Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Why Use Tables and Graphs for Knowledge Discovery

JOHN FEO

NORTHWEST INSTITUTE FOR ADVANCED COMPUTING
PACIFIC NORTHWEST NATIONAL LABORATORY

HPC USER FORUM
SEPTEMBER, 2016

Why graphs? Good for representing unstructured data

Mary called her sister Sally to discuss buying her 6-year daughter a pony for Christmas.

1) Mary called Sally
2) Mary has a sister named Sally
3) Sally has a sister named Mary
4) Either Mary or Sally has a daughter
5) The daughter is 6 years old
6) Mary wants to buy a pony

Sally rented Joe's condo in Hawaii for a two week vacation. She paid $1200 rent.

1) Sally traveled to Hawaii
2) Sally vacationed in Hawaii
3) Joe owns a condo
4) Joe's condo is in Hawaii
5) Sally rented Joe's condo
6) Joe rented his condo for $600 per week

?

NAME	SIBLING	CHILD	AGE	CALLED	FUTURE PURCHASES
Mary	Sally	?		Sally	Pony
Sally	Mary	?			
?			6		

Why graphs? Good for representing unstructured data

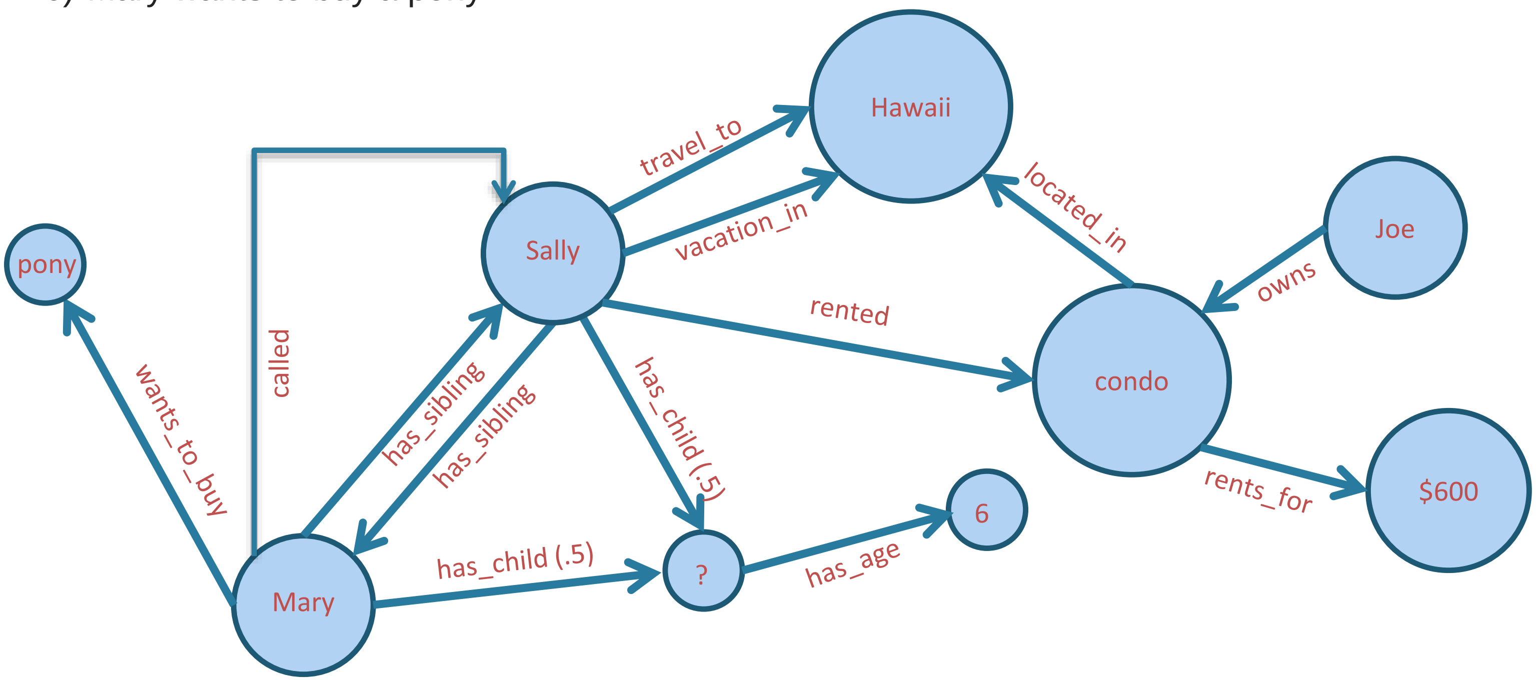
Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Mary called her sister Sally to discuss buying her 6-year daughter a pony for Christmas.

1) Mary called Sally
2) Mary has a sister named Sally
3) Sally has a sister named Mary
4) Either Mary or Sally has a daughter
5) The daughter is 6 years old
6) Mary wants to buy a pony

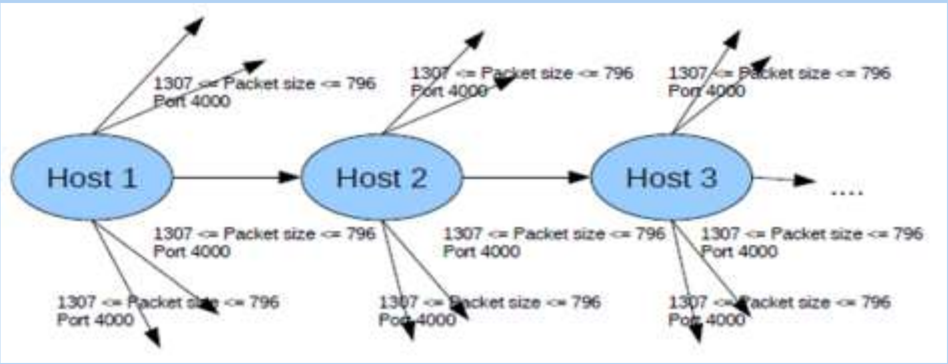
Sally rented Joe's condo in Hawaii for a two week vacation. She paid $1200 rent.

