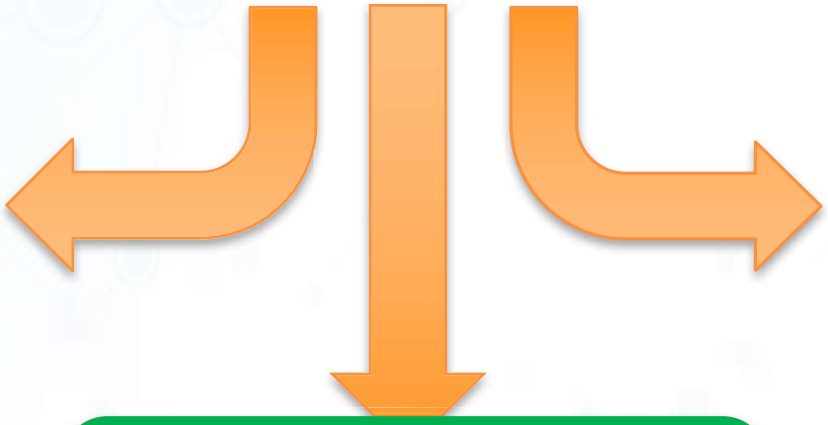
Graph Analytics Outline

- What graph databases are, and how they're different
  - Summary: graph databases are useful when you're at least as interested in the relationships between data items as you are in the data items themselves
- Types of Graph Databases
  - Semantic ("triple store")
  - Property Graph
- Use case examples
  - Summary: you can get new information out of old data

Analytics Architectures Today

Big Data

- Structured
- Semi-structured
- Unstructured



Scale-out Analytics

- Unstructured Data
- Partitionable Datasets
- Hadoop, Spark
- Examples: Hbase, MongoDB, Titan, Neo4j

Data Warehouses BI Tools

- Structured Data
- OLAP Cubes
- Regular Queries, Known Variables
- Example: Oracle Exadata

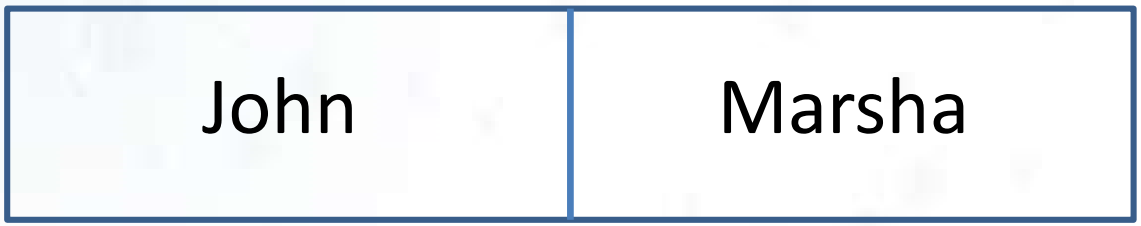
In-memory Analytics

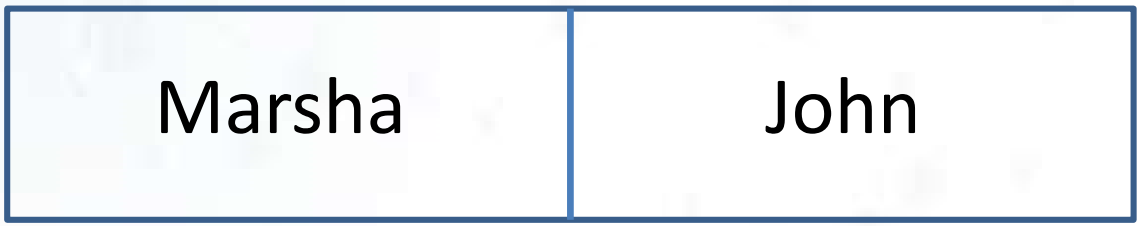
- Structured AND Unstructured Data
- Non-partitionable
- Complex Queries ("Pattern Matching")
- Examples: Jena, Sesame,
- Cray Graph Engine

Graph Databases 101

- What's a graph-oriented database?
- What's a *graph*?
- Why should I be interested?

