Groups
- Finding Influencers in a Community

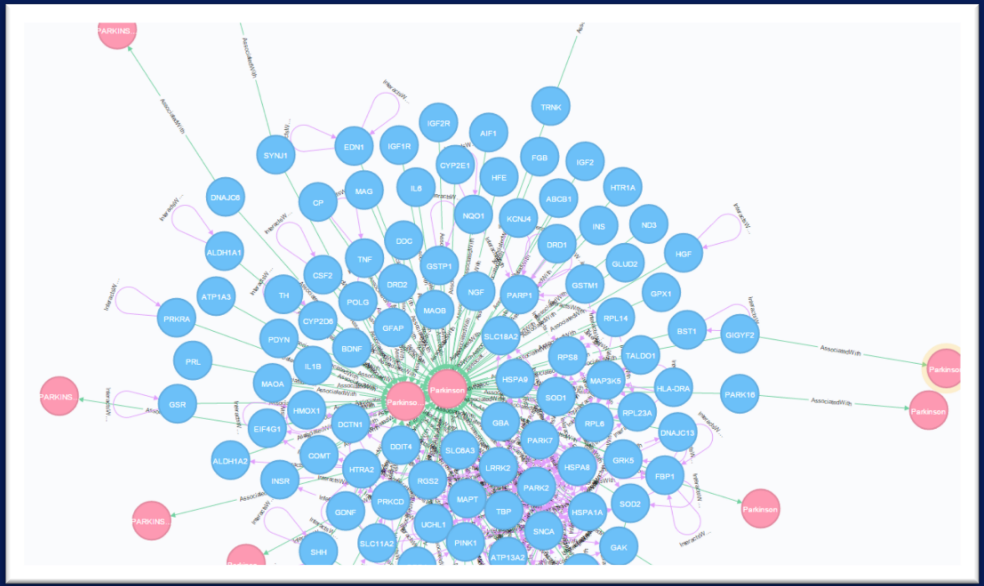


pause

Why Graphs?

