Network Analysis with R and Neo4j

Geoffrey Hannigan, PhD
University of Michigan Medical School
@iprophage

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Goals

- It's a lot of ground to cover in meeting.
- Provide an introduction to network analysis.
- Cover examples of how to use networks.
- Equip everyone to continue learning about this after the meeting.

Outline

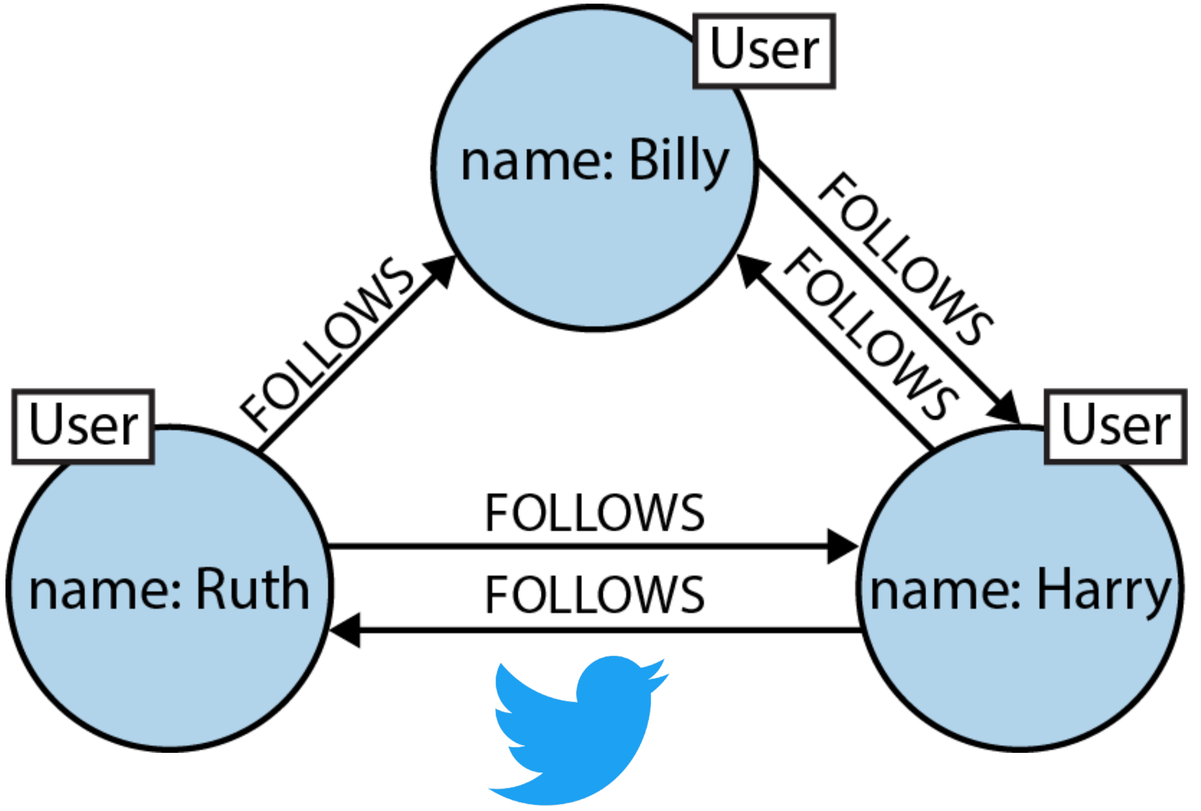
- What are graphs and why use them?
- How do I build a graph database?
- How do I analyze a graph database?
- Applying graphs to biological problems.
- Wrapping it up.

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

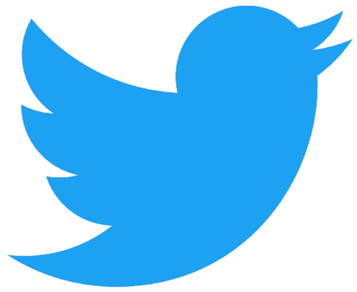
What are graphs and why use them?

What is a Graph?

- Graphs are collections of data nodes and the relationships that connect them.
- Data points and their properties are represented as node.
- Structure focused on relationships between nodes. Relationships also have properties.



Understanding Data Through Tables

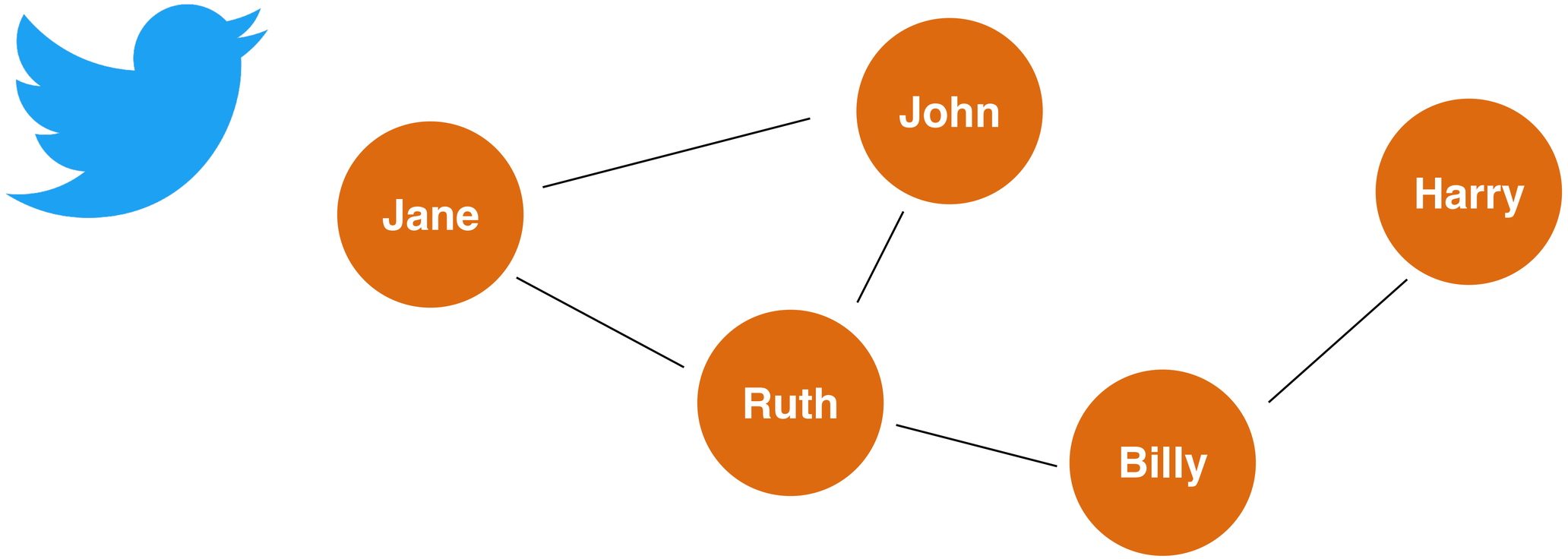


Name	Age	Tweets Last Week	Followers
Jane	23	102	2
John	57	302	1
Ruth	98	43	5
Harry	74	6	3
Billy	62	72	3

What is the average age of the population?
How many users have been active in the past week?
What is the average number of followers per user?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

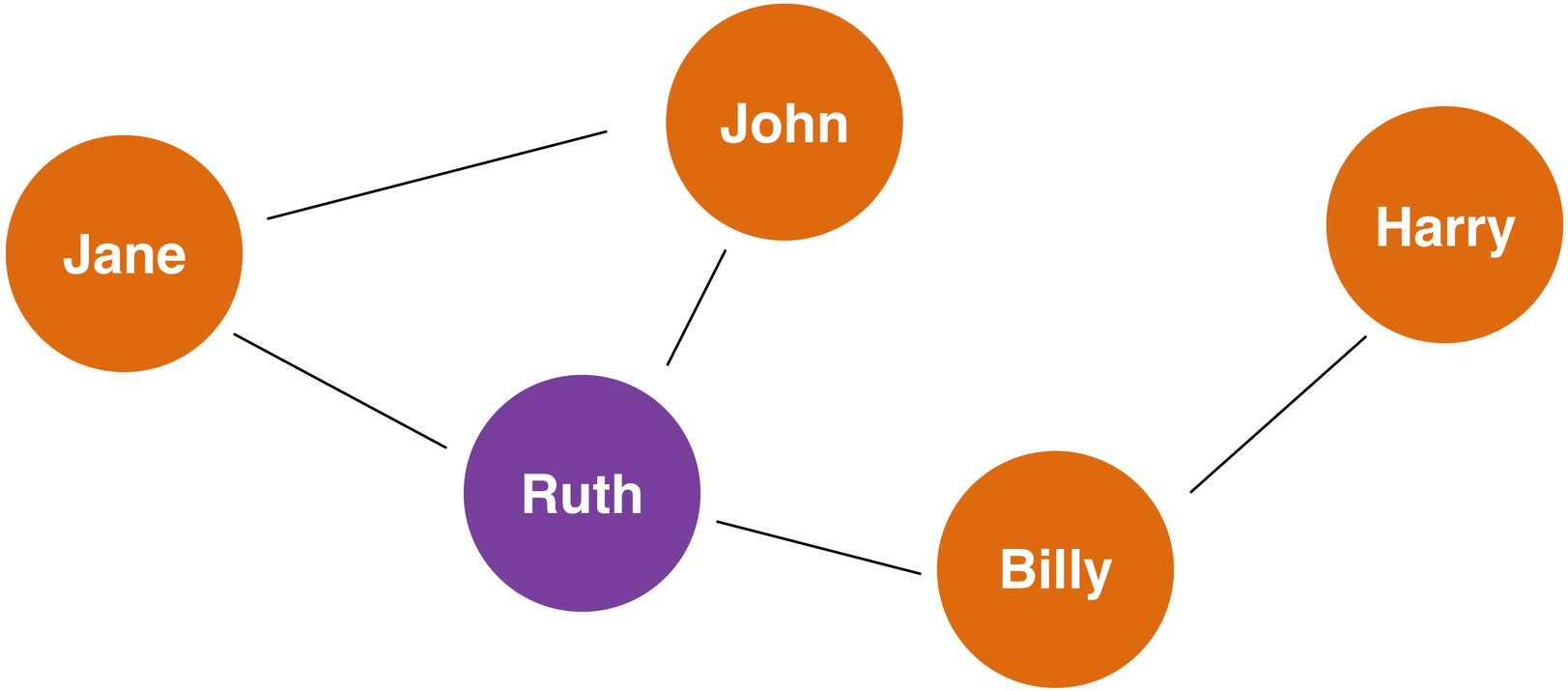
Graphs Focus on Data **Relationships**



Who is the most influential user?
Who should we suggest Harry follows?
Who is friends with Ruth?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