Example 1

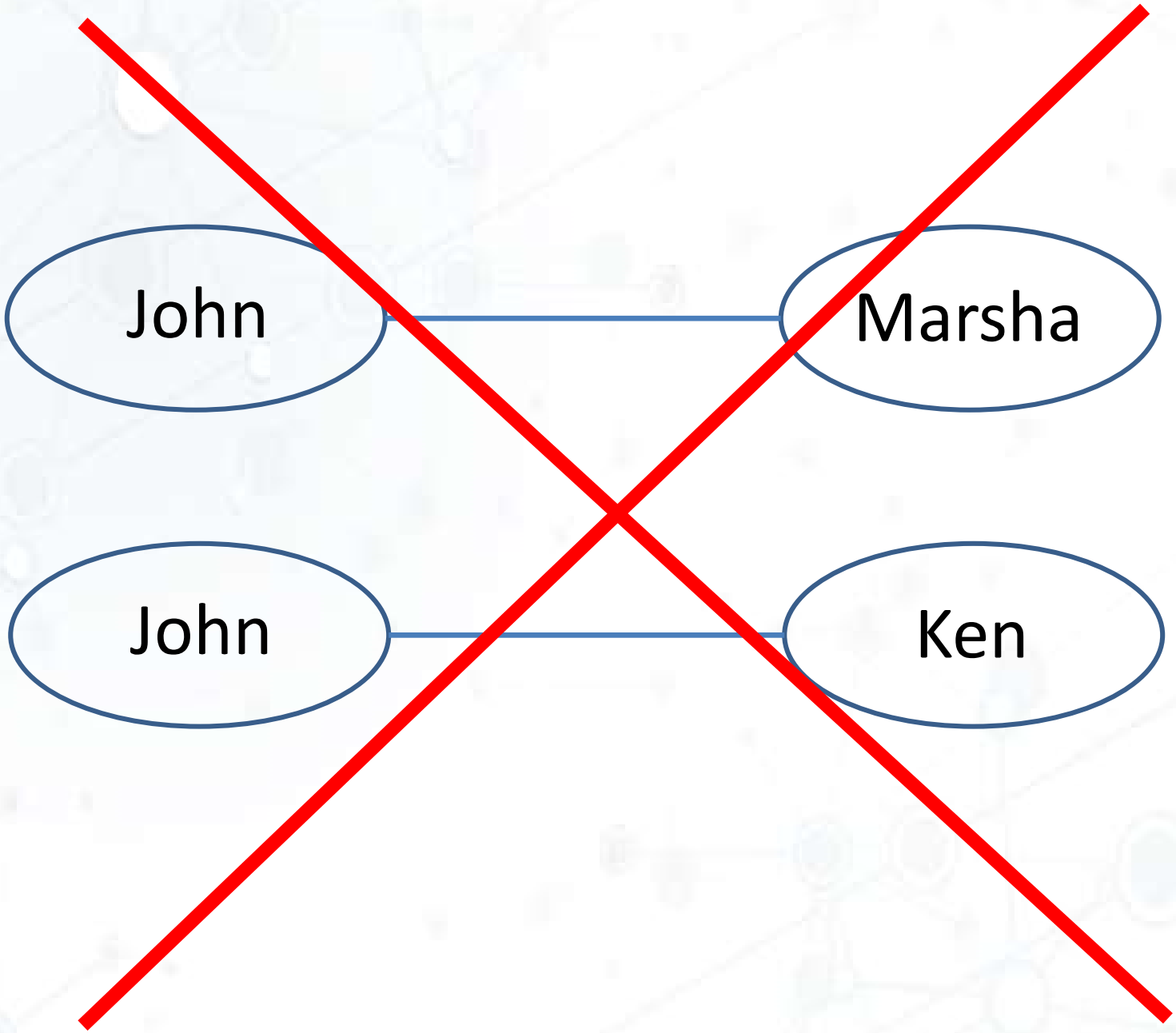

John
Marsha

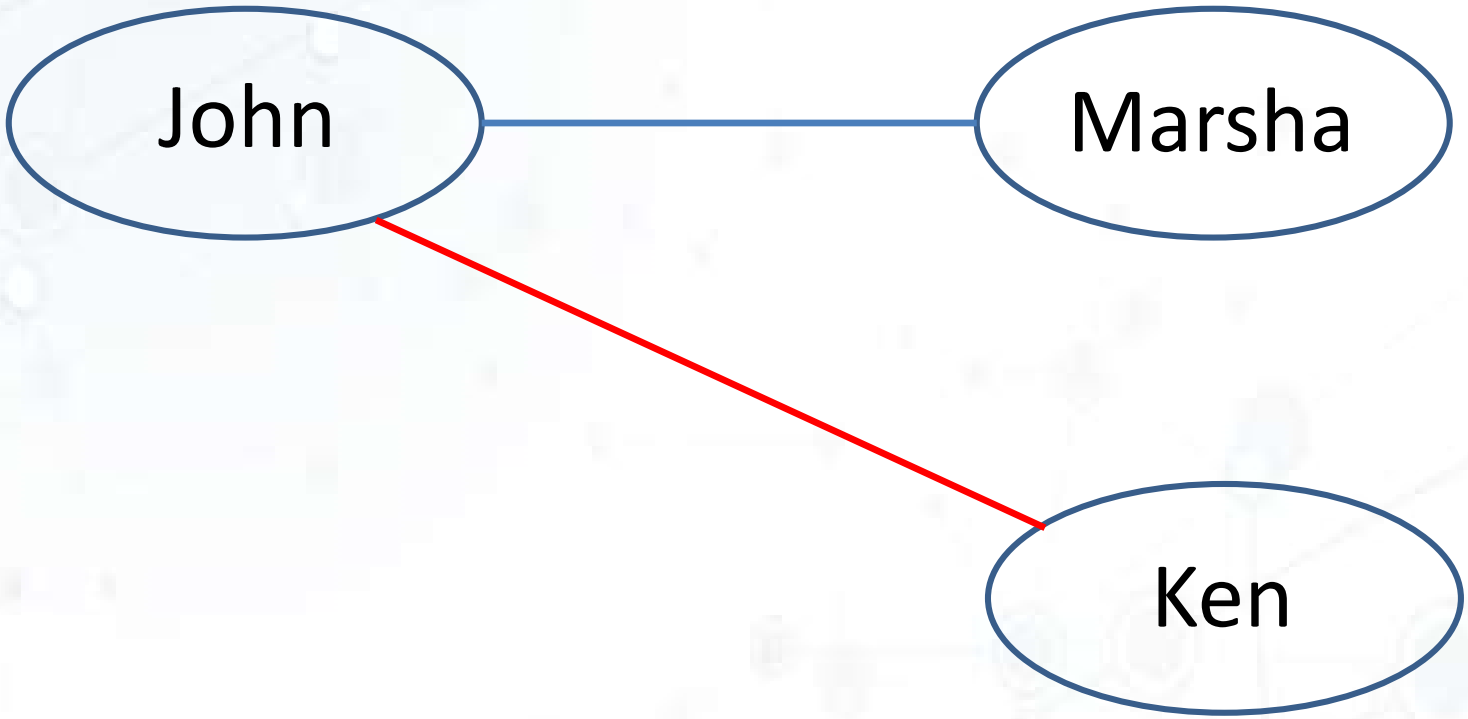

Marsha
John

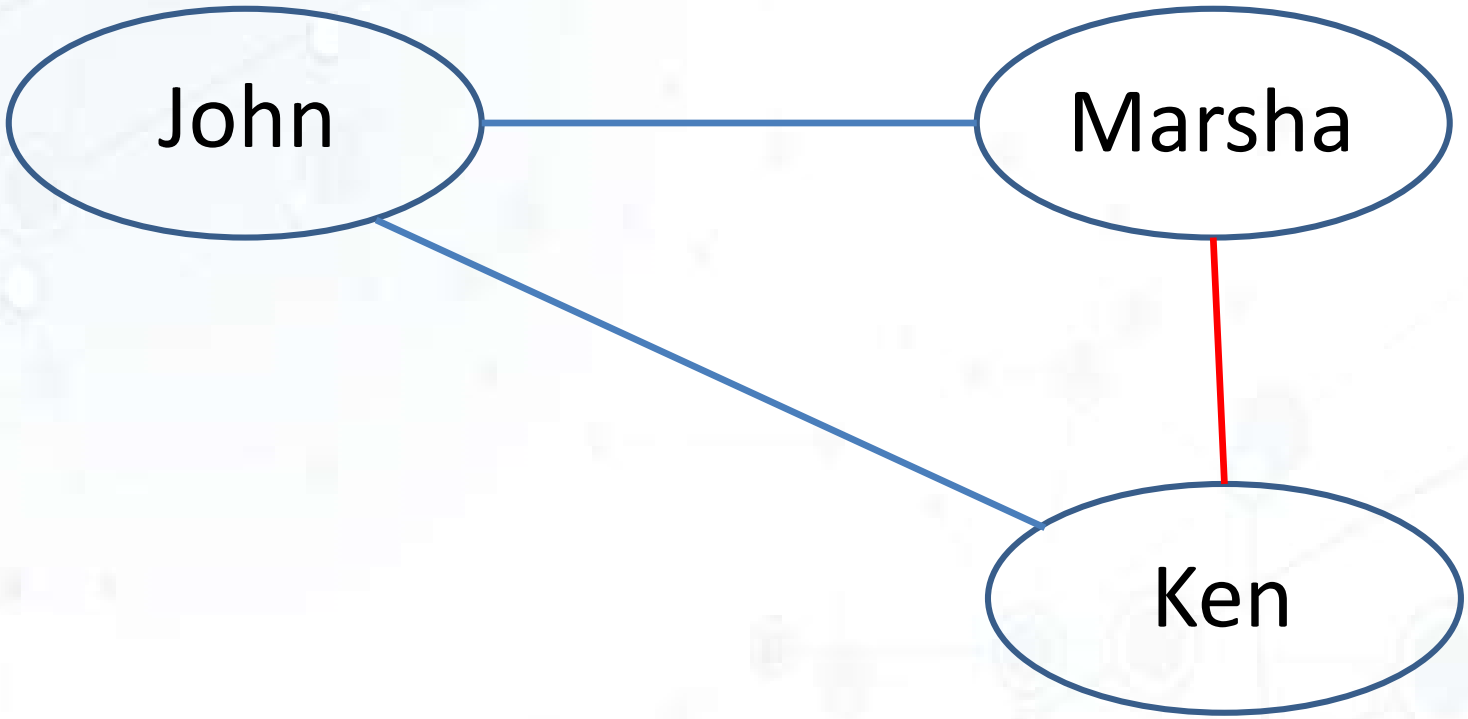
Alice	Bob
Henry	Joanne
Charlie	Elaine
Pam	Oscar
Bob	Henry
Mike	Nick
Bob	Elaine
Joanne	Ken
Fred	Bob
Joanne	Pam
Iris	Bob
Gail	Bob
Iris	Gail
Iris	Henry
Quentin	Joanne
Nick	Laura
Fred	Elaine
Oscar	Mike
Joanne	Nick
Charlie	Dan
Pam	Mike
Iris	Elaine
Oscar	Nick
Ken	Quentin
Ken	Nick
Bob	Dan
Ken	Mike
Bob	Charlie
Nick	Quentin
Gail	Henry
Pam	Nick
Fred	Gail
Oscar	Ken
Elaine	Henry
Quentin	Pam

Alice	Bob
Henry	Joanne
Charlie	Elaine
Pam	Oscar
Bob	Henry
Mike	Nick
Bob	Elaine
Joanne	Ken
Fred	Bob
Joanne	Pam
Iris	Bob
Gail	Bob
Iris	Gail
Iris	Henry
Quentin	Joanne
Nick	Laura
Fred	Elaine
Oscar	Mike
Joanne	Nick
Charlie	Dan
Pam	Mike
Iris	Elaine
Oscar	Nick
Ken	Quentin
Ken	Nick
Bob	Dan
Ken	Mike
Bob	Charlie
Nick	Quentin
Gail	Henry
Pam	Nick
Fred	Gail
Oscar	Ken
Elaine	Henry
Quentin	Pam