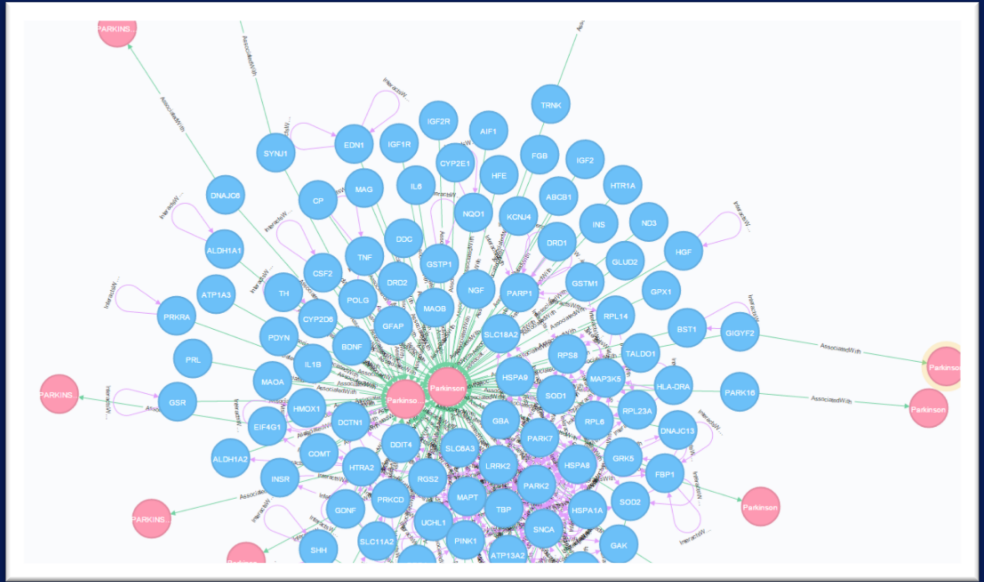
Example 2: Biology

Biological Entities Interact



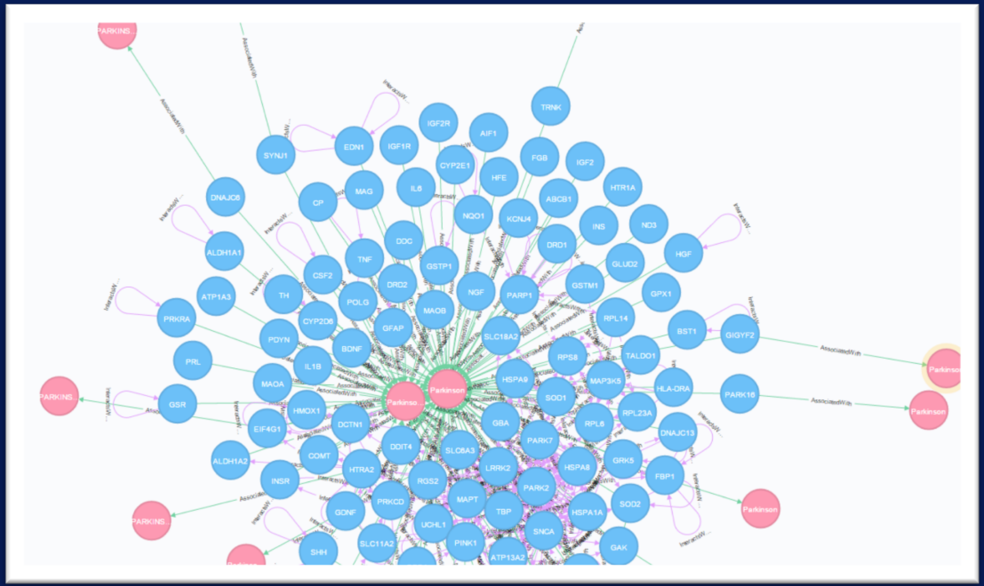
- **Networks arising from**
  - Experimental observations
    - Gene-protein relationships
    - Gene-gene interactions
    - Cell-cell signaling
    - Gene-phenotype-disease relationships
  - Human Knowledge

Biological Entities Interact



- **Networks arising from**
  - Experimental observations
    - Gene-protein relationships
    - Gene-gene interactions
    - Cell-cell signaling
    - Gene-phenotype-disease relationships
  - Human Knowledge
    - Anatomical knowledge

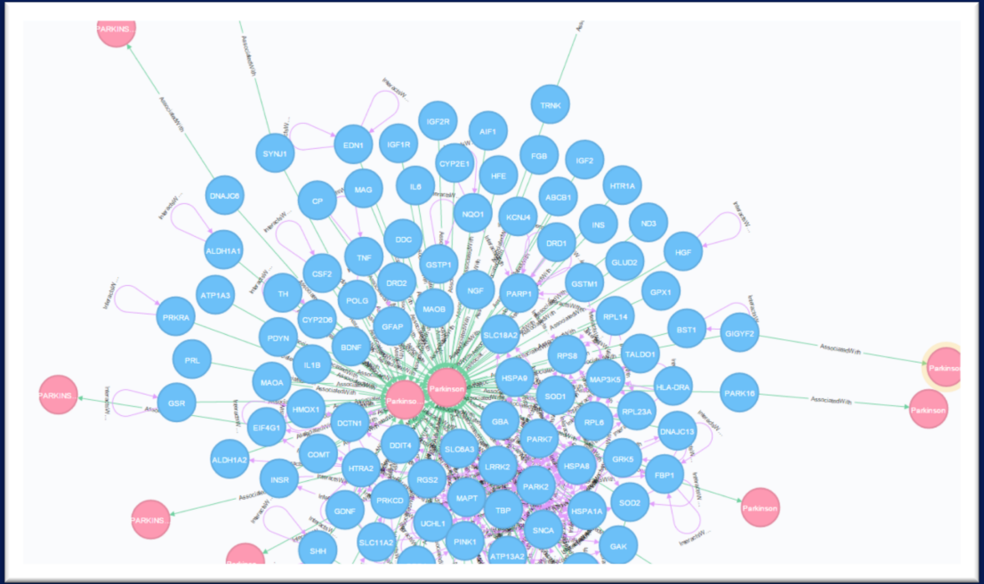
Biological Entities Interact



- **Networks arising from**
  - Experimental observations
    - Gene-protein relationships
    - Gene-gene interactions
    - Cell-cell signaling
    - Gene-phenotype-disease relationships
  - Human Knowledge
    - Anatomical knowledge
    - Taxonomy of the animal kingdom …

Computed Relations

- **Networks arising from**
  - Computational Techniques
    - Bioinformatics algorithms
    - Literature mining