We Can Evaluate Influence by Relationship Abundance



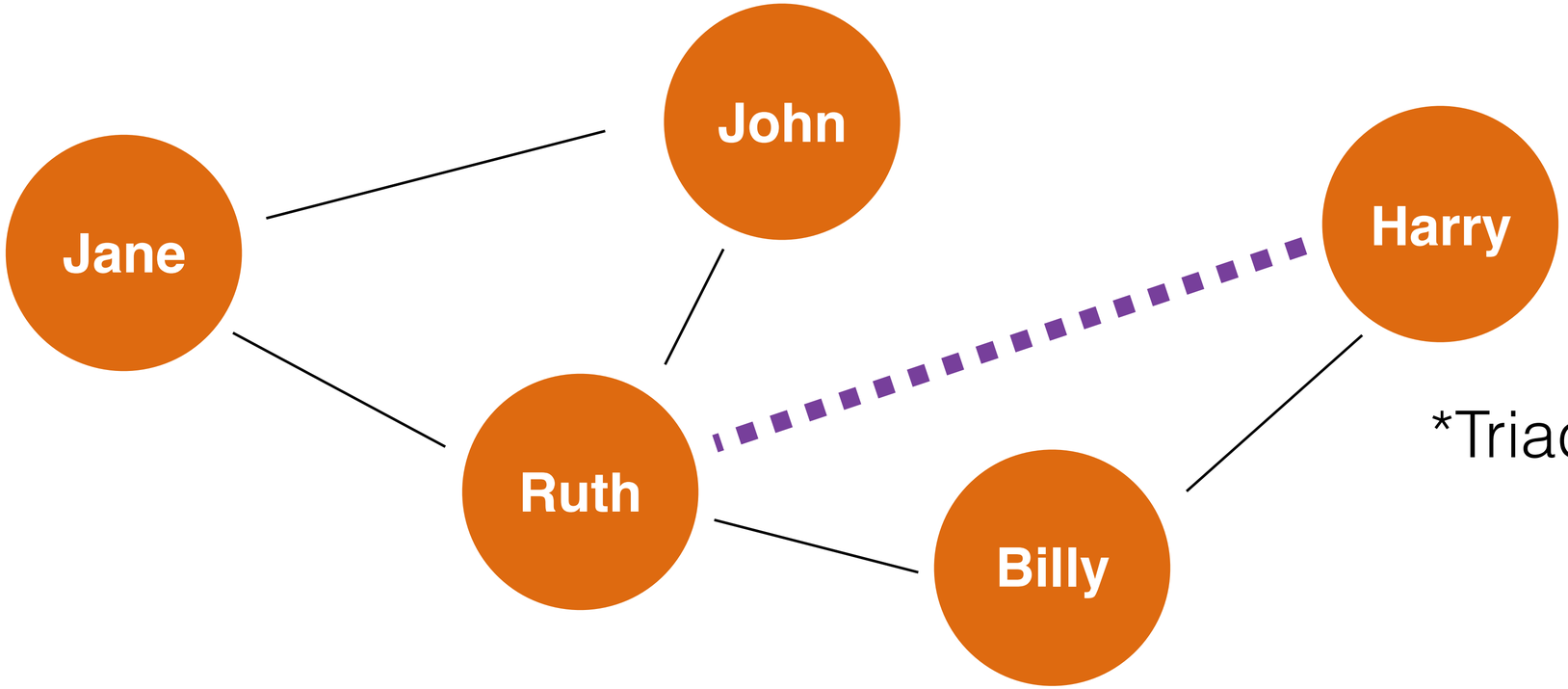
Who is the most influential user?

Who should we suggest Harry follows?

Who is friends with Ruth?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

We Can Predict Unrecorded Relationships



*Triadic Closure

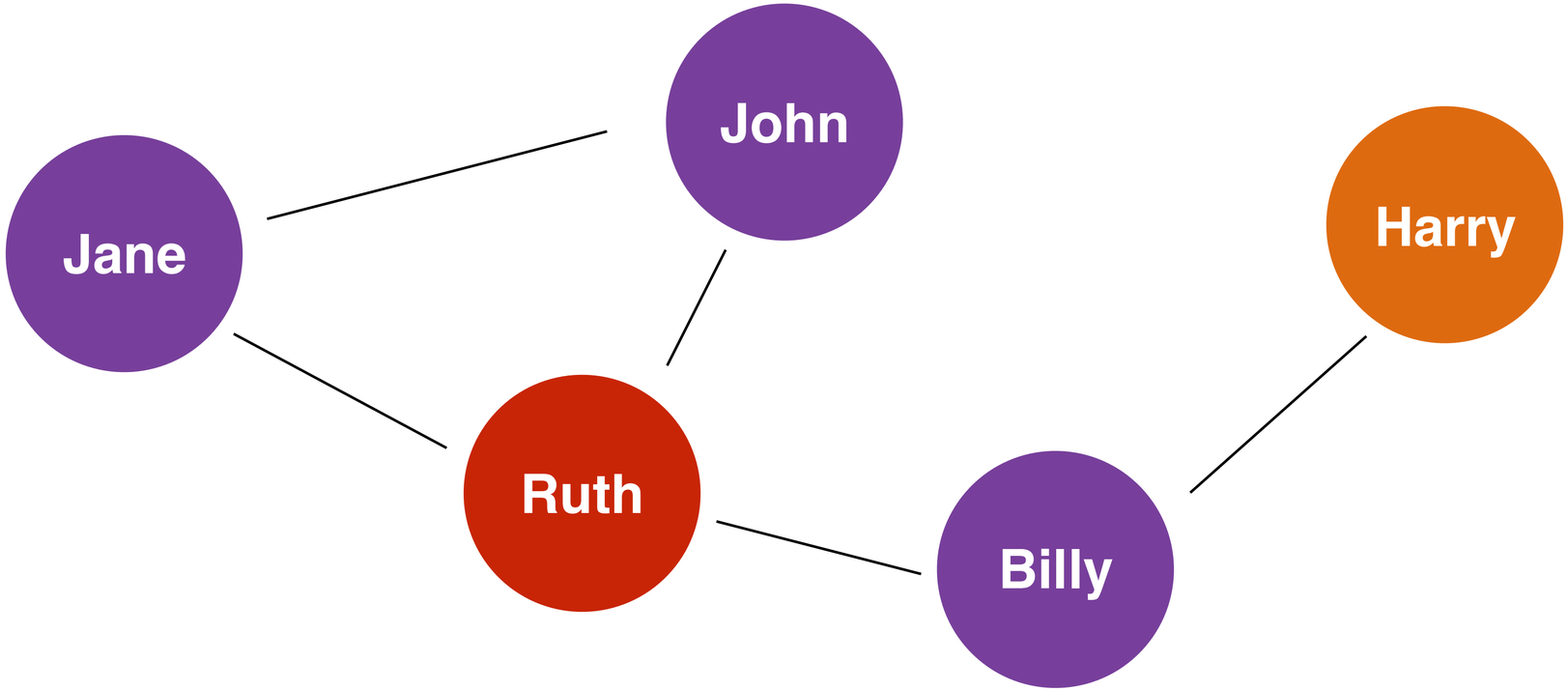
Who is the most influential user?

Who should we suggest Harry follows?

Who is friends with Ruth?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

We Can Quickly Retrieve Relationships



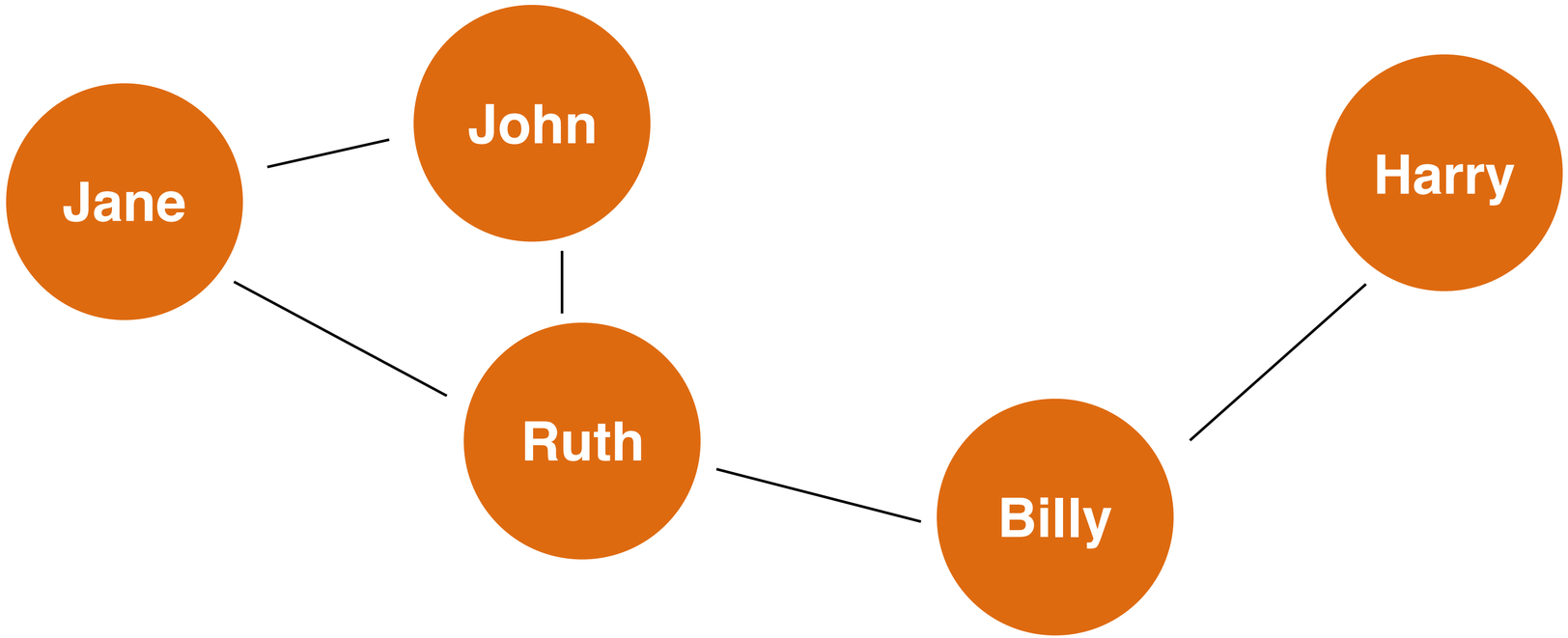
Who is the most influential user?