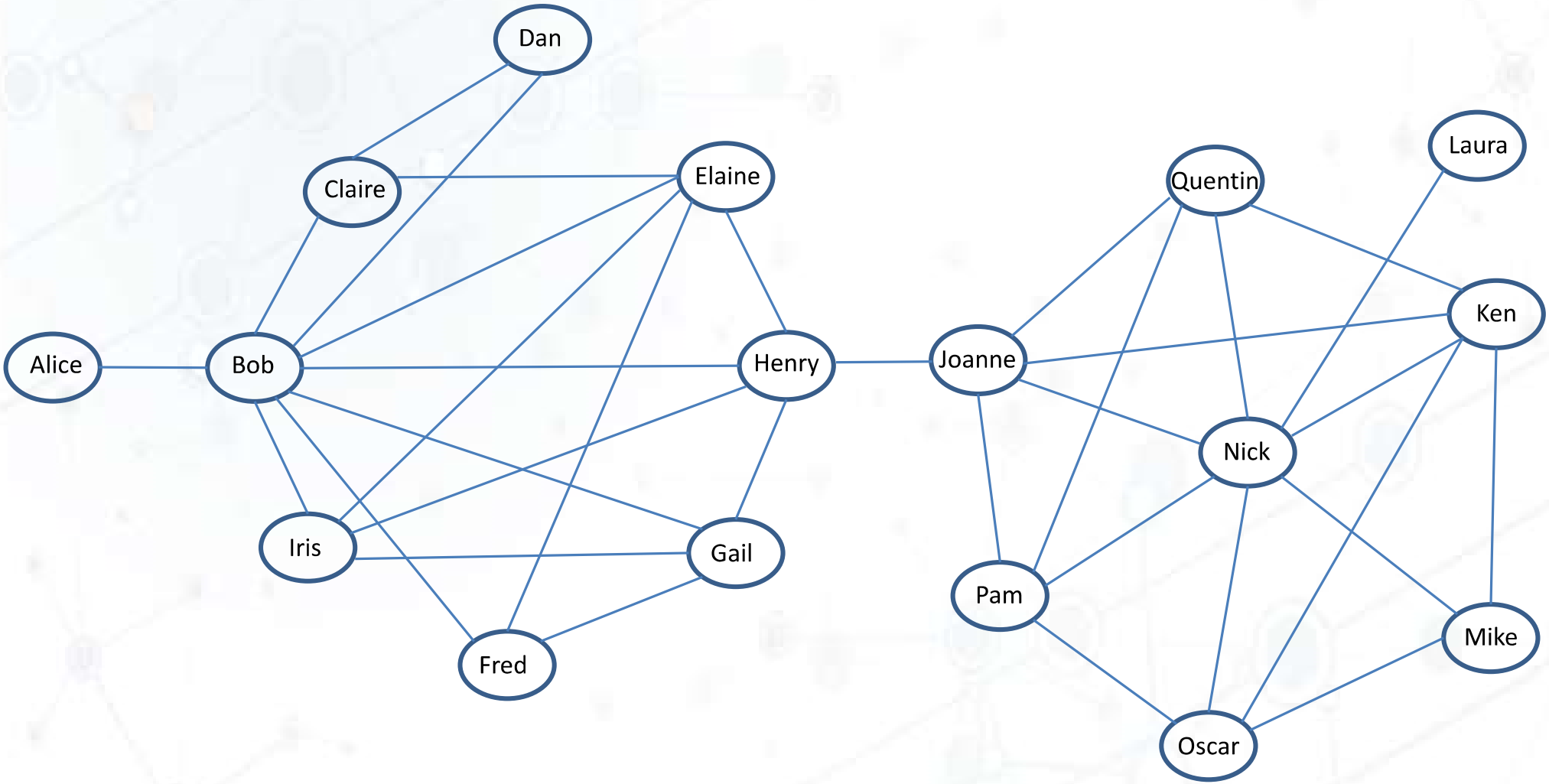

John
Marsha


John
Marsha
John
Ken


John
Marsha
Ken


John
Marsha
Ken

Alice	Bob
Henry	Joanne
Charlie	Elaine
Pam	Oscar
Bob	Henry
Mike	Nick
Bob	Elaine
Joanne	Ken
Fred	Bob
Joanne	Pam
Iris	Bob
Gail	Bob
Iris	Gail
Iris	Henry
Quentin	Joanne
Nick	Laura
Fred	Elaine
Oscar	Mike
Joanne	Nick
Charlie	Dan
Pam	Mike
Iris	Elaine
Oscar	Nick
Ken	Quentin
Ken	Nick
Bob	Dan
Ken	Mike
Bob	Charlie
Nick	Quentin
Gail	Henry
Pam	Nick
Fred	Gail
Oscar	Ken
Elaine	Henry
Quentin	Pam


Dan
Claire
Elaine
Alice
Bob
Henry
Iris
Gail
Fred
Laura
Quentin
Ken
Joanne
Nick
Pam
Mike
Oscar