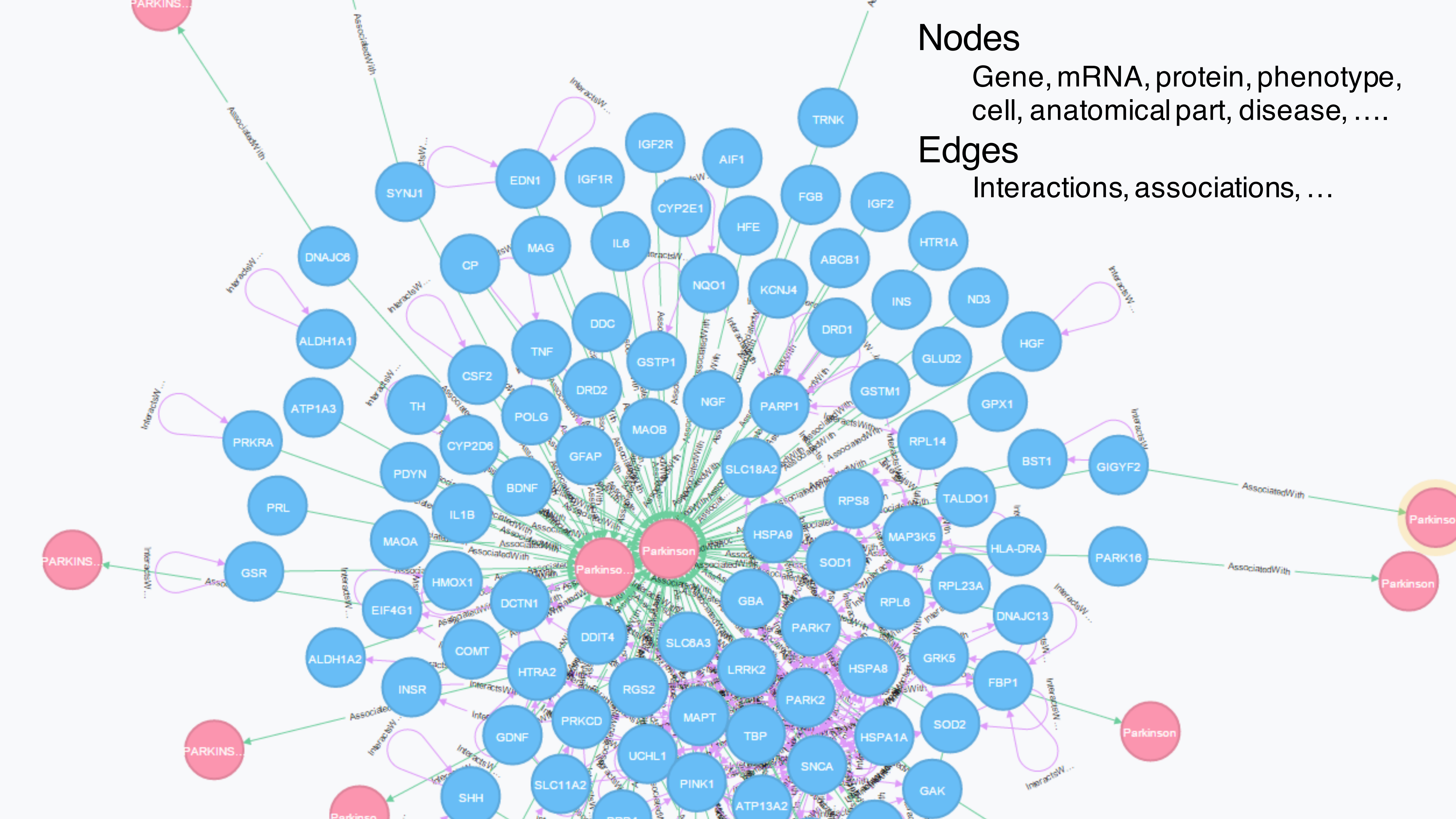


Nodes

Gene, mRNA, protein, phenotype, cell, anatomical part, disease, ….