Who should we suggest Harry follows?

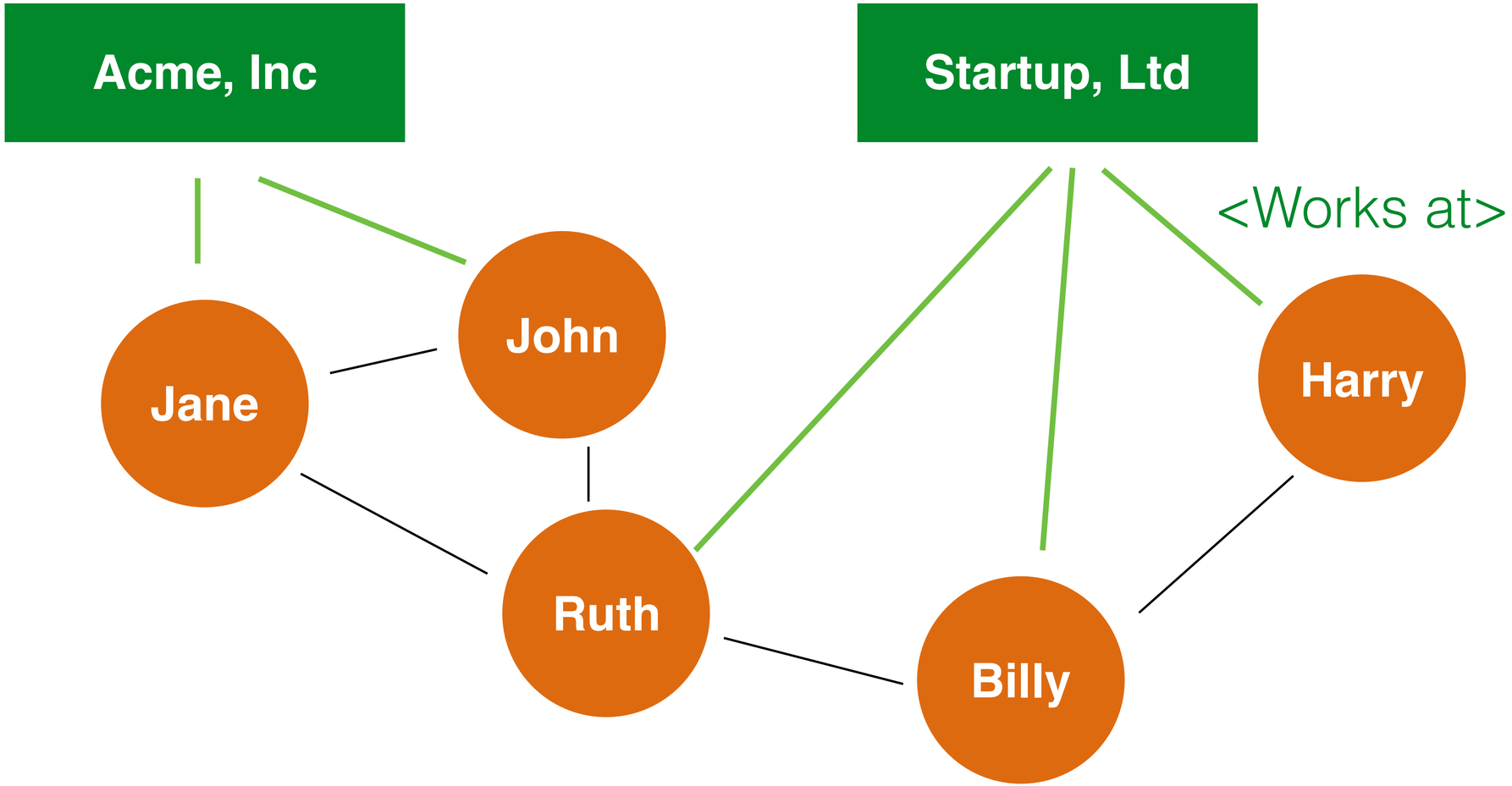
Who is friends with Ruth?