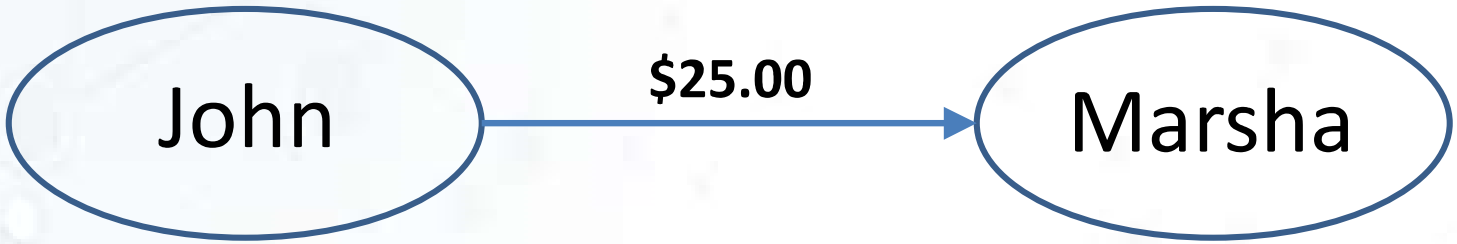
Example 2

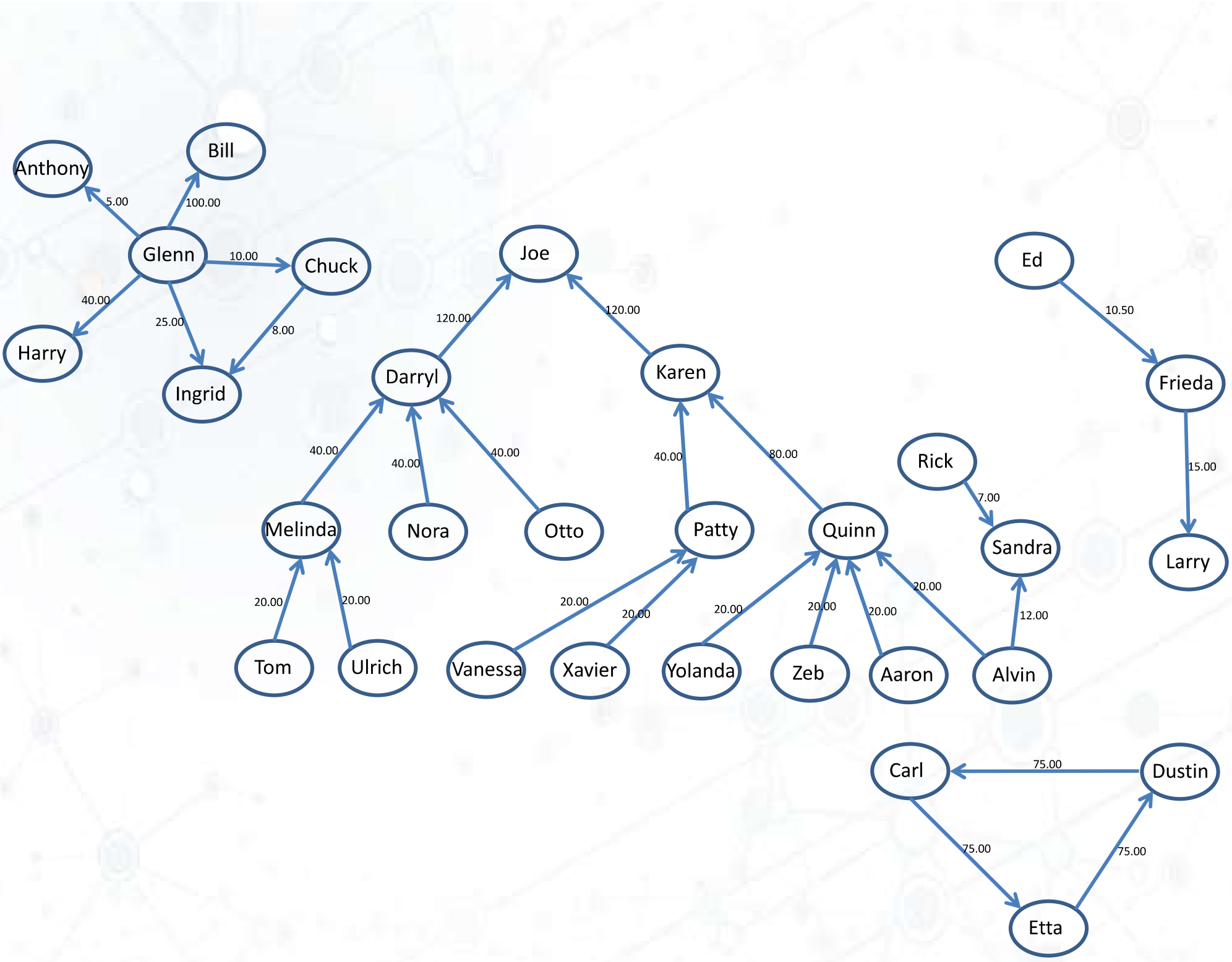
John	25.00	Marsha

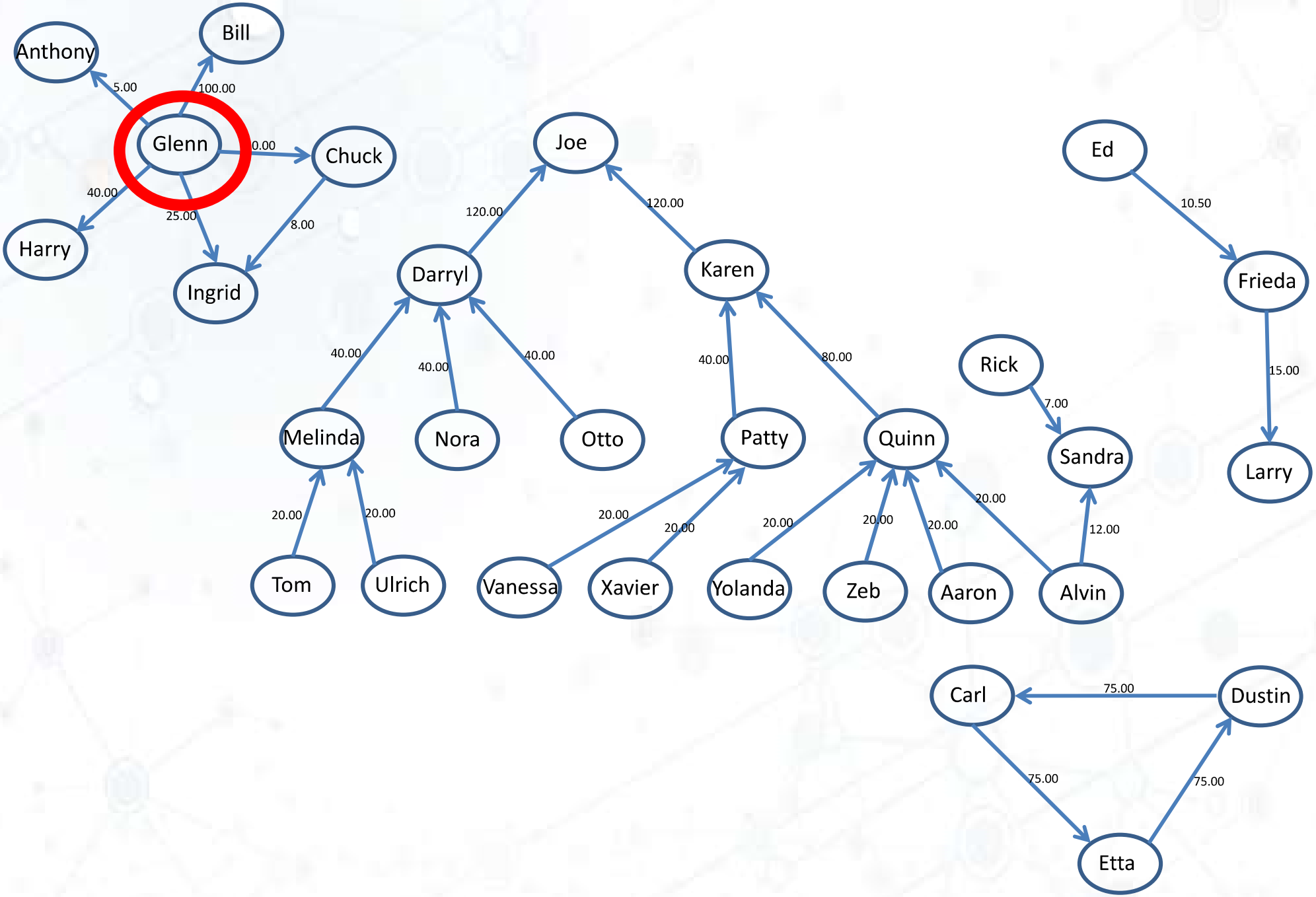
Glenn	5.00	Anthony
Zeb	20.00	Quinn
Ed	10.50	Frieda
Ulrich	20.00	Melinda
Glenn	10.00	Chuck
Rick	7.00	Sandra
Chuck	8.00	Ingrid
Frieda	15.00	Larry
Glenn	100.00	Bill
Vanessa	20.00	Patty
Karen	120.00	Joe
Yolanda	20.00	Quinn
Alvin	12.00	Sandra
Xavier	20.00	Patty
Glenn	40.00	Harry
Melinda	40.00	Darryl
Dustin	75.00	Carl
Alvin	20.00	Quinn
Nora	40.00	Darryl
Etta	75.00	Dustin
Glenn	25.00	Ingrid
Otto	40.00	Darryl
Aaron	20.00	Quinn
Quinn	80.00	Karen
Carl	75.00	Etta
Patty	40.00	Karen
Darryl	120.00	Joe
Tom	20.00	Melinda