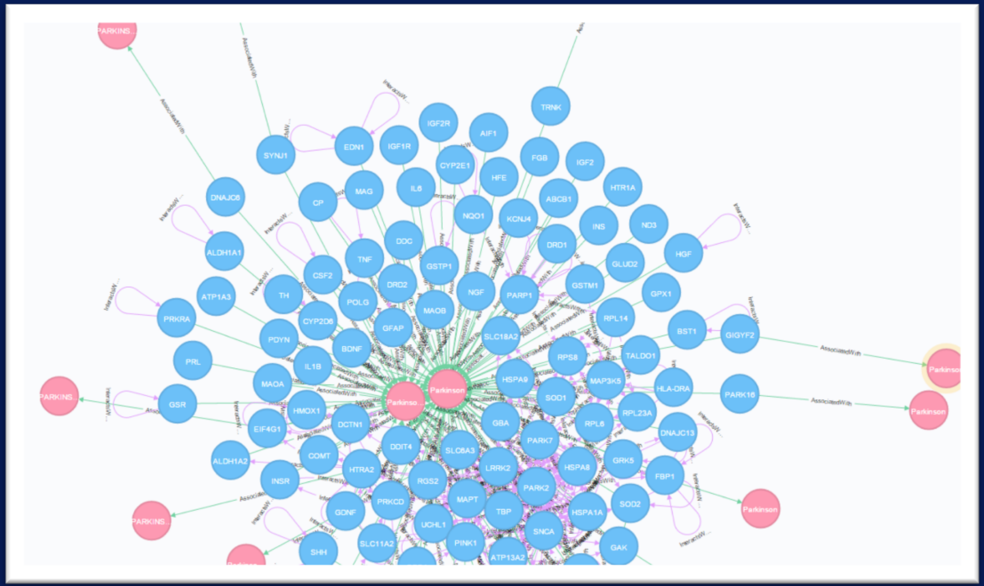
Edges

Interactions, associations, …



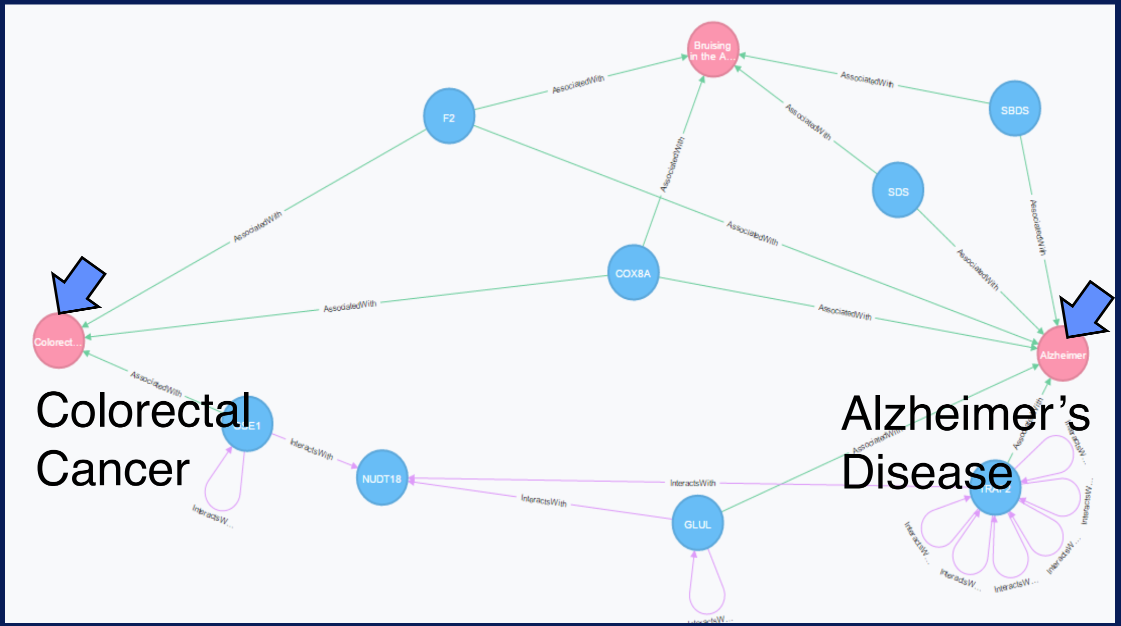
Data Integration

- Data sets must be assembled
- More data integration = larger data volume