1) Sally traveled to Hawaii
2) Sally vacationed in Hawaii
3) Joe owns a condo
4) Joe's condo is in Hawaii
5) Sally rented Joe's condo
6) Joe rented his condo for $600 per week

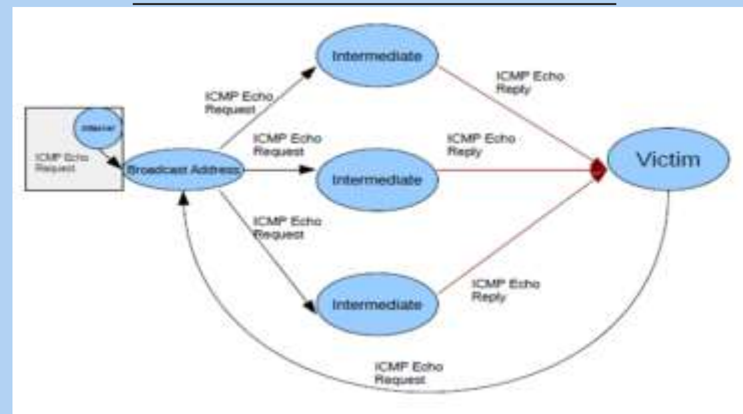


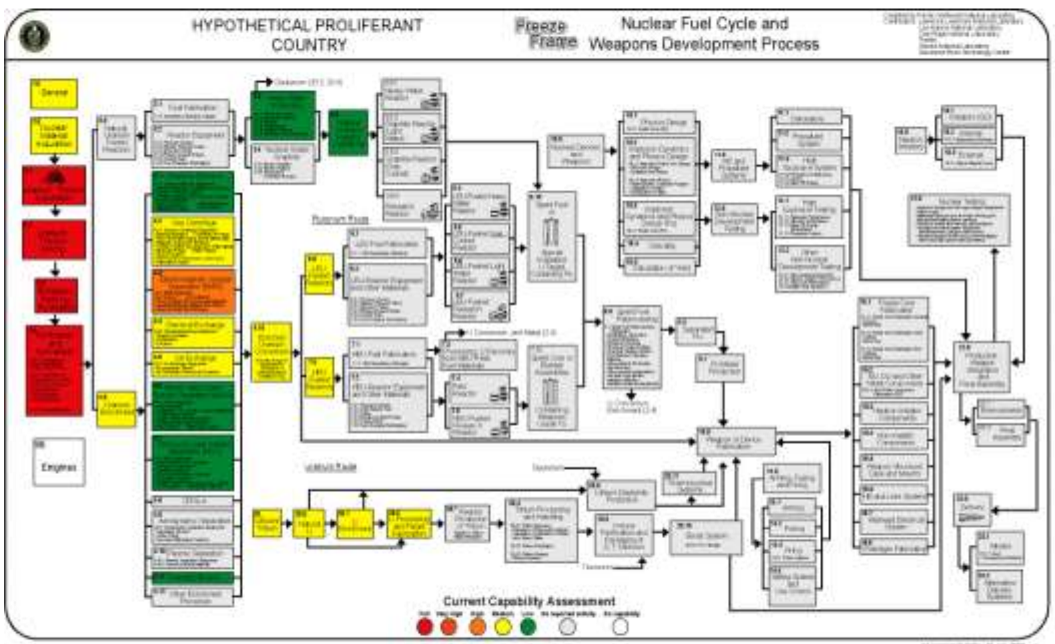
Why graphs? Good for finding patterns

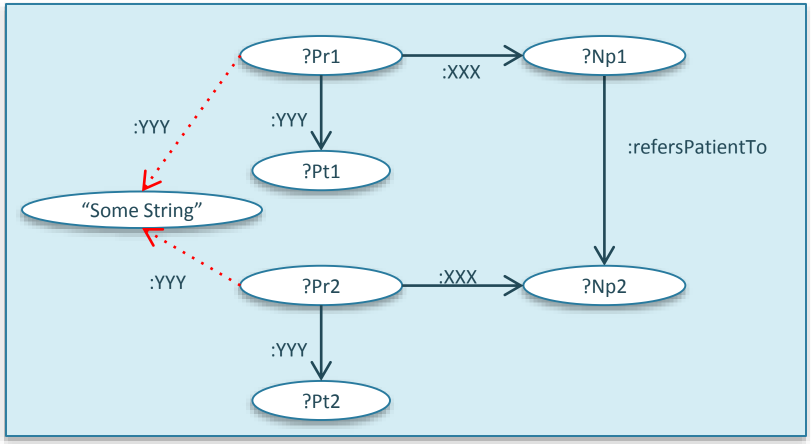
Witty Worm



Distributed DoS Smurf Attack



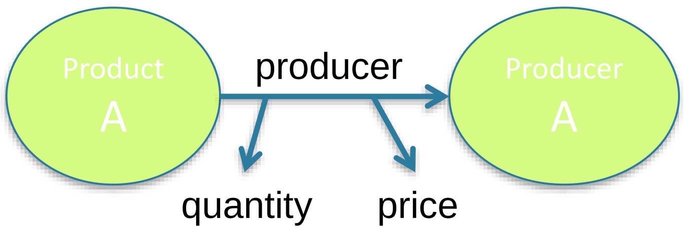