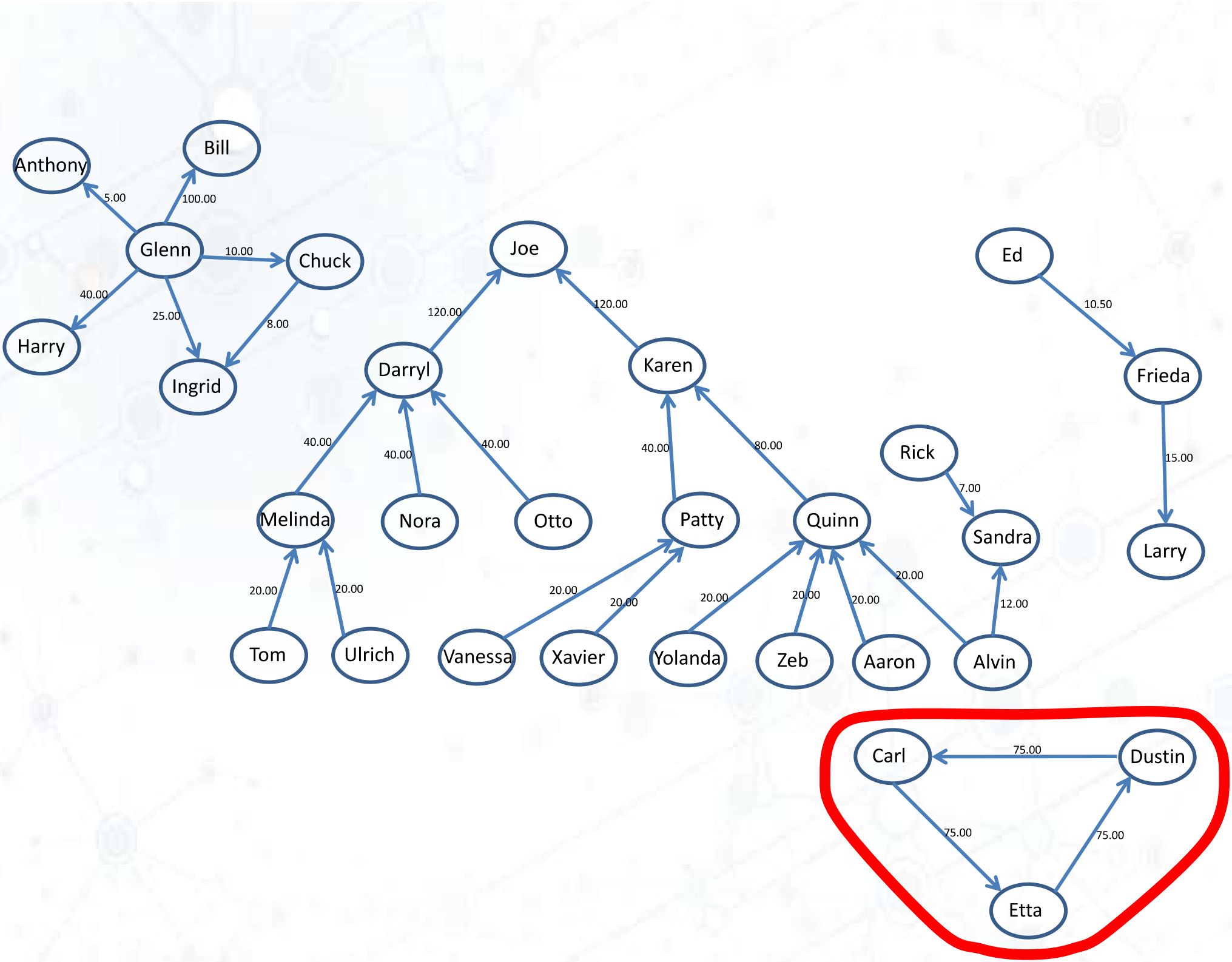
Glenn	5.00	Anthony
Zeb	20.00	Quinn
Ed	10.50	Frieda
Ulrich	20.00	Melinda
Glenn	10.00	Chuck
Rick	7.00	Sandra
Chuck	8.00	Ingrid
Frieda	15.00	Larry
Glenn	100.00	Bill
Vanessa	20.00	Patty
Karen	120.00	Joe
Yolanda	20.00	Quinn
Alvin	12.00	Sandra
Xavier	20.00	Patty
Glenn	40.00	Harry
Melinda	40.00	Darryl
Dustin	75.00	Carl
Alvin	20.00	Quinn
Nora	40.00	Darryl
Etta	75.00	Dustin
Glenn	25.00	Ingrid
Otto	40.00	Darryl
Aaron	20.00	Quinn
Quinn	80.00	Karen
Carl	75.00	Etta
Patty	40.00	Karen
Darryl	120.00	Joe
Tom	20.00	Melinda

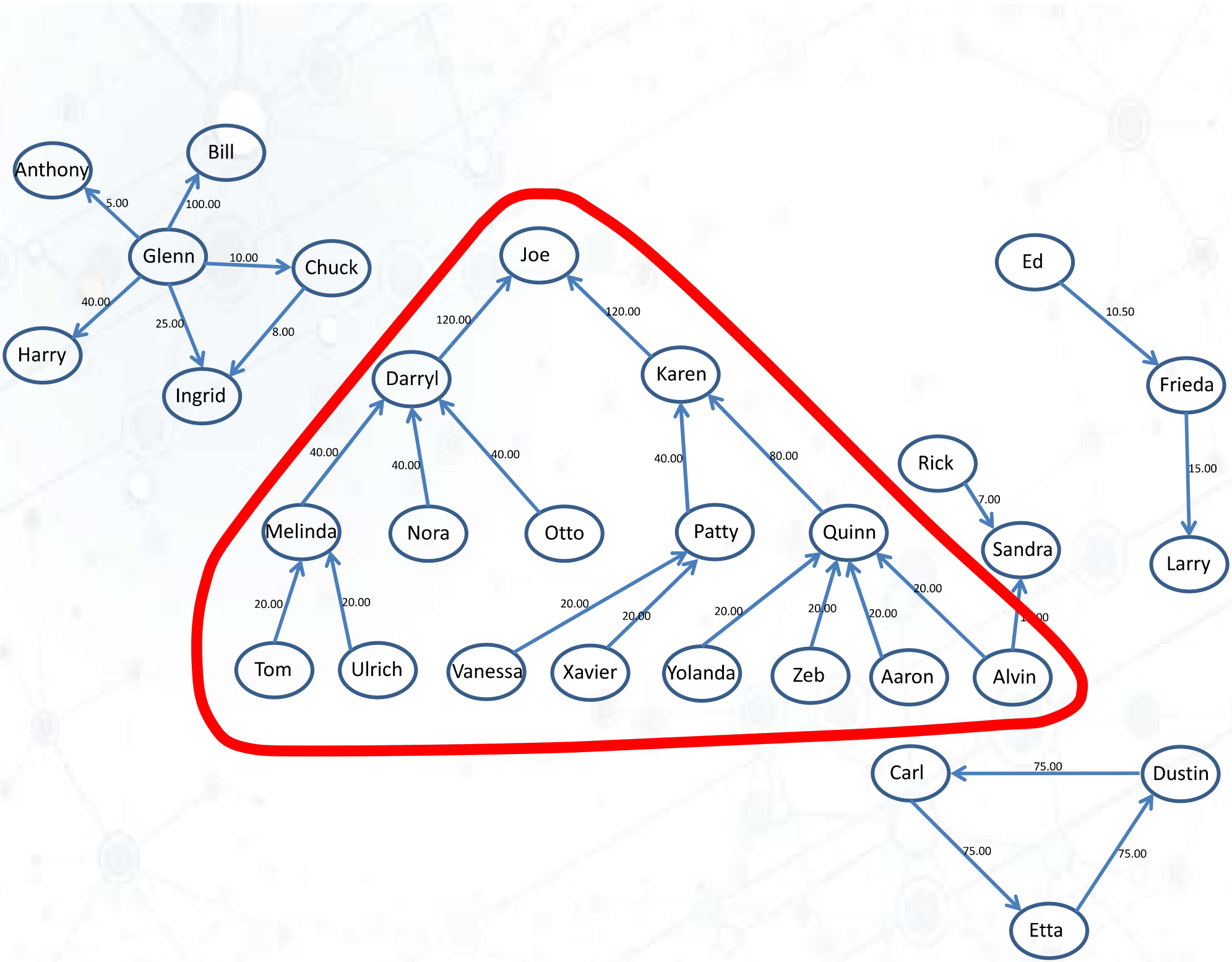
Glenn	**5.00**	Anthony
Zeb	20.00	Quinn
Ed	10.50	Frieda
Ulrich	20.00	Melinda
Glenn	**10.00**	Chuck
Rick	7.00	Sandra
Chuck	8.00	Ingrid
Frieda	15.00	Larry
Glenn	**100.00**	Bill
Vanessa	20.00	Patty
Karen	120.00	Joe
Yolanda	20.00	Quinn
Alvin	12.00	Sandra
Xavier	20.00	Patty
Glenn	**40.00**	Harry
Melinda	40.00	Darryl
Dustin	75.00	Carl
Alvin	20.00	Quinn
Nora	40.00	Darryl
Etta	75.00	Dustin
Glenn	**25.00**	Ingrid
Otto	40.00	Darryl
Aaron	20.00	Quinn
Quinn	80.00	Karen
Carl	75.00	Etta
Patty	40.00	Karen
Darryl	120.00	Joe
Tom	20.00	Melinda