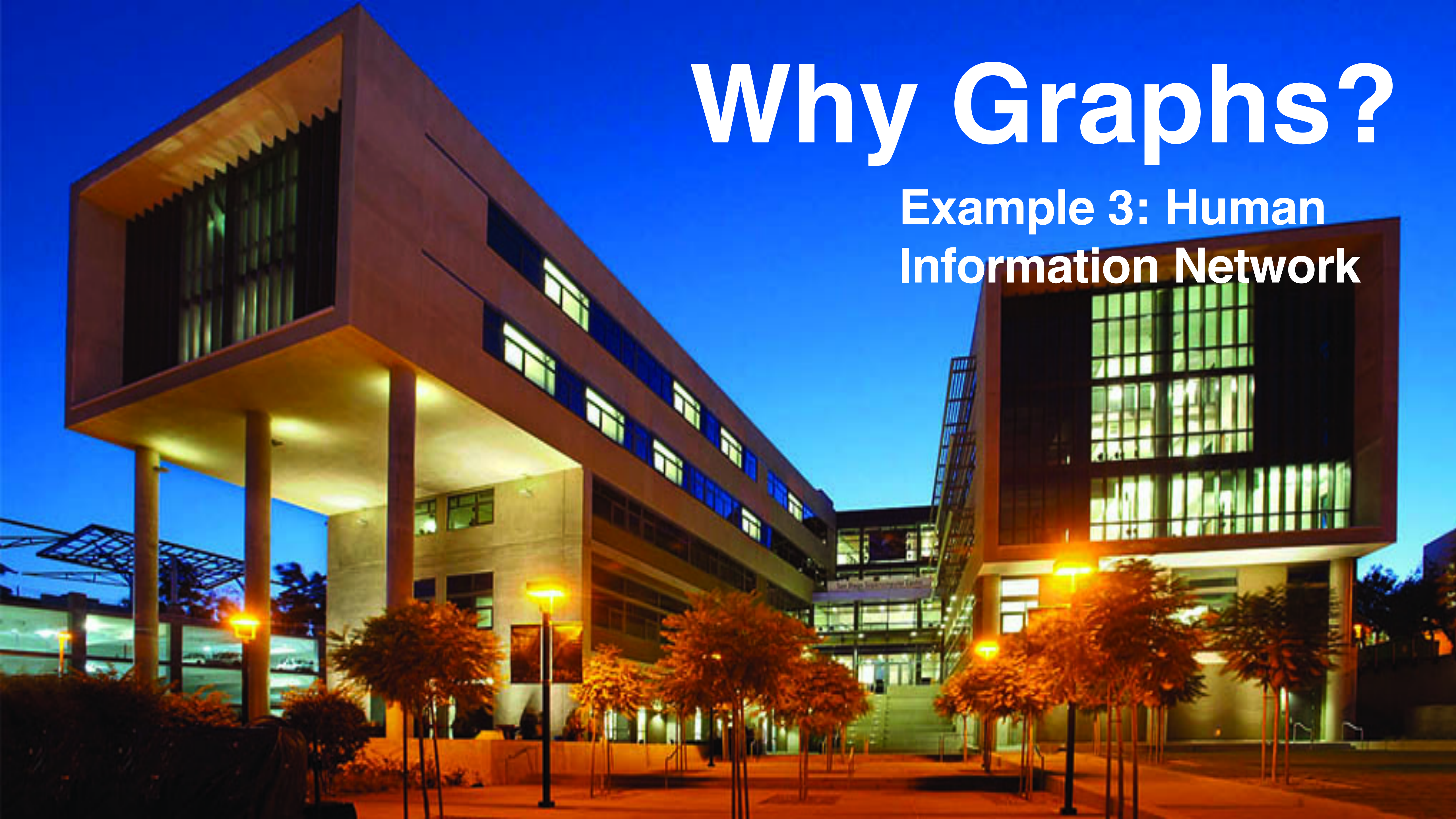
Why Graphs?

- Discovering Unknown Relationships
  - Connecting the dots
    - Indirect associations between diseases
    - Exploratory analysis



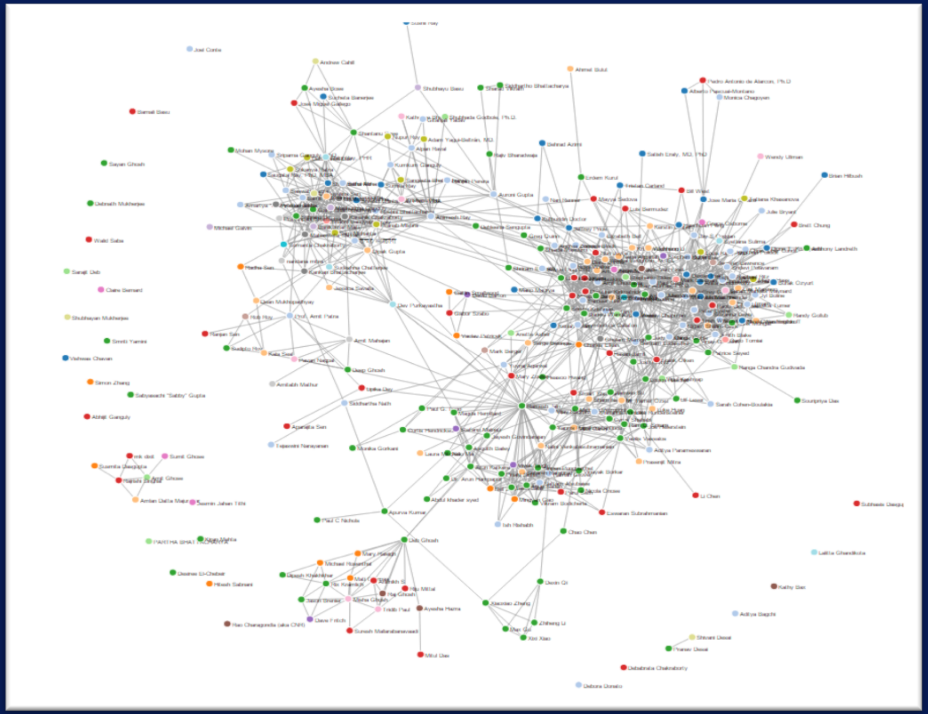
Why Graphs?