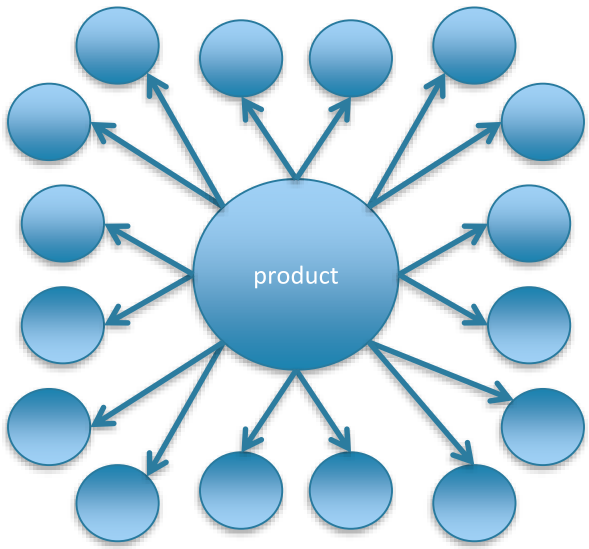


Why not graphs?

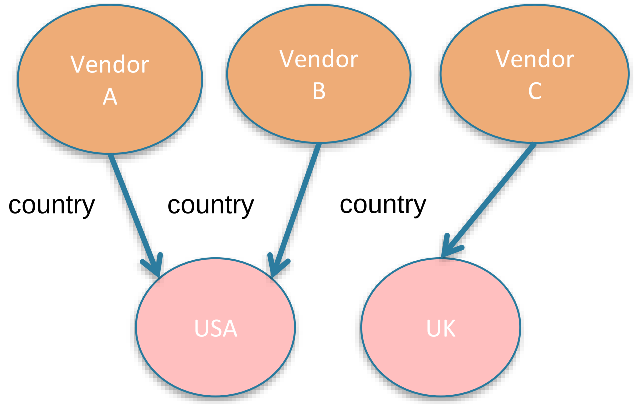
id	srcaddr	dstaddr	protocol	srcport	dstport	start	end
1	62.175.55.106	131.215.122.25	6	25	5641	5682	9084
2	67.72.170.48	131.215.122.25	6	57947	3724	17364	24602
3	67.72.170.48	131.215.122.25	6	60373	3724	28494	31684
4	67.72.170.48	179.137.146.48	6	80	64845	41239	43766
5	67.72.170.48	131.215.122.25	17	58020	9050	45516	46136