John
$25.00
Marsha


Anthony
Bill
5.00
100.00
Glenn
10.00
Chuck
40.00
25.00
8.00
Harry
Ingrid
Joe
120.00
120.00
Darryl
Karen
40.00
40.00
40.00
40.00
80.00
Melinda
Nora
Otto
Patty
Quinn
20.00
20.00
20.00
20.00
20.00
20.00
20.00
Tom
Ulrich
Vanessa
Xavier
Yolanda
Zeb
Aaron
Alvin
Ed
10.50
Frieda
15.00
Larry
Rick
7.00
Sandra
12.00
Carl
75.00
Dustin
75.00
75.00
Etta

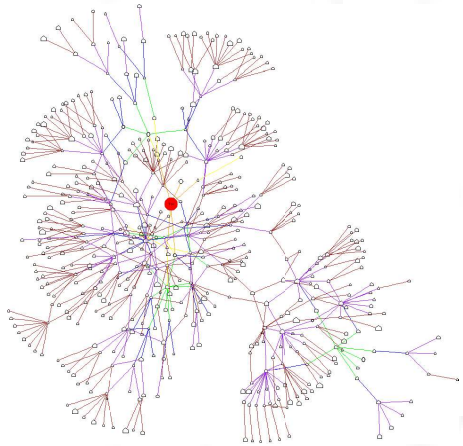

Anthony
Bill
Glenn
Chuck
Harry
Ingrid
5.00
100.00
0.00
40.00
25.00
8.00
Joe
Darryl
Karen
120.00
120.00
40.00
40.00
40.00
40.00
80.00
Melinda
Nora
Otto
Patty
Quinn
20.00
20.00
20.00
20.00
20.00
20.00
20.00
Tom
Ulrich
Vanessa
Xavier
Yolanda
Zeb
Aaron
Alvin
Ed
Frieda
Larry
10.50
15.00
Rick
Sandra
7.00
12.00
Carl
Dustin
Etta
75.00
75.00
75.00


Anthony
Bill
5.00
100.00
Glenn
10.00
Chuck
40.00
25.00
8.00
Harry
Ingrid
Joe
120.00
120.00
Darryl
Karen
40.00
40.00
40.00
40.00
80.00
Melinda
Nora
Otto
Patty
Quinn
20.00
20.00
20.00
20.00
20.00
20.00
20.00
Tom
Ulrich
Vanessa
Xavier
Yolanda
Zeb
Aaron
Alvin
Ed
10.50
Frieda
15.00
Larry
Rick
7.00
Sandra
12.00
Carl
75.00
Dustin
75.00
75.00
Etta


Anthony
Bill
5.00
100.00
Glenn
10.00
Chuck
40.00
25.00
8.00
Harry
Ingrid
Joe
120.00
120.00
Darryl
Karen
40.00
40.00
40.00
40.00
80.00
Melinda
Nora
Otto
Patty
Quinn
20.00
20.00
20.00
20.00
20.00
20.00
20.00
Tom
Ulrich
Vanessa
Xavier
Yolanda
Zeb
Aaron
Alvin
Ed
10.50
Frieda
15.00
Larry
Rick
7.00
Sandra
Carl
75.00
Dustin
75.00
75.00
Etta

Takeaway #1

Relational tables can hold a lot of information and give you rapid, sophisticated access to it.



Graphs are the best representation to use if you are at least as interested in the *pattern of relationships between data items* as you are in the data itself.

Takeaway #2

Things that are hard to spot in a small table are slow to compute in a relational database.

Things that are easy to spot in a small graph can be computed rapidly in a graph-oriented database.

So graph is different, valuable and should be added to advanced analytics environments!

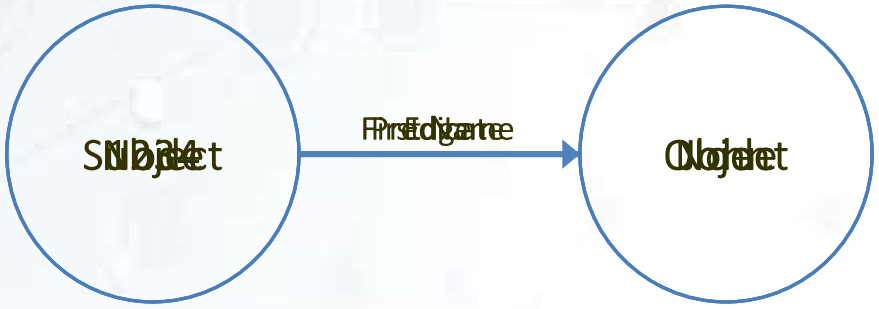
But there are different approaches to graph

- Semantic databases (”triple stores”)
- Property graphs

Semantic databases

- Practically synonymous with “RDF triples” databases
- A standard, subject-predicate-object way of representing every fact in the database
- Based on W3C standards
- Less spatially efficient than a relational database

Simple Graph
“RDF Triple”



RDF Triple		
Subject	Predicate	Object
1234	First Name	John

Semantic Formats and Languages

RDF (Resource Description Framework)

- Stores each data item as a triple containing a subject-predicate –-object
- Self-defining — You can easily figure out what a triple is talking about
- Unique across the entire Internet
- Inherently graph- oriented
- Works with SPARQL!

SPARQL (SPARQL Protocol and RDF Query Language)

- Built to access RDF data
- Graph oriented
- Fully specified query language, with operators: FILTER, GROUP BY, ORDER BY, UNION, OPTIONAL, etc.
- Syntactically quite similar to SQL —which people are familiar with
- W3C standard, with a full suite of compliance tests

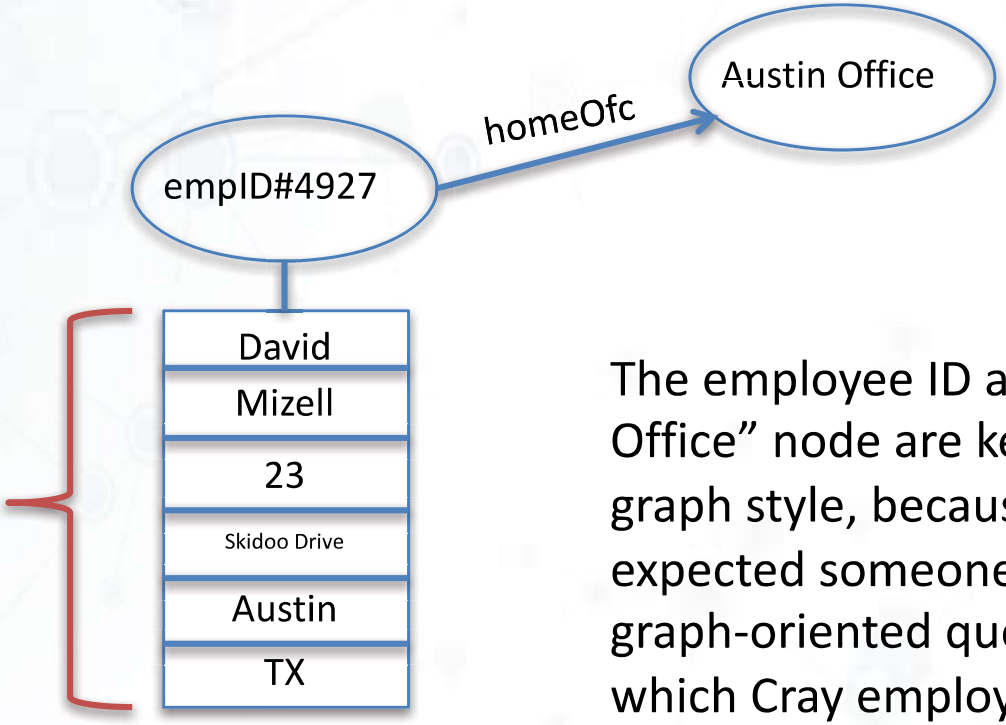
Property graphs

- Designed to be a compromise between semantic databases and relational database
- Data that you don’t expect will ever be linked to in any kind of graph analysis can be attached to a node or an edge like a little relational tuple

PROPERTY GRAPH

So this set of "properties" of the employee ID node is attached to the node in such a way that they are easy to access only from the node.