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

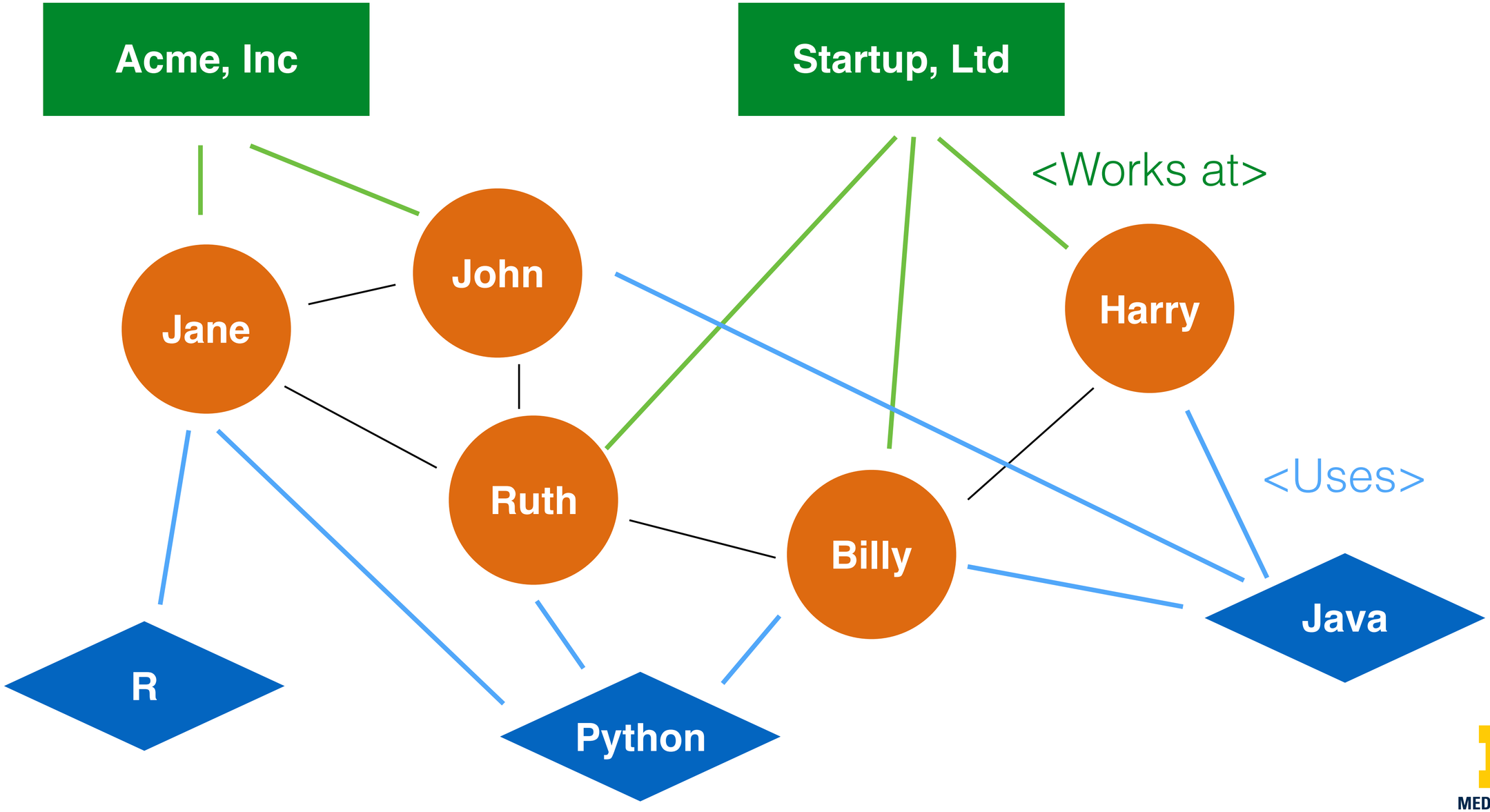
Graphs Can Manage Complex Datasets



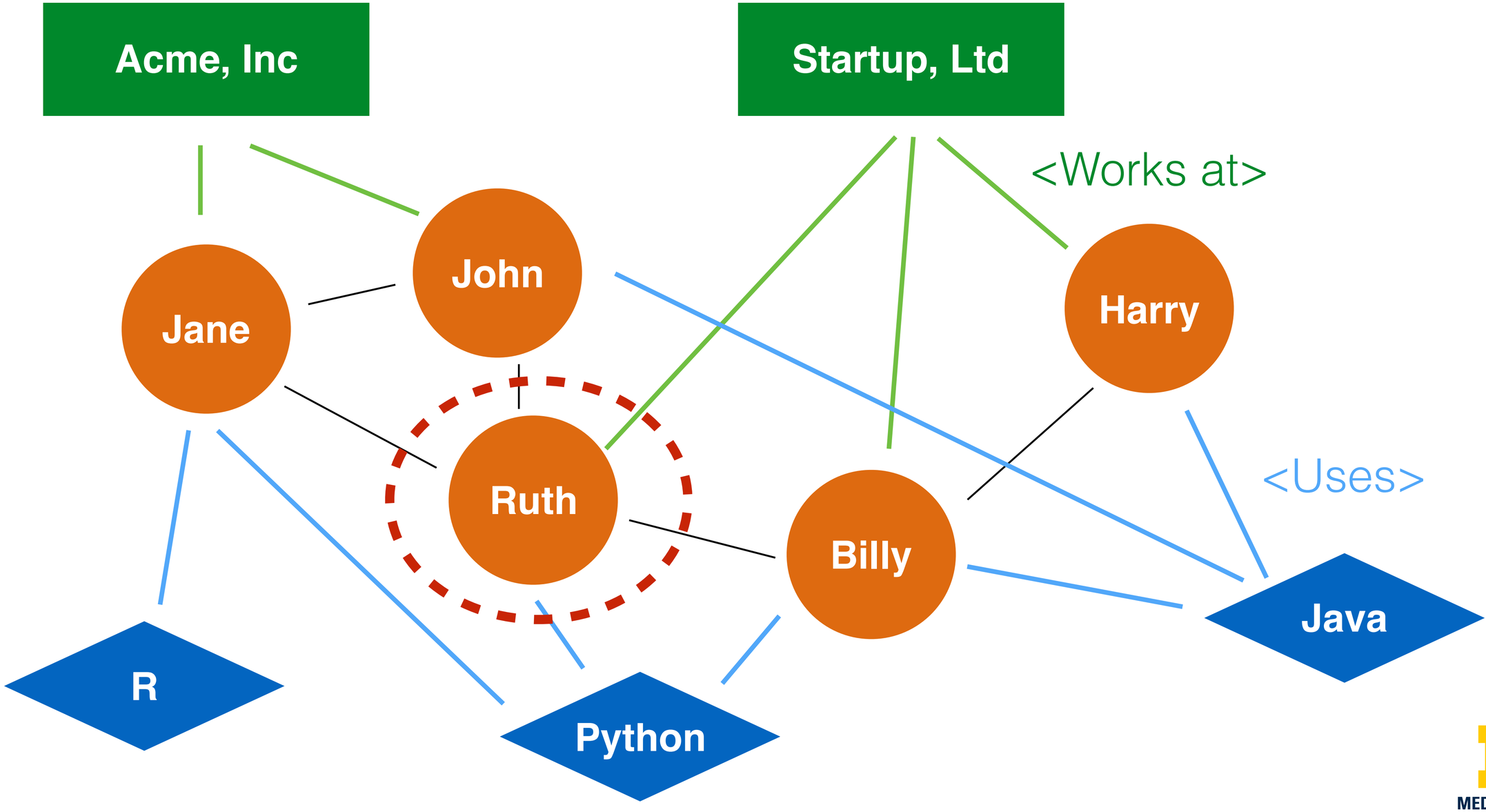
Add Employment Information



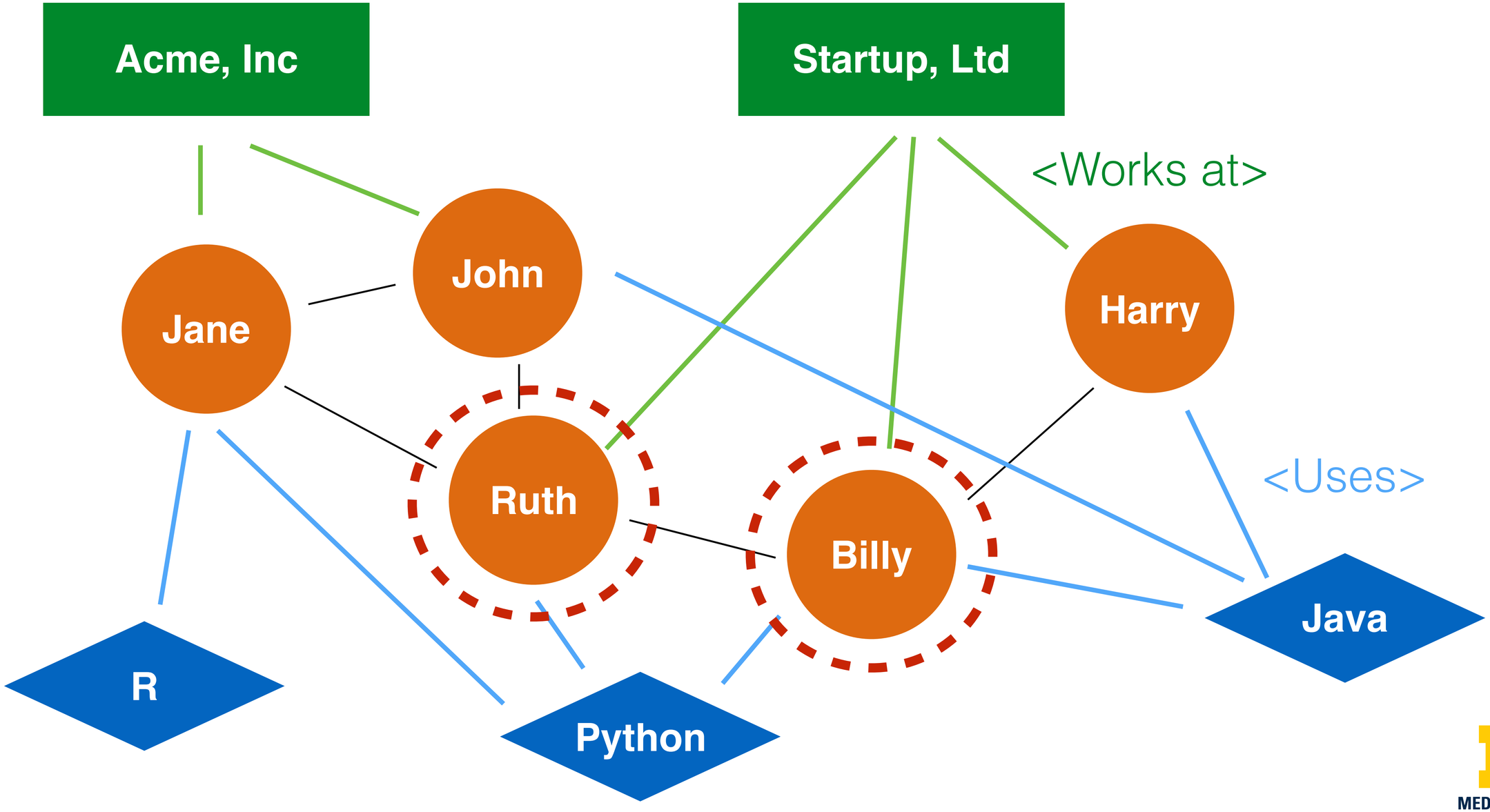
Add Programming Skills



What Startup Employee Might Know About Acme?



Who Shares Jane's Interest in Python?



MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

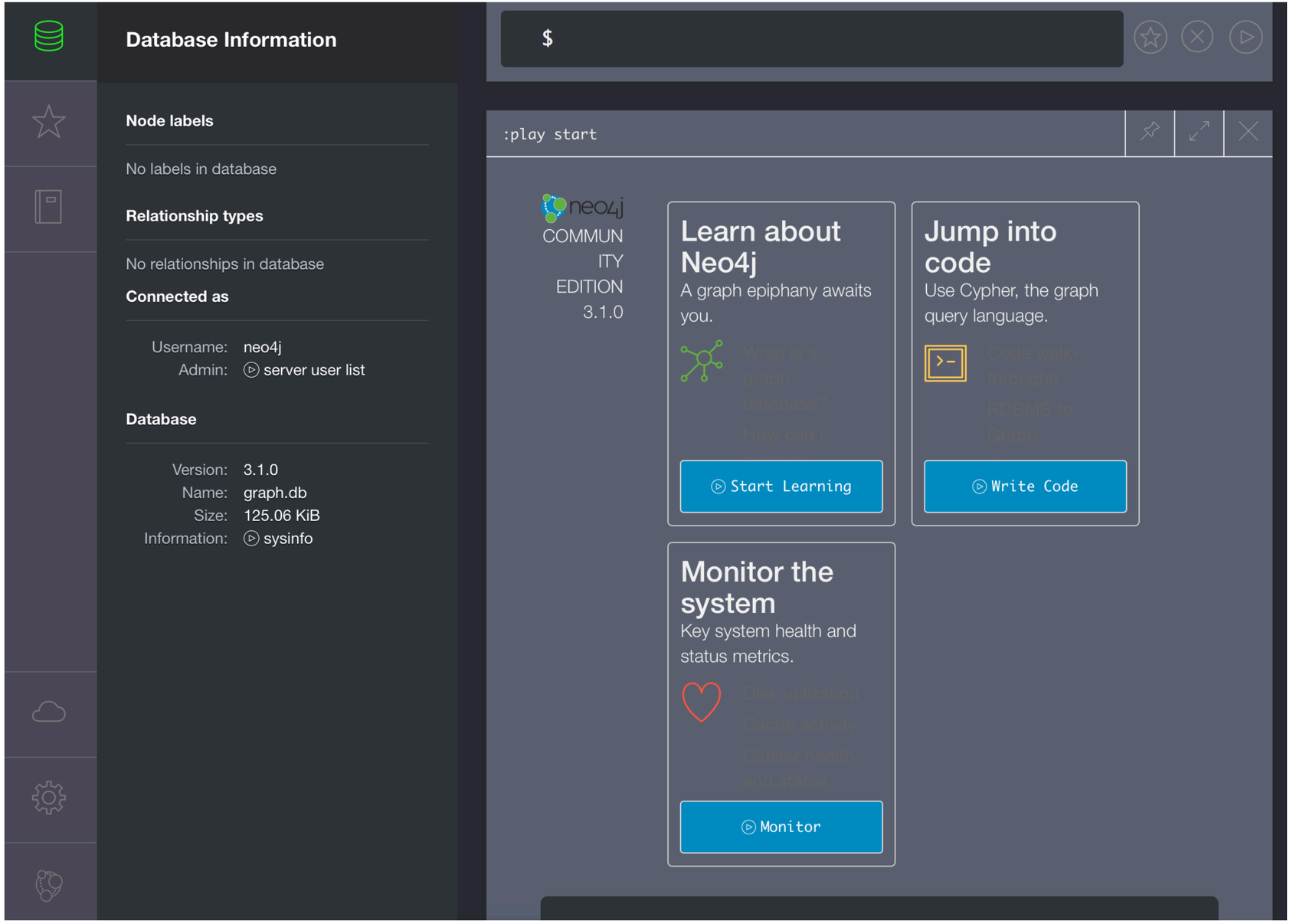
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

How do I build a graph database?