Edge attributes





Starification



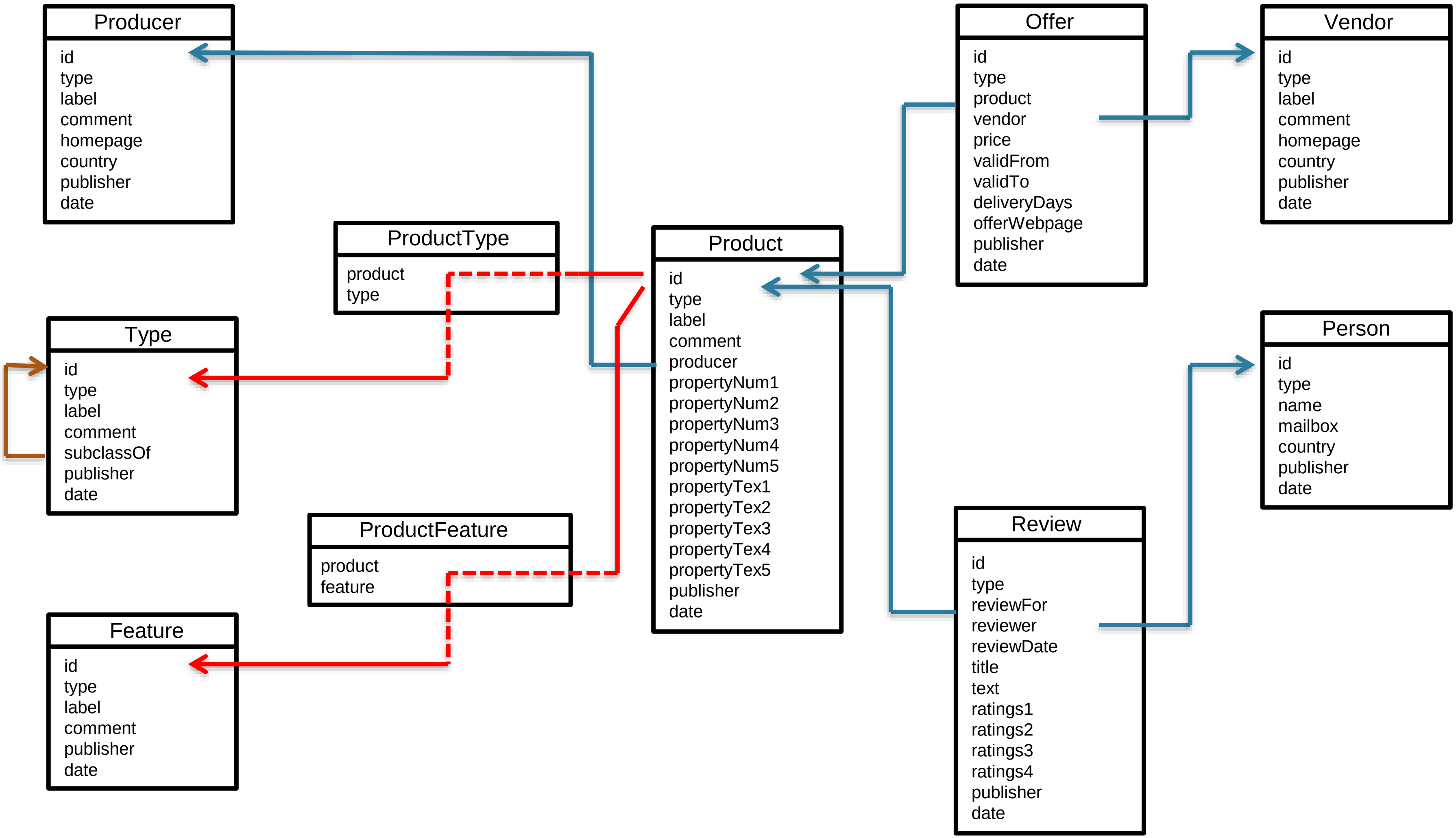
Inefficient selects

So, may be we need …

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Table and graph data structures
- Table operations and graph methods
- Relational and graph/knowledge databases
- New query languages? **No,** just a few extensions
- Disruptive platforms? **No,** just integrate new features
- New workflows? **No,** just new options

Some attributes are edges



Bizer and Schultz, “The Berlin SPARQL Benchmarks,” *International Journal on Semantic Web and Information Systems*, 2009

GraQL – adding a graph view to SQL

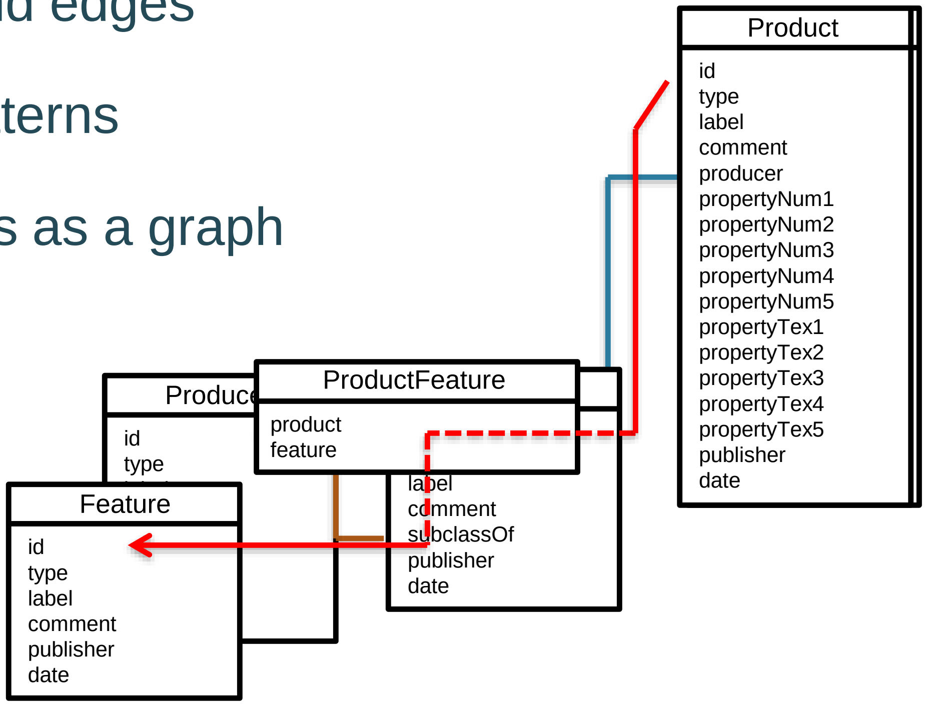
- Extensions to declare vertices and edges
- Extensions to describe graph patterns
- Extensions to return query results as a graph

```
create vertex ProductVtx(id)
from table Product

create vertex ProducerVtx(id)
from table Producer

create vertex TypeVtx(id)
from table Type

create vertex FeatureVtx(id)
from table Feature
```