Example 3: Human Information Network



Human Information Network Analytics

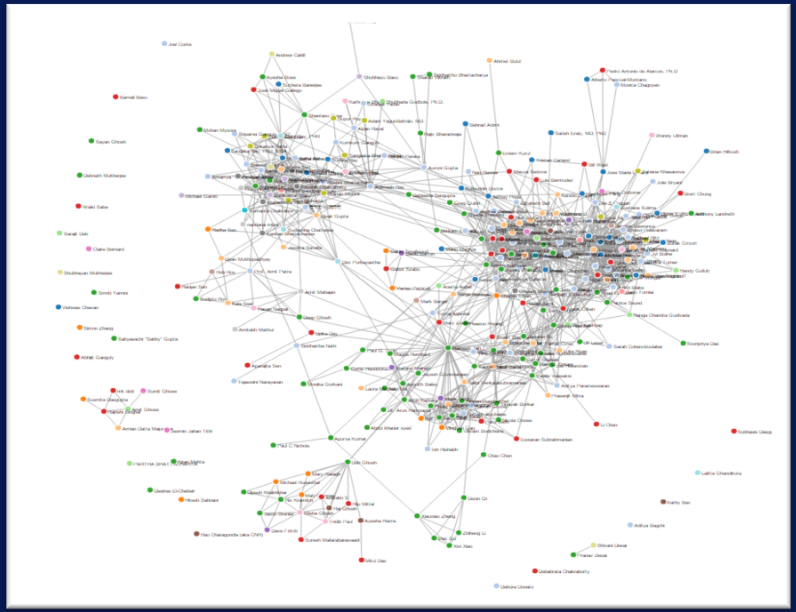
- **Combining**
  - Professional information (e.g., LinkedIn)
  - Personal information made public (e.g., FB, Google Plus)
  - Calendar information
  - Contact/Relationship information
  - Public co-activities with other people
  - Organized by time and events
  - Annotated with events



Why Graphs?

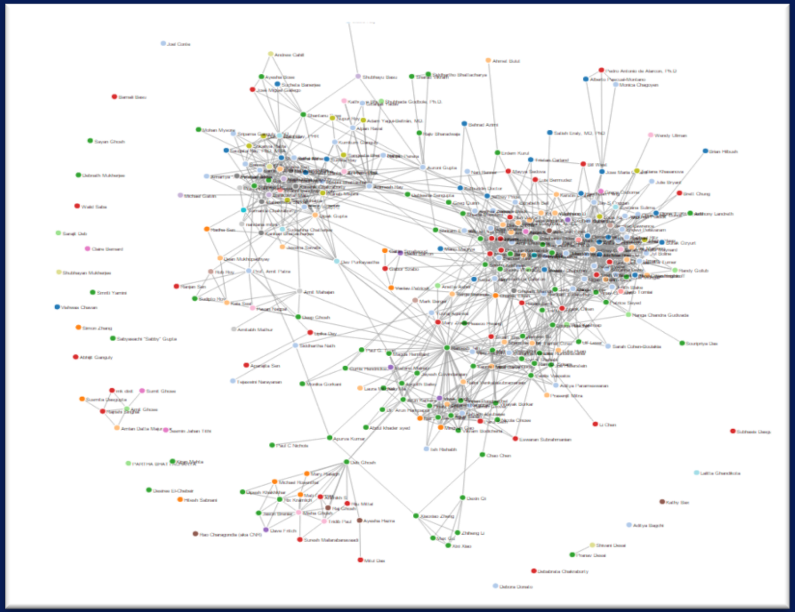
Example 4: Human Information Network Analytics