Use Neo4j to Build Graph Databases

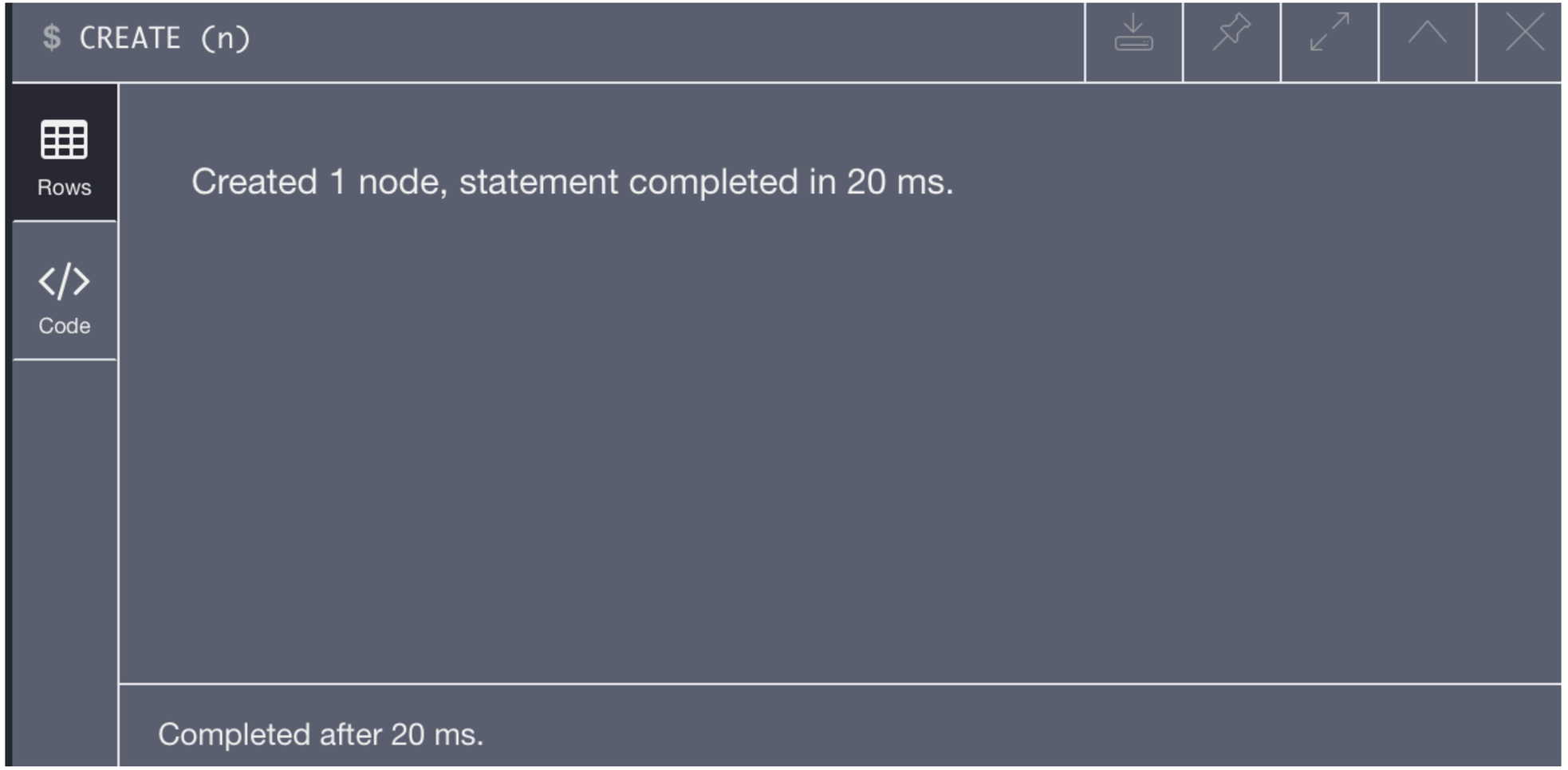
- Neo4j is robust, free (community edition) software that powers graph databasing.
- Runs similar to MySQL and other similar relational databases.
- Uses the query language **Cypher**.
- Free introductory O'Rielly book available for more info on their website.