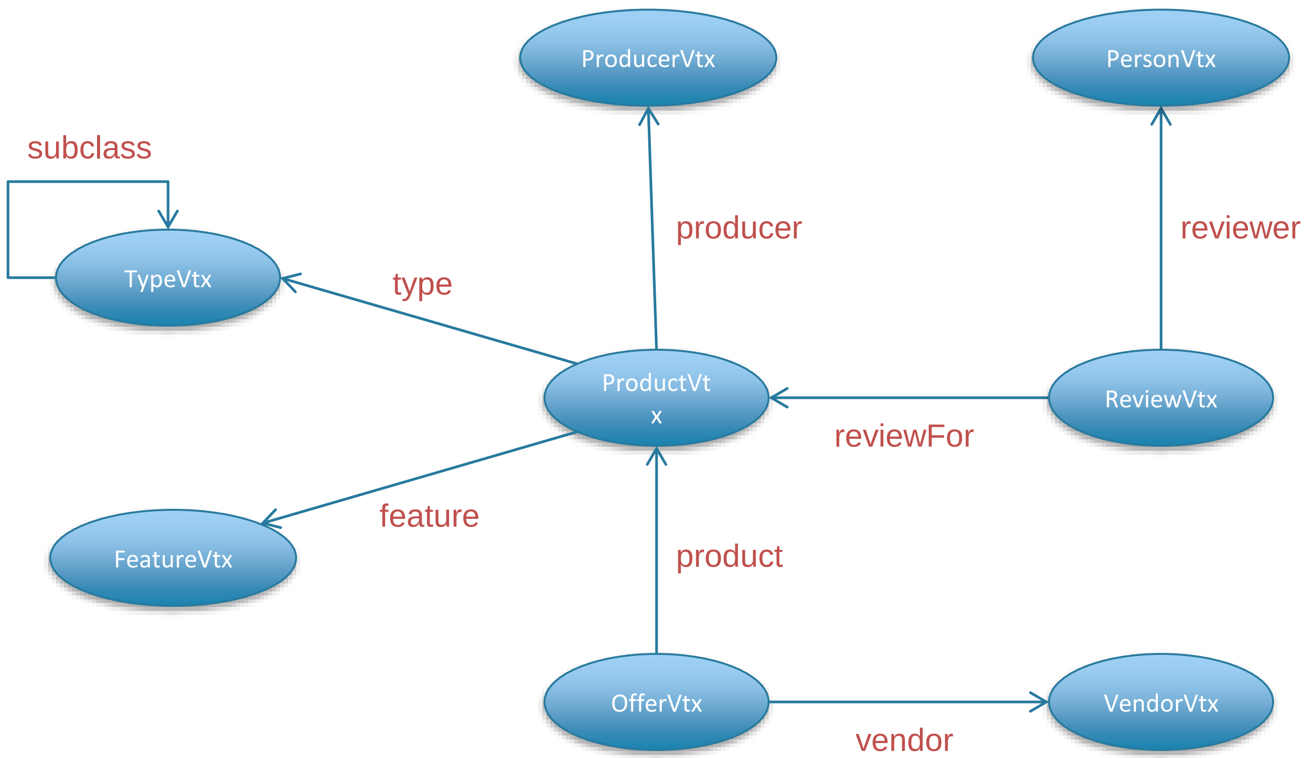



```
create edge feature with vertices (ProductVtx, FeatureVtx)
from table ProductFeature
where ProductFeature.product = ProductVtx.id and ProductFeatures.feature =
FeatureVtx.id
```

Chavarria Castellana, Morari, Haglin, and Feo, *"GraQL: A Query Language for High-Performance Attributed Graph Databases,"* to appear ParLearning 2016.

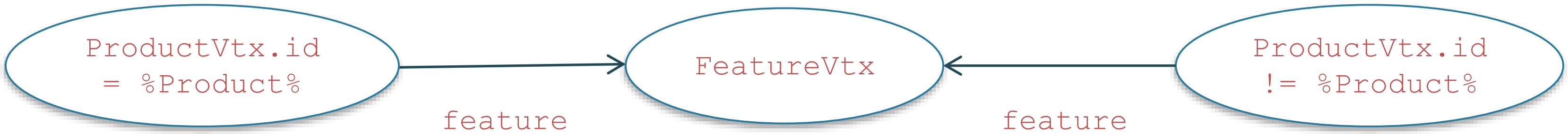
A graph view

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*



A graph makes it easy to find patterns

SELECT THE TOP 10 PRODUCTS MOST SIMILAR TO A SPECIFIC PRODUCT,
RATED BY THE COUNT OF FEATURES THEY HAVE IN COMMON




```
select y.id
from graph ProductVtx(id = %Product%) -feature-> FeatureVtx()
           <-feature- foreach y: ProductVtx(id != %Product%)
into table T1

select top 10
   id, count(*) as sameFeature
from table T1
     group by id
     order by count(*) desc
```

```
SELECT ?otherProduct ?sameFeatures {
  ?otherProduct type Product .
  FILTER(?otherProduct != %Product%)
  {SELECT ?otherProduct (count(%otherFeature) as ?sameFeatures)
     {%Product%     productFeature  ?feature .
      ?otherProduct productFeature  ?otherFeature .
      FILTER(?feature = ?otherFeature)
     }
     Group By ?otherProduct
} }
Order By desc(?sameFeatures) ?otherProduct
Limit 10
```

More complex definitions and queries


```
create vertex server(ipAddr)
from table Netflow
with ip = srcaddr or ip = dstaddr
```

```
create edge comm with vertices (server as src, server as dst)
from table Netflow
with attributes (id, protocol, srcport, dstport, size, start, end)
where src.ipAddr = srcaddr and dst.ipAddr = dstaddr
```

```
from graph
     ( server -comm(protocol = 1)-> foreach ip1: server -comm(protocol = 1)->
       server -comm(protocol = 1)-> foreach ip3: server -comm(protocol = 1)-> ip1
     )
  and
     (ip1 -comm(protocol = 1)-> server -comm(protocol = 1)-> ip3
     )
  and
     (ip1 -comm(protocol = 1)-> server -comm(protocol = 1)-> ip3
     )
```