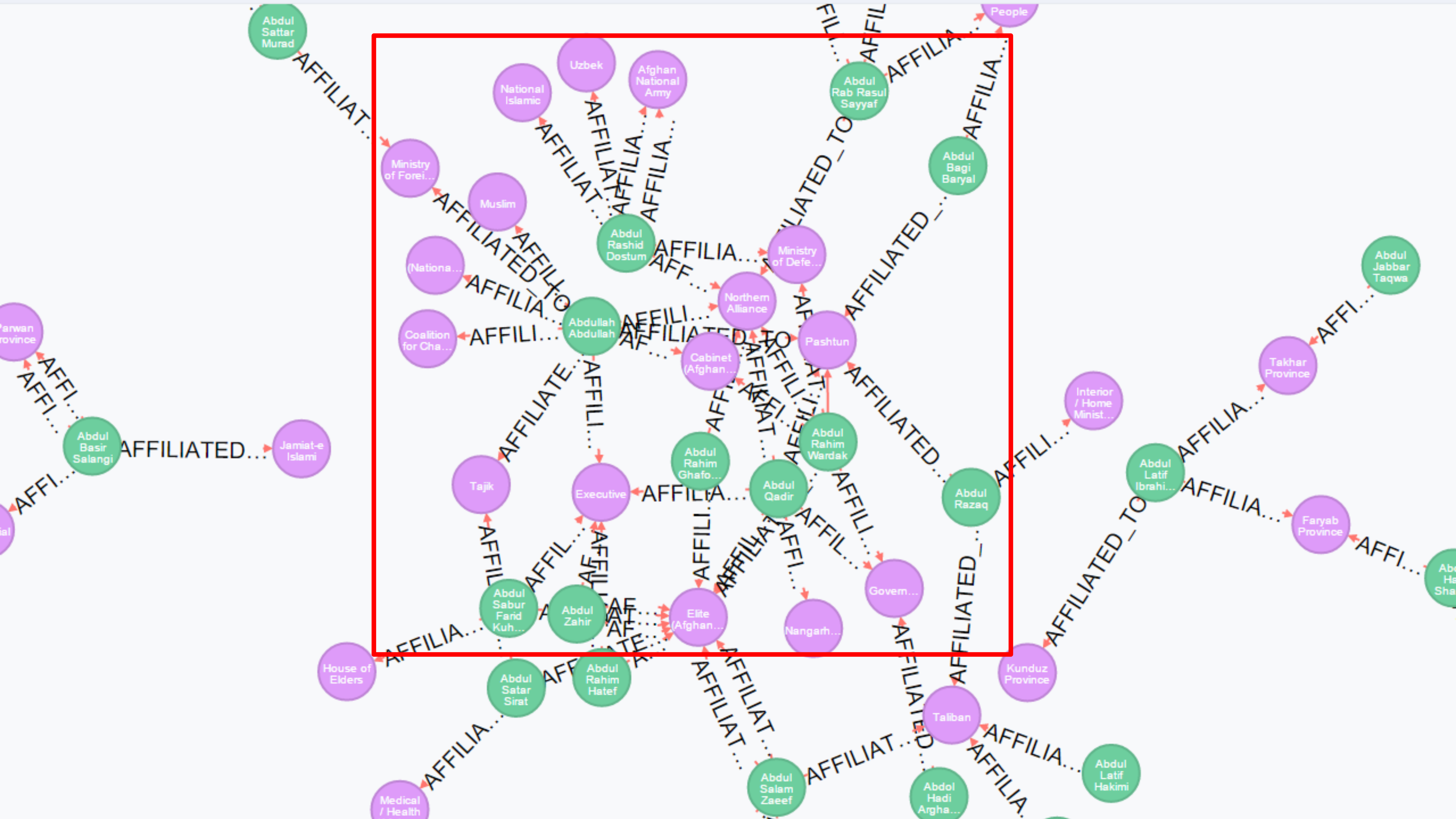
- **Behavioral information**
  - Financial and business transactions
  - Performance in activities
  - Fitness habits
  - Location information



Why Graphs?