They are not part of the graph, because they are never expected to be linked to anything else in the graph.



The employee ID and "Austin Office" node are kept linked in graph style, because it is expected someone will ask graph-oriented questions about which Cray employee is in which office.

Property Graph Languages

Cypher (Neo4J)

- Can ask about linked data or about properties attached to a node or an edge.

Gremlin (Titan)

- Can access property lists (in their format) or linked data

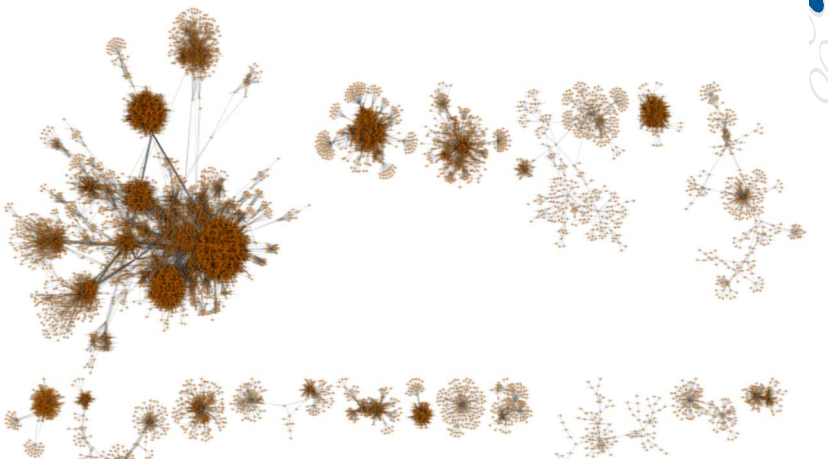
Scala (GraphX)

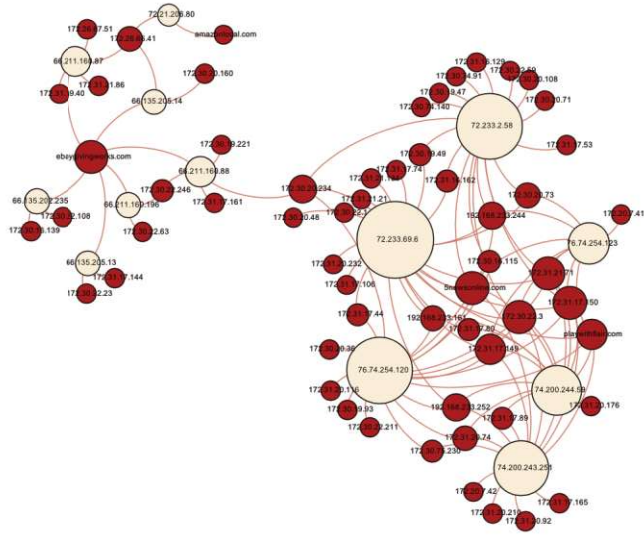
- Library of Graph primitives to manipulate Spark data structures (Graphframes)

Graph Analytics Use Cases

Cybersecurity Use Case: Discovering new Cyber Threats

- Goal: Proactively identify unknown cyber threats by examining all possible relationships
- Data sets: IP, MAC, BGP, Firewall, DNS, Netflow, Whois, NVD, CIDR…
- Technical Challenges: Volume and Velocity of data; Temporal dependencies; Real-time response
- Users: Cyber Analysts
- Usage model: Iterative analysis of all patterns across all traffic to explore deviations in frequency of occurrence, derivative patterns of known threats and linking patterns through relationships in offline data
- Augmenting: Existing data appliances

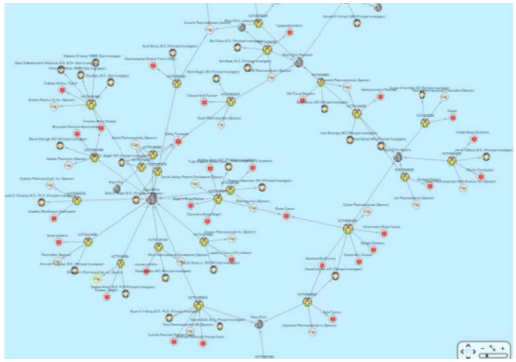


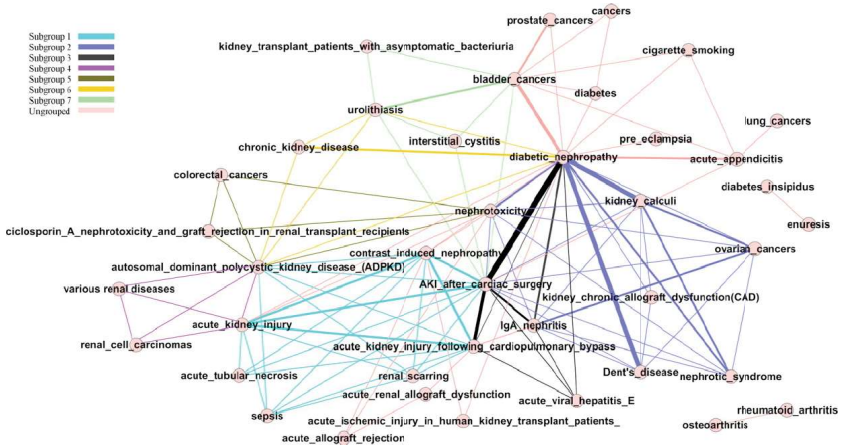


DTG	SIP	DIP	PROT	SPORT	DPORT	PKTS	BYTES	AVG BYTES per PKT	blacklist
2012-07-10T07:52:29	172.31.21.234	212.117.170.54	6	51200	80	6	558	93.00	Tor exit node
2012-07-10T08:37:59	172.31.21.234	212.117.170.54	6	51845	80	6	558	93.00	Tor exit node
2012-07-10T08:38:25	172.31.21.234	212.117.170.54	6	51846	80	8	944	118.00	Tor exit node
2012-07-10T10:59:25	172.31.21.29	212.117.170.54	6	52834	80	7	598	85.43	Tor exit node
2012-07-10T12:04:43	172.31.21.234	212.117.170.54	6	53341	80	6	558	93.00	Tor exit node
2012-07-10T14:04:19	172.31.21.29	212.117.170.54	6	54373	80	6	558	93.00	Tor exit node
2012-07-10T15:59:32	172.31.21.29	212.117.170.54	6	54830	80	6	558	93.00	Tor exit node
2012-07-10T16:45:57	172.31.20.53	65.164.25.47	6	47155	80	33	2347	71.12	SSH Brute Force
2012-07-10T16:45:58	172.31.20.53	65.164.25.47	6	47166	80	8	932	116.50	SSH Brute Force
2012-07-10T16:45:58	172.31.20.53	65.164.25.47	6	47165	80	8	933	116.63	SSH Brute Force