Neo4j Provides a Helpful GUI

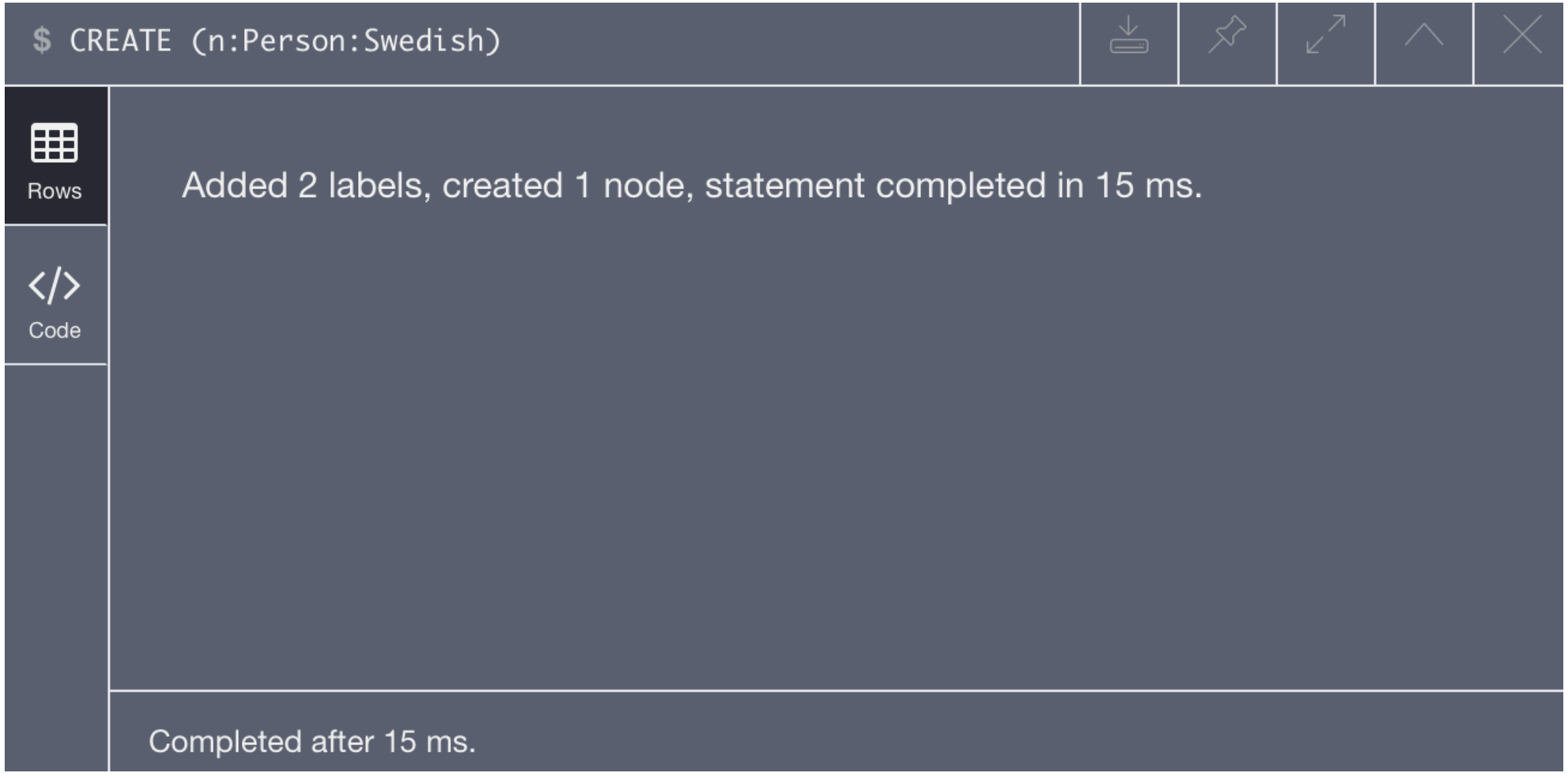


Create a Node Using Cypher

```
CREATE (n)
```



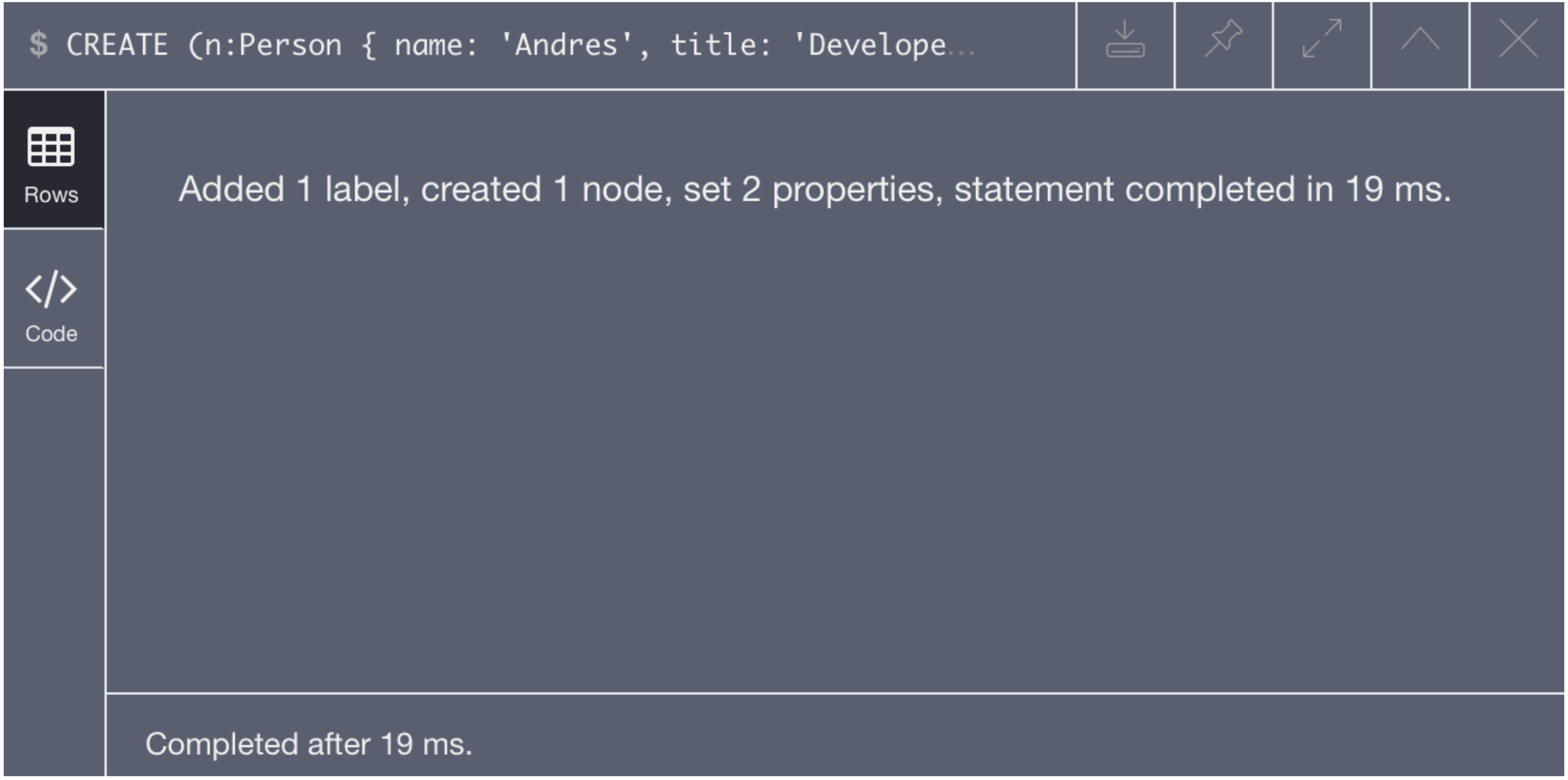
Create Nodes With Labels



Labels represent the role the node plays in the graph (allows grouping).
This node represents a **person** who is **Swedish**.

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

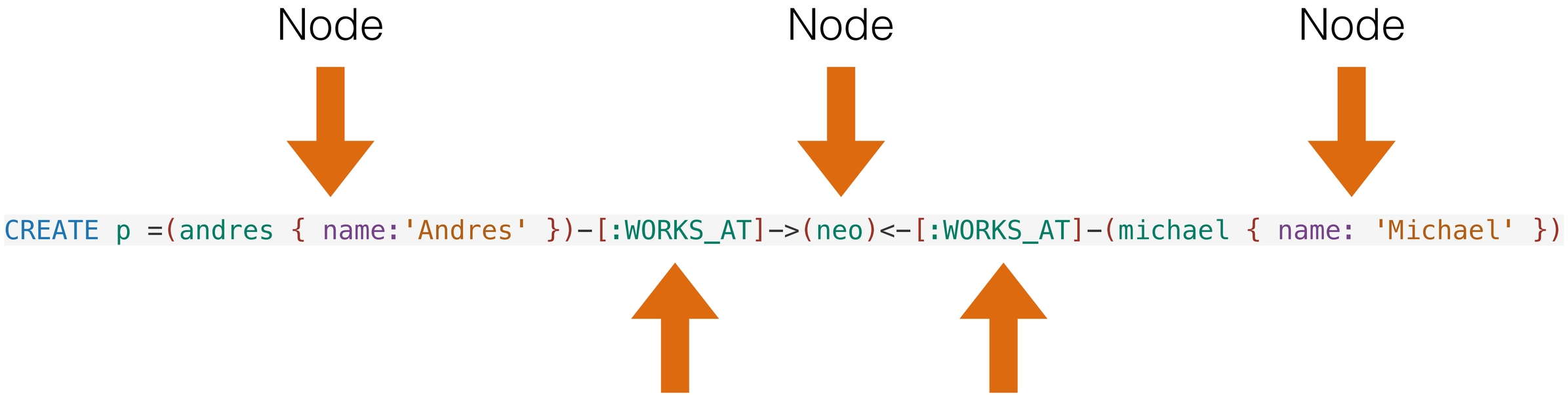
We Can Create Nodes With Labels & Properties



Properties are more unique attributes stored in key-value pairs.
This node represents a person **named Andres** whose **title** is **developer**.

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

We Can Connect Nodes with Relationships

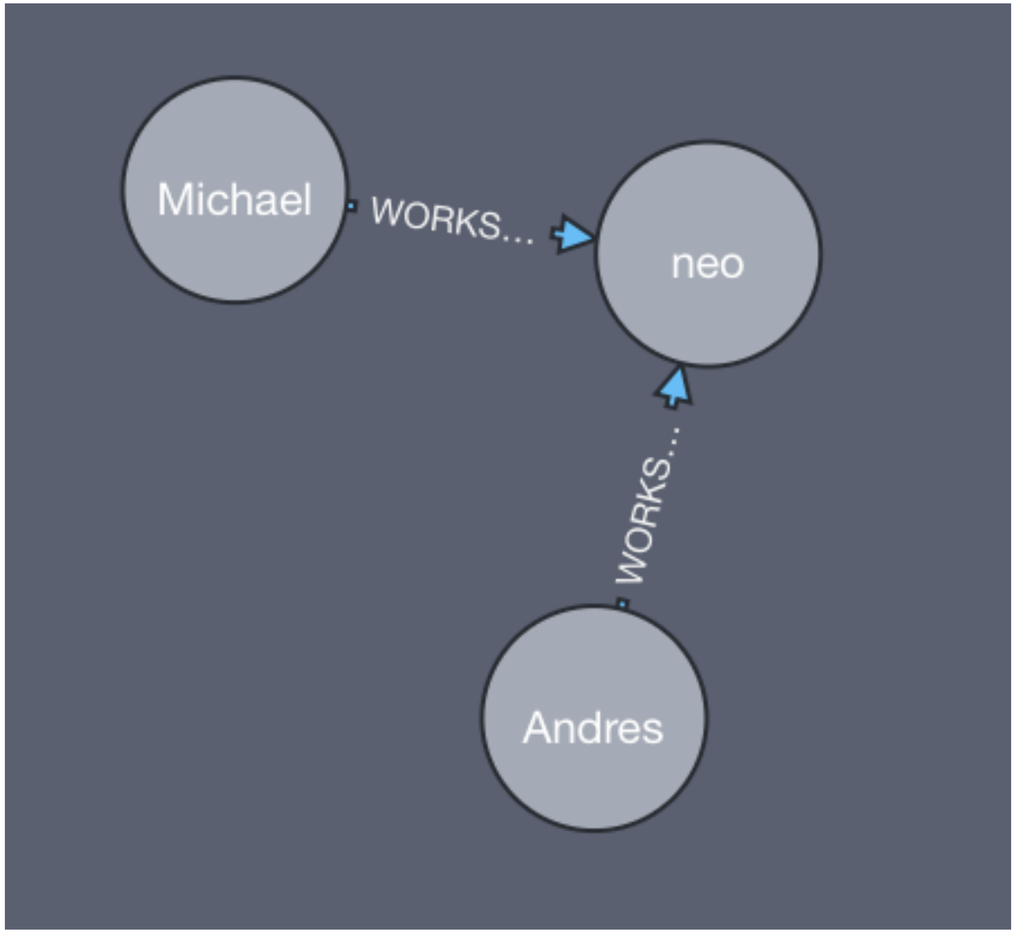


Andres and Michael both work at neo.

We Can Connect Nodes with Relationships

```
CREATE p =(n:person { name:'Andres' })-[:WORKS_AT]->(a:company { name: 'neo' })<-[:WORKS_AT]-(m:michael { name: 'Michael' })

RETURN p
```



Larger Datasets Can Be Imported From CSV

name	born	type	roles	title	released	tagline
Emil Eifrem	1978	ACTED_IN	Emil	The Matrix	1999	Welcome to the Real World
Joel Silver	1952	PRODUCED		The Matrix	1999	Welcome to the Real World
Lana Wachowski	1965	DIRECTED		The Matrix	1999	Welcome to the Real World
Andy Wachowski	1967	DIRECTED		The Matrix	1999	Welcome to the Real World

An example dataset containing movie information.