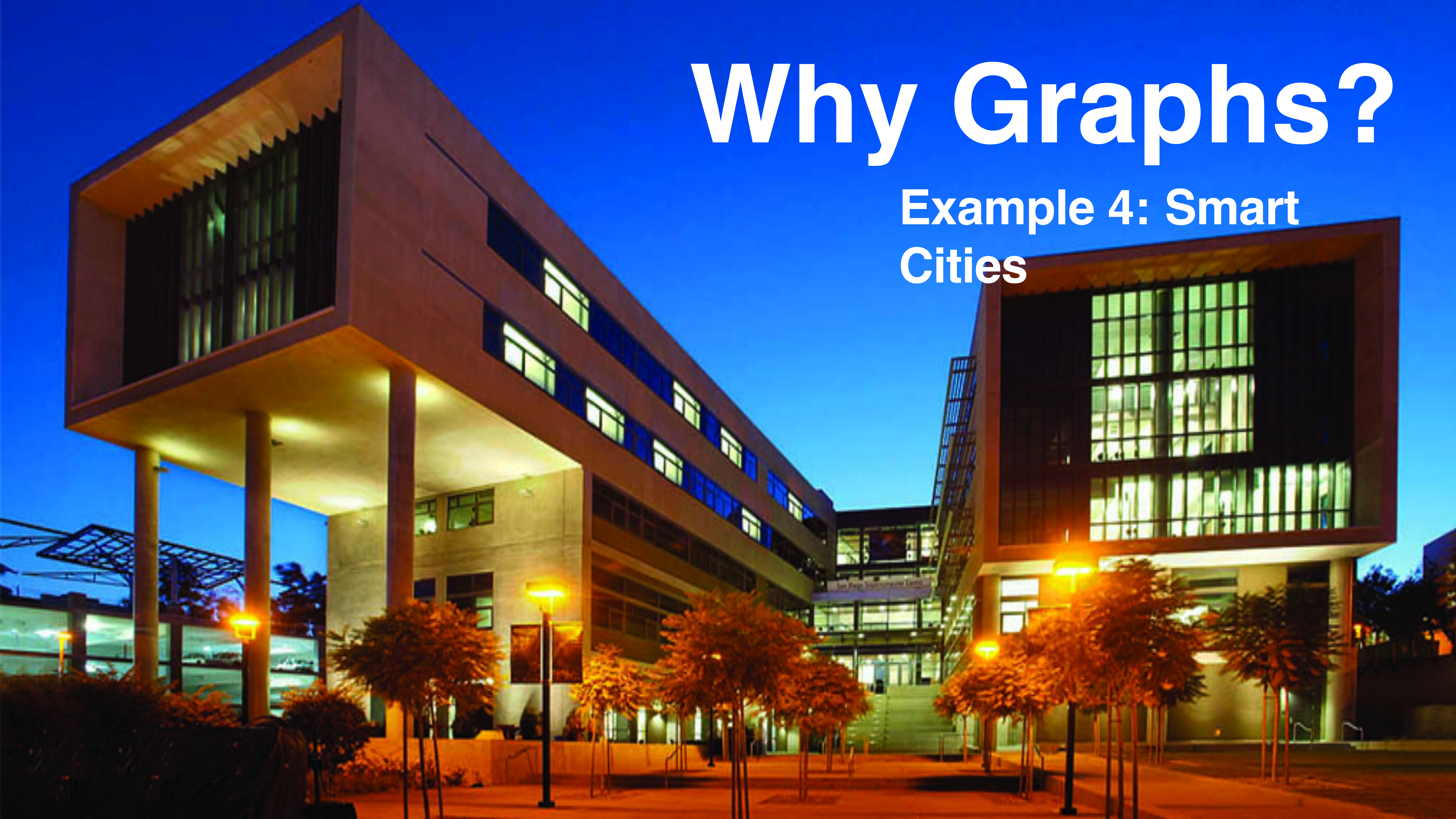
- **Match making problems**
  - Job-candidate pairing
- **Topical influencer analysis**
  - Who would have maximal reach for task X?
- **Situation detection, assessment**
  - Threat detection




Abdul Sattar Murad
Ministry of Forei…
National Islamic
Uzbek
Afghan National Army
Abdul Rab Rasul Sayyaf
People
Abdul Bagi Baryal
Muslim
Abdul Rashid Dostum
Ministry of Defe…
(Nationa…
Northern Alliance
Abdullah Abdullah
Coalition for Cha…
Cabinet (Afghan…
Pashtun
Parwan Province
Abdul Basir Salangi
Jamiat-e Islami
Abdul Jabbar Taqwa
Takhar Province
Interior / Home Minist…
Abdul Rahim Wardak
Abdul Rahim Ghafo…
Tajik
Executive
Abdul Qadir
Abdul Razaq
Abdul Latif Ibrahi…
Faryab Province
Govern…
Abdul Sabur Farid Kuh…
Abdul Zahir
Elite (Afghan…
Nangarh…
House of Elders
Abdul Satar Sirat
Abdul Rahim Hatef
Kunduz Province
Taliban
Abdul Latif Hakimi
Abdul Salam Zaeef
Abdol Hadi Argha…
Medical / Health
AFFILIATED_TO

Why Graphs?

Example 4: Smart Cities

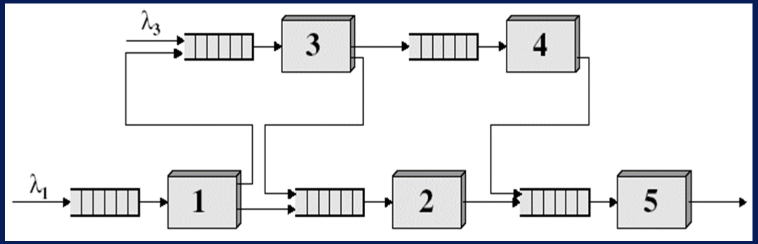


Cities have Networks

- **Multiple interacting networks <u>over the same spatial domain</u>**
  - Transportation networks
    - Multiple modalities
  - Water and sewage network
  - Power transmission network
  - Broadband IP and M2M networks

Networks Should be Modeled

- **Multiple functionalities of each network**
  - Physical infrastructure
  - Commodity flow
    - Material
    - Energy
    - People
- **Need to create “network models”**
  - Analytical traffic model
    - Congestion models



Why Graphs?

- Planning for "smart hubs"
- Estimating congestion patterns
- Demand Response Analyses
- Energy-optimal routing