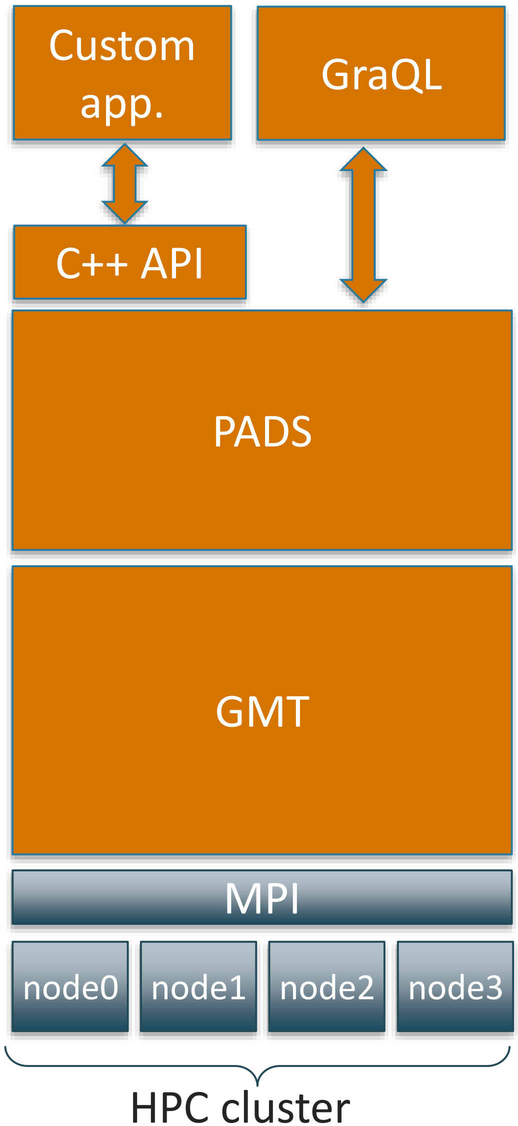
```
-x->             out-edge
<-x-              in-edge
~x->     conditional edge
|-x->             no edge
```

Graph Engine for Multithreaded Systems (GEMS)

Provide an in-memory property-graph database capable of doing deep analytic query processing over multi-terabyte datasets with human-interactive response times on commodity computing parts

- Focus on
  - Blending *relational* and *graph* data representations, methodologies, algorithms, and queries
  - Providing an application programming interface (API) for custom applications using embedded graph databases
  - Supporting data feeds that supply data in batches (e.g., hourly)
  - Serving a range of user profiles: application programmer, subject matter expert analyst

GEMS 2.0



Scalable, in-memory hybrid search engine for pattern discovery

Castellana, et. al, In-Memory Graph Databases for Web-Scale Data. *IEEE Computer* 48(3): 24-35 (2015)

Parallel and Distributed Algorithms and data Structures:

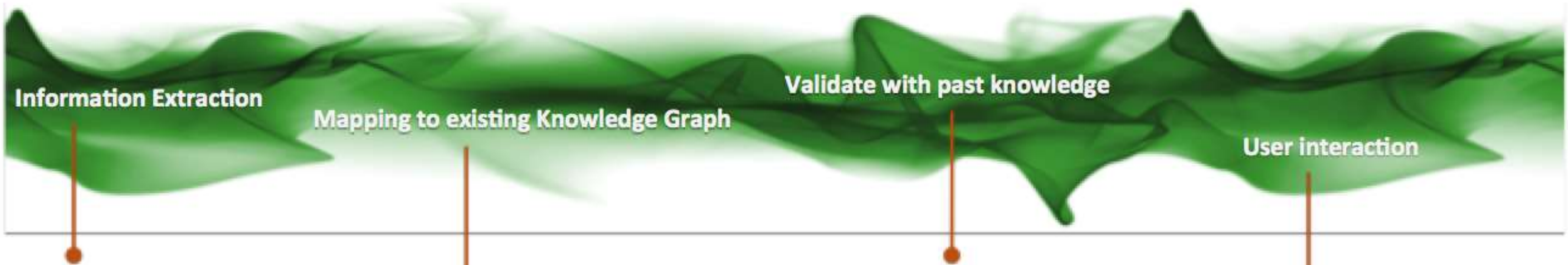
- Table to store tuples
- Index (Graph) for fast neighbours access
- Dictionary for encoding
- CSV and RDF parser

Global Memory and Threading runtime system:

- Runtime library implemented in C
- Requirements: MPI, Linux, x86
  - Single node uses Intel TBB
- Partitioned Global Address Space (PGAS)
- Parallel loops program structure (*parFor*)
- Massive user-level asynchronous tasks
- Software multi-threading to hide latency
- Message aggregation

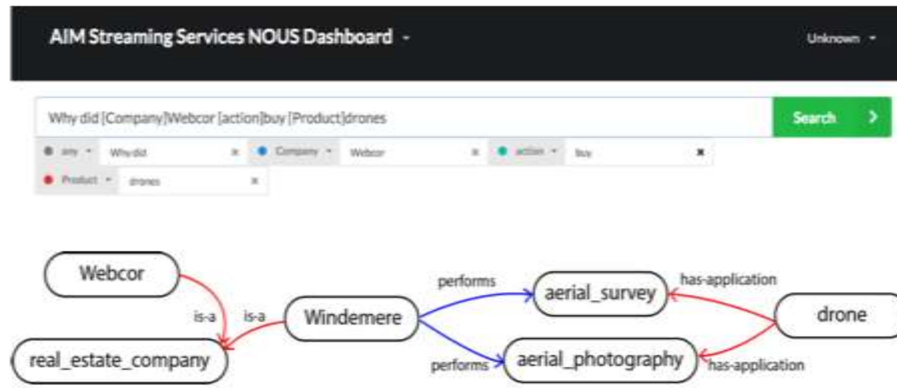
Morari, et.al. Scaling Irregular Applications through Data Aggregation and Software Multithreading. IPDPS 2014: 1126-1135