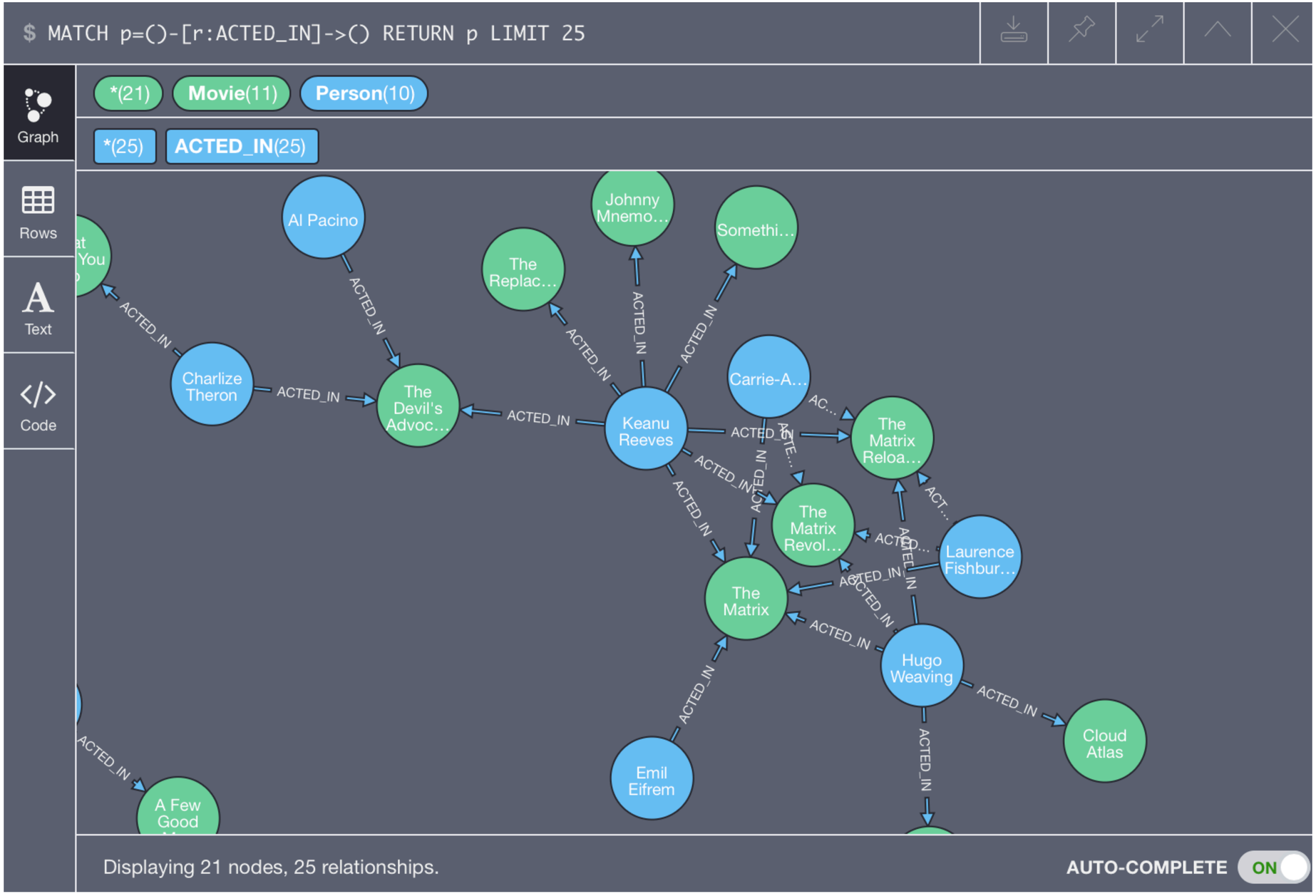
Larger Datasets Can Be Imported From CSV

NODE		RELATIONSHIP		NODE		
name	born	type	roles	title	released	tagline
Emil Eifrem	1978	ACTED_IN	Emil	The Matrix	1999	Welcome to the Real World
Joel Silver	1952	PRODUCED		The Matrix	1999	Welcome to the Real World
Lana Wachowski	1965	DIRECTED		The Matrix	1999	Welcome to the Real World
Andy Wachowski	1967	DIRECTED		The Matrix	1999	Welcome to the Real World

```
LOAD CSV WITH HEADERS FROM "https://dl.dropboxusercontent.com/u/14493611/movies_setup.csv"
AS row
MERGE (m:Movie {title:row.title}) ON CREATE SET m.released = toInt(row.released),
m.tagline = row.tagline
MERGE (p:Person  {name:row.name}) ON CREATE SET p.born     = toInt(row.born)
WITH   m,p,row WHERE row.type = "ACTED_IN"
MERGE (p)-[r:ACTED_IN]->(m) ON CREATE SET r.roles = split(row.roles,";")[0..-1]
```

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Quick Look at What People Acted In



MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

How do I analyze a graph database?

Getting Started with Graph Analysis

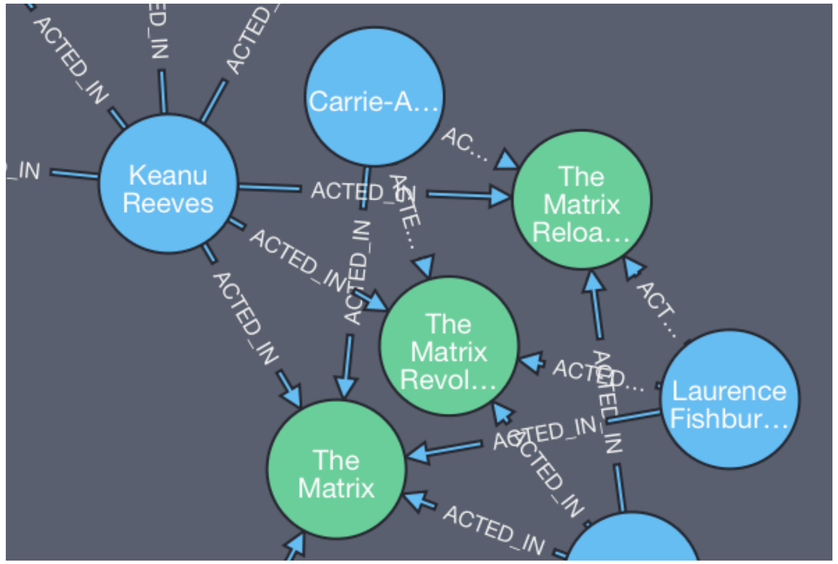
- **Neo4j/Cypher** analysis only gets you so far.
- **RNeo4j** provides a driver to link R to the database.
- **igraph** provides an extensive analytical toolkit for graph-based analyses.
- These three tools work together to provide almost everything you need to analyze graphs in R!

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Getting Started with Graph Analysis

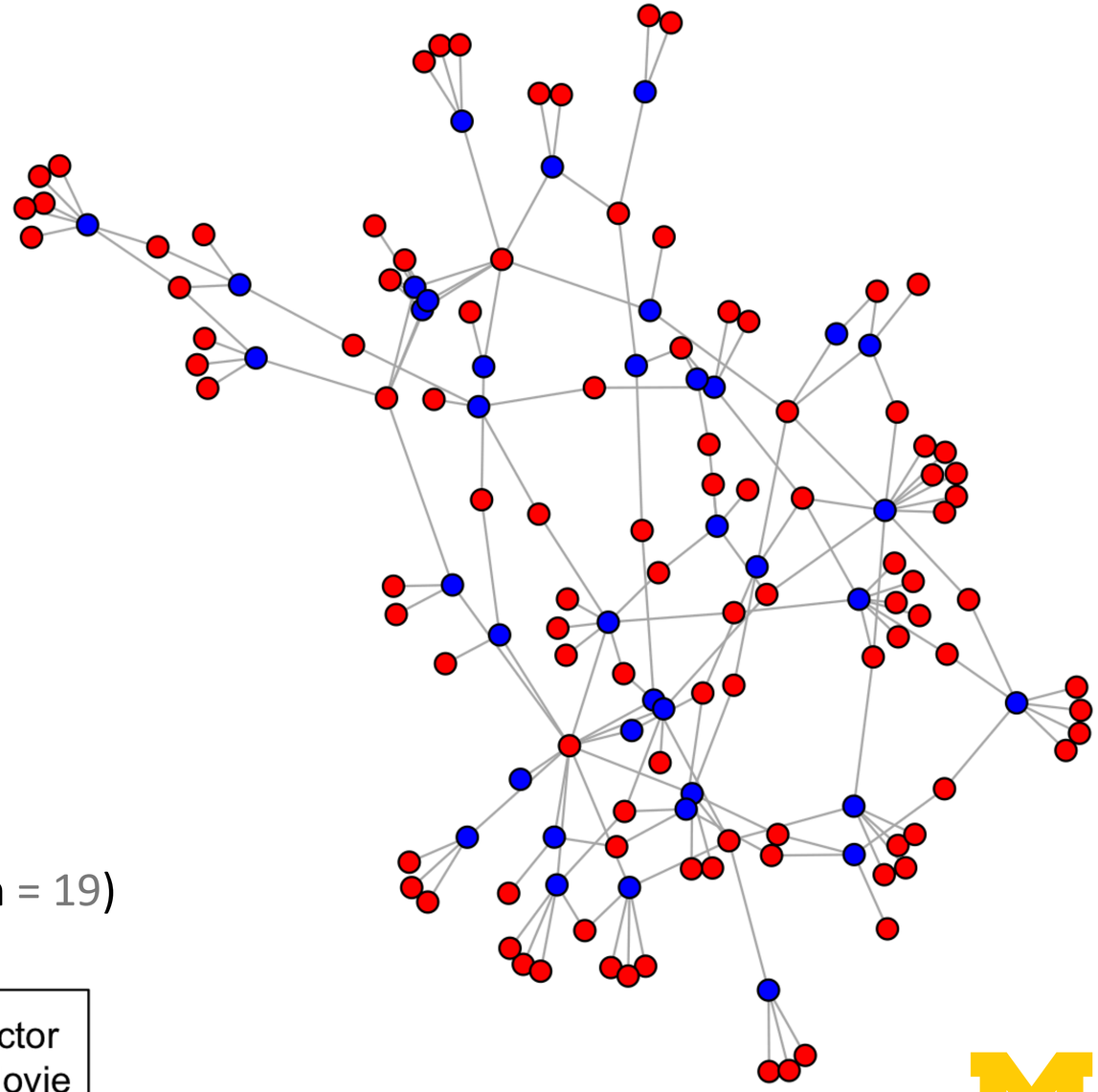
```
# Install the RNeo4j library
library(RNeo4j)
# Set the connection to the running neo4j database
graph <- startGraph("http://localhost:7474/db/data/", "neo4j", "root")
# Write out the desired cypher query
query <- "
MATCH (a)-[r:ACTED_IN]->(b)
RETURN a.name AS from, b.title AS to;
" # Run the query against the database
edges <- cypher(graph, query)
```

```
> edges
                    from                    to
1           Keanu Reeves          The Matrix
2      Laurence Fishburne          The Matrix
3        Carrie-Anne Moss          The Matrix
4           Hugo Weaving          The Matrix
```