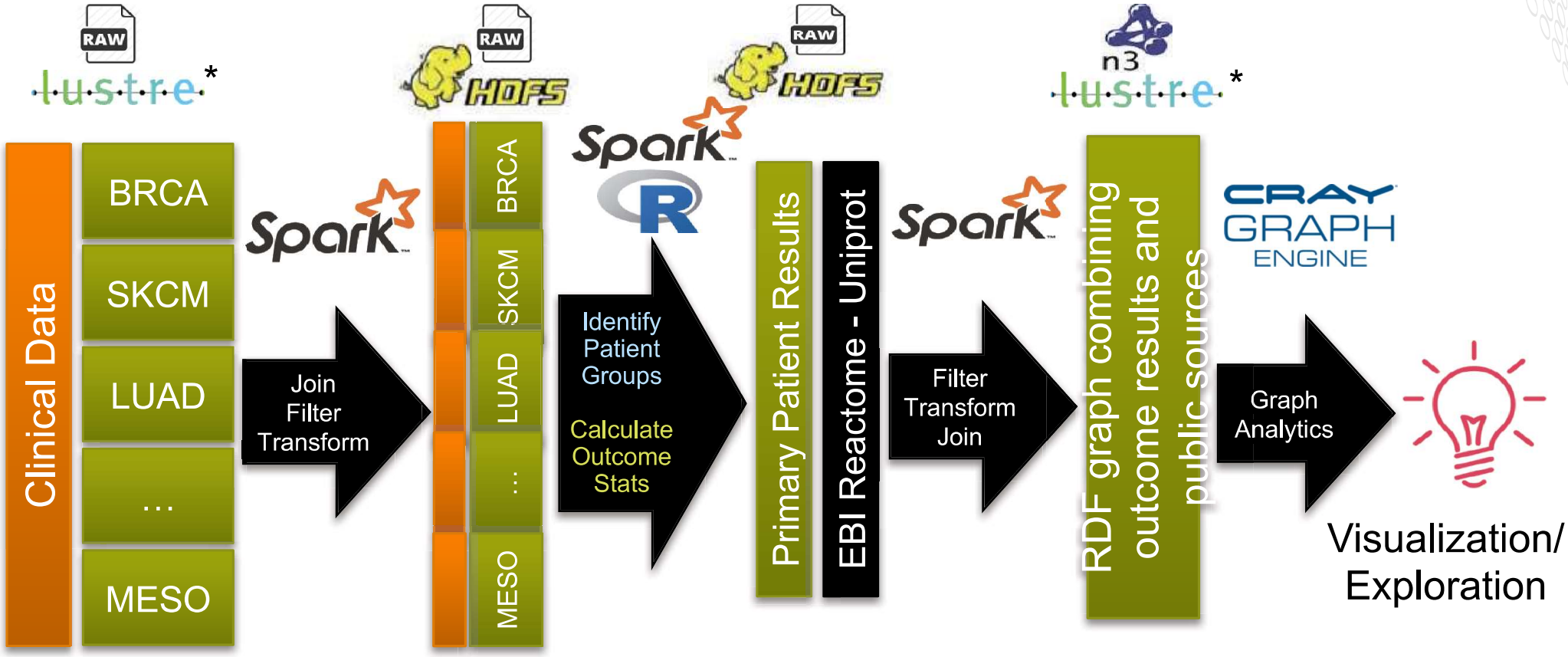
Life Sciences Use Case: Discovering new uses for existing drugs

- Goal: Re-purpose drugs for a different disease or cancer
- Data sets: Medline, PubMed, TCGA, Uniprot, Pfam, CRO, Clinical Trials…
- Technical Challenges: Volume and Velocity of data; Complex inter-relationships; Entity resolution; Probabilistic relationships
- Users: Life sciences researchers both in discovery and development
- Usage model: Identify the complex relationships between the disease state, patient biomarkers, literature and drug pathways and use the information for identifying new candidate disease targets or companion diagnostics
- Augmenting: Existing data warehouse, search engines





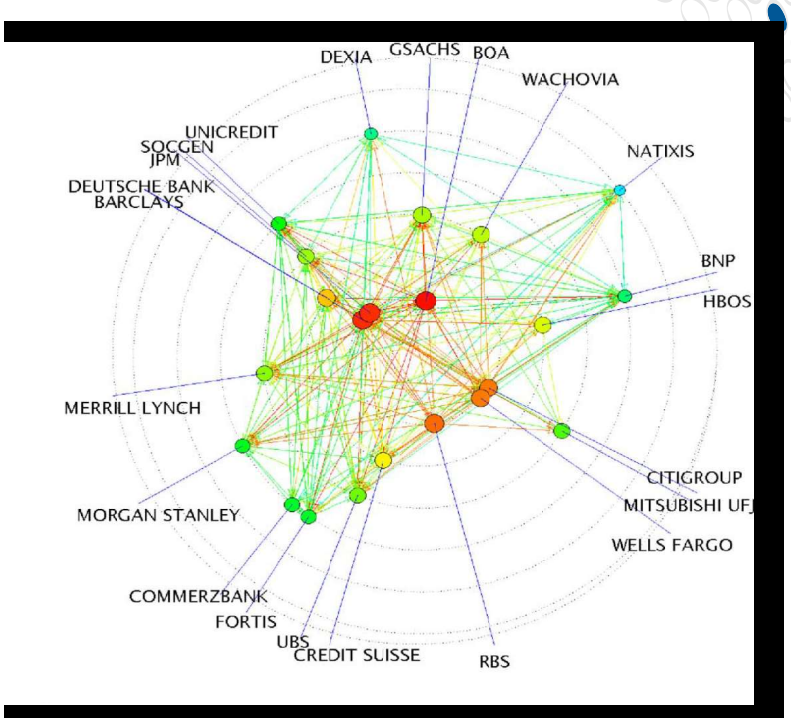
Life Sciences Architecture

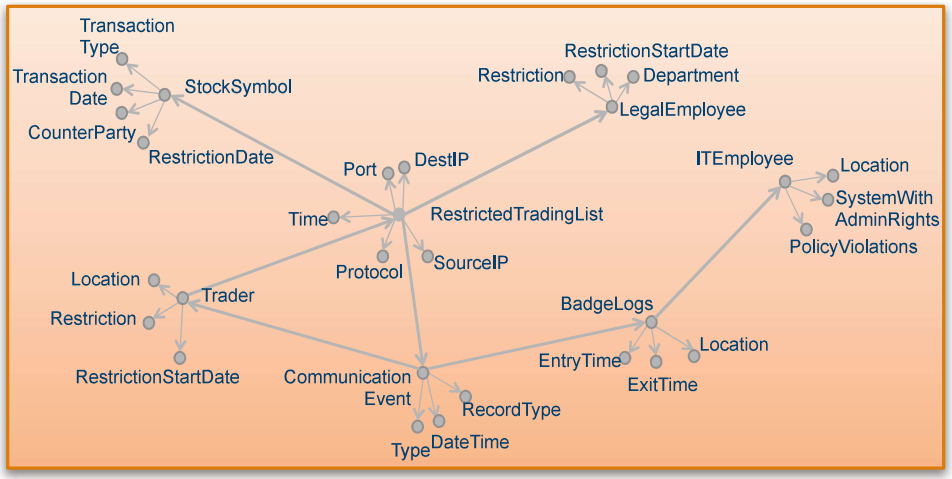


*Optional

Financial Services Use Case: Discovering new Risk/Compliance events

- Goal: Find detection patterns and improve efficiency of the investigation process by reducing false positives
- Data sets: Accounts, Customer Transactions, 3rd party data feeds, Detection and Case Management systems
- Technical Challenges: Rigid detection system schemas and rules; Constantly degrading performance as new data comes in; Hard to tune performance with new data; Long data on-boarding timeframes; Manual disposition of benign alerts
- Users: Investigators, Analysts
- Usage model: Tune detection system models via data discovery; Enhance, improve and augment the alert investigations process
- Augmenting: Existing detection systems





In a nutshell

- **Relational Databases**
  - Relational tables can hold a lot of information and give you rapid, sophisticated access to it
  - Examples: Oracle Exascale
- **Semantic Databases ("triple stores")**
  - Graphs are the best representation to use if you are at least as interested in the *pattern of relationships between data items* as you are in the data itself
  - Examples: Cray Graph Engine, Apache Jena, Stardog, Blazegraph
- **Property Graphs**
  - Compromise between semantic databases and relational databases
  - Examples: Neo4j, Titan, (Spark GraphX and Graphframes)
- **https://en.wikipedia.org/wiki/Graph_database**