NOUS: Incremental Knowledge Graphs



- Natural Language Processing using Semantic Role Labeling
- Entity Disambiguation, Entity classification , and Predicate
- Link prediction
- Rule learning
- Rule based verification
- Graph search
- Trend discovery

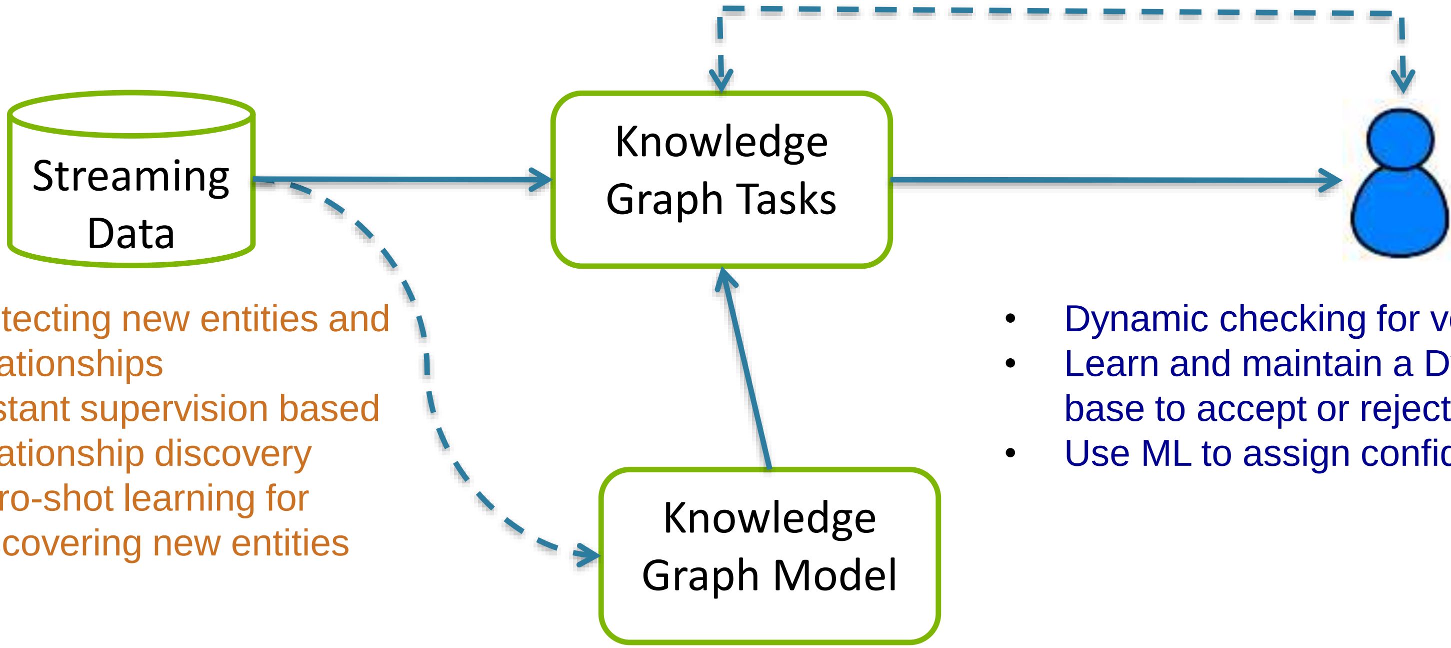
- Incrementally construct KGs from streaming data and multiple sources
- Support advanced trending and explanatory questions
- Automatically generate hypotheses to guide users in information discovery
- Web-based, interactive UI



Zhang B, S Choudhury, et.al. *"Trust from the past: Bayesian Personalized Ranking based Link Prediction in Knowledge Graphs,"* in 2016 SIAM Data Mining Workshop on Mining Networks and Graphs: A Big Data Analytic Challenge

NOUS features

- Continuous pattern (or rult) discovery, search, and resaoning
- Algorithms that monitor emerging “solution sketches” or “pathways”
- Integrates machine-learning models (e.g., LDA, RNN) to steer the search



- Detecting new entities and relationships
- Distant supervision based relationship discovery
- Zero-shot learning for discovering new entities

- Dynamic checking for verification
- Learn and maintain a Dynamic Rule-base to accept or reject facts
- Use ML to assign confidence

- Combine structure and attributes
- Track provenance

Ingest

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Apple's New Challenge: Learning How the U.S. Cracked Its iPhone

SAN FRANCISCO — Now that the United States government has cracked open an iPhone that belonged to a gunman in the San Bernardino, Calif., mass shooting without Apple's help, the tech company is under pressure to find and fix the flaw.

But unlike other cases where security vulnerabilities have cropped up, Apple may face a higher set of hurdles in ferreting out and repairing the particular iPhone hole that the government hacked.