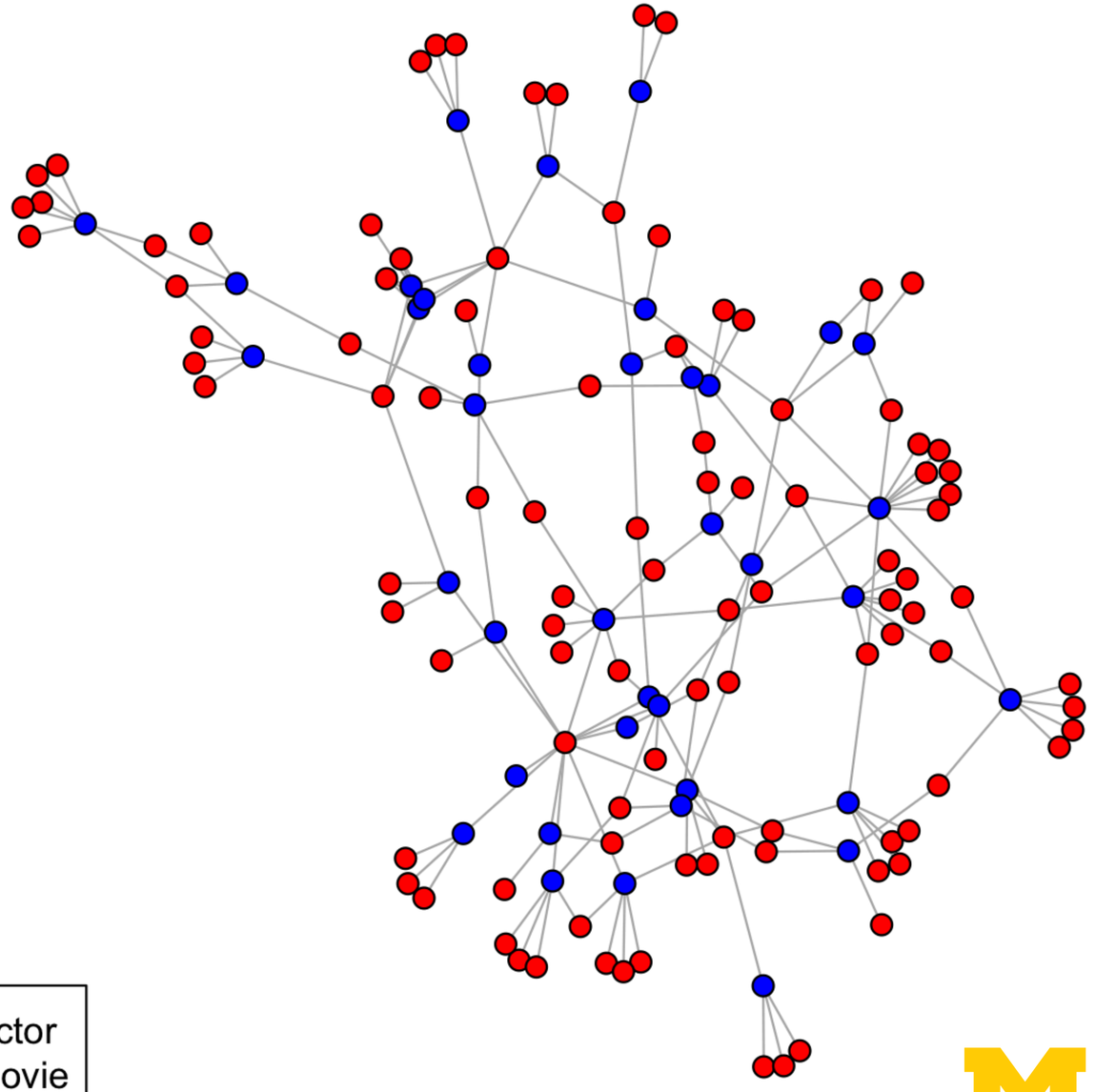
We Can Easily Plot Our Graph

```
# Create a graph object from the edge list
library(igraph)
# Get the node IDs
nodes <- data.frame(id=unique(c(edges$from, edges$to)))
nodes$label <- nodes$id
# Make the graph
ig <- graph.data.frame(edges, directed=FALSE)
V(ig)$color <- ifelse(nodes$label %in% edges[,1], "red", “blue")
V(ig)$label <- NA
V(ig)$size <- 4
plot(ig)
legend('bottomleft', c("Actor", "Movie"), col = c("red", "blue"), pch = 19)
```



Which Actor Is Most Influential On Other Actors?

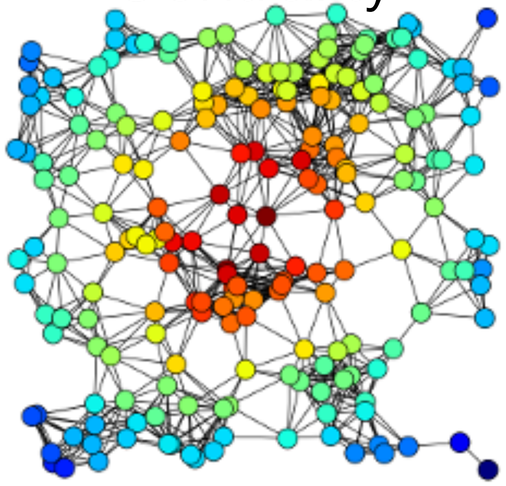
- Which actor(s) had the most opportunities to influence other actors while working on movies?
- Do we measure this as the number of movies they acted in?
- How do we account for the number of actors associated with those movies?
- This is a question of **centrality**.



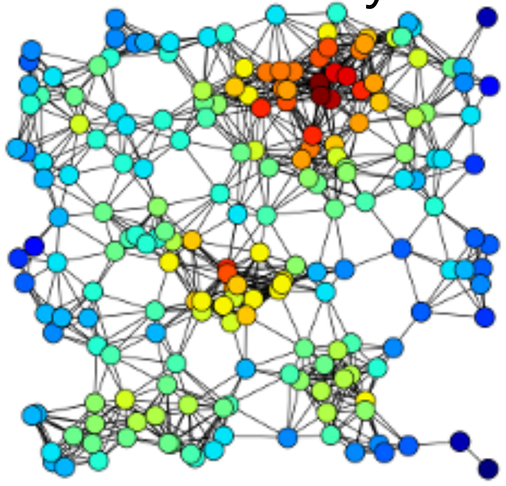
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Many Ways to Measure Centrality

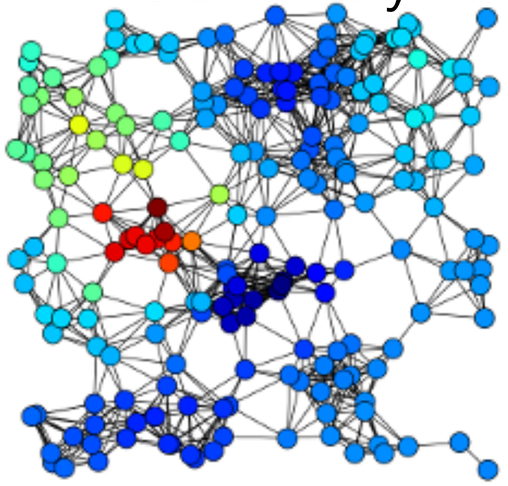
Closeness Centrality



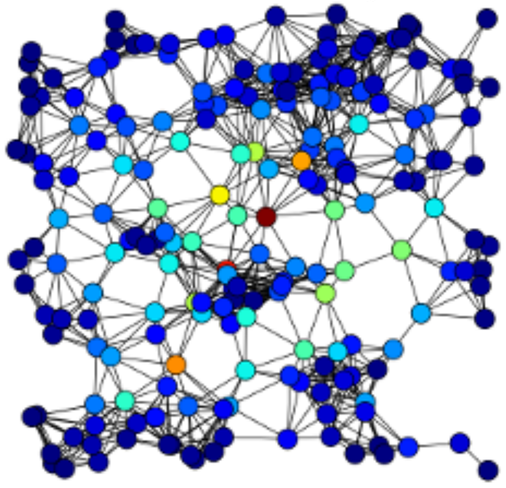
Degree Centrality



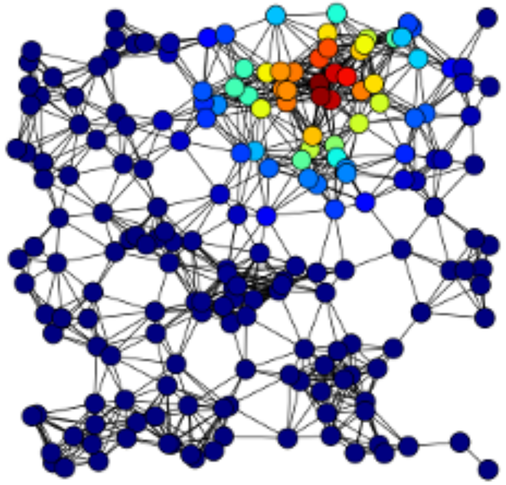
Katz Centrality