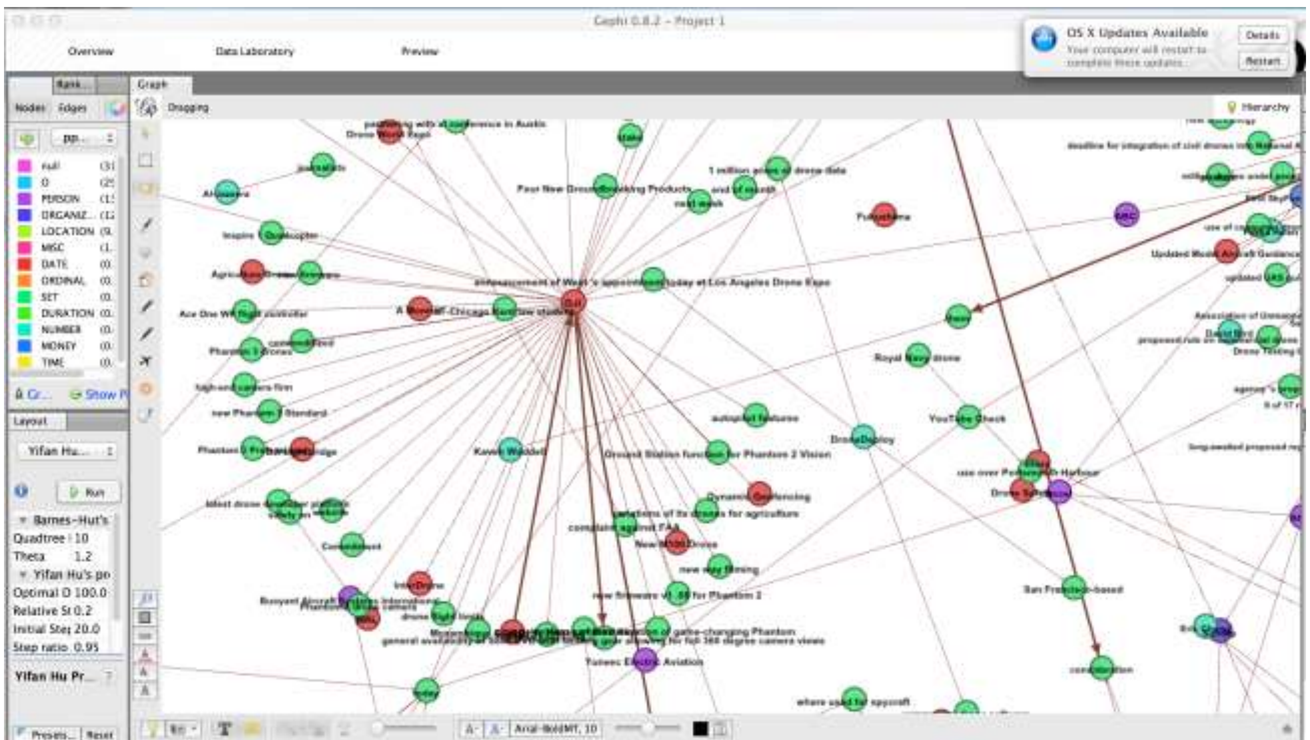
United States government	crack	an iPhone that belonged to a gunman in the San Bernardino
Apple	repair	the particular iPhone hole that the government hacked.
Federal officials	specify	the procedure used to open the iPhone
Federal officials	deny	to specify the procedure used to open the iPhone.
Jay Kaplan, chief executive of Synack and a former NSA analyst.	says	Apple has to earn the trust of Apples customers,"
F.B.I.	cracks	Mr. Farook's
LegbaCore, which previously found and fixed flaws for Apple.	found and fixed	flaws for Apple.

- Built on top of Stanford CoreNLP and OpenIE
- Adds a layer with heuristics to minimize the noise in triples

Post processing raw tripes

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- Entity Disambiguation
- Event Detection
- Relationship Discovery



Navigating data

Find news about autonomous drones

What are their capabilities?

What are their software components?

Ardupilot: An Open Source award winning platform

Navigating data (cont.)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Which vendors are mentioned with Ardupilot?

Arduino

This little board is truly disruptive because it breaks down several barriers that existed in the hardware world. I'll illustrate this with an example: in order to develop a hardware product (a drone, a smart thermostat, a phone...) one could buy an ARM processor on a board and program it

Where is Arduino?

Who sells Arduino?

YKS Upgraded Arduino APM 2.8 Flight Controller Board built-in Compass w/ Shock Absorber for RC Quadcopter
by YKS
3 customer reviews

$48.69 & FREE Shipping on orders over $49. Details

In Stock.
Want it tomorrow, May ... Order within ... and choose One-Day Shipping at checkout.
Details

- APM2.8 Flight Controller, APM2.5 2.8 Upgraded version,newest APM version.
- Includes 3-axis gyro, accelerometer and magnetometer, along with a high-performance barometer. Barometric pressure sensor upgraded to MS5611-01BA03, from Measurement Specialties.
- Onboard 4 MegaByte Dataflash chip for automatic datalogging.Optional off-board GPS, uBlox LEA-6 module with Compass.
- One of the first open source autopilot systems to use Invensense's 6 DoF Accelerometer/Gyro MPU-

Summary

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

- Big data is changing our world
- Winning platforms will implement a varitety of methods and structures
- PNNL has developed *an in-memory, scalable table-graph engine* capable of supporting a trillion attributed vertices and edges
- … and *extended SQL* to define graph views and queries

Gartner Hype Cycle, https://en.wikipedia.org/wiki/Hype_cycle

- … and built *a knowledge graph* capable of processing two million web crawls and analyzing trends, connecting dots, and presenting hypotheses

Contacts

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- GraQL – Daniel Chavaria, daniel.chavaria@pnnl.gov
- GEMS – David Haglin, david.haglin@pnnl.gov
- NOUS – Sutanay Choudhury, sutanay.choudhury@pnnl.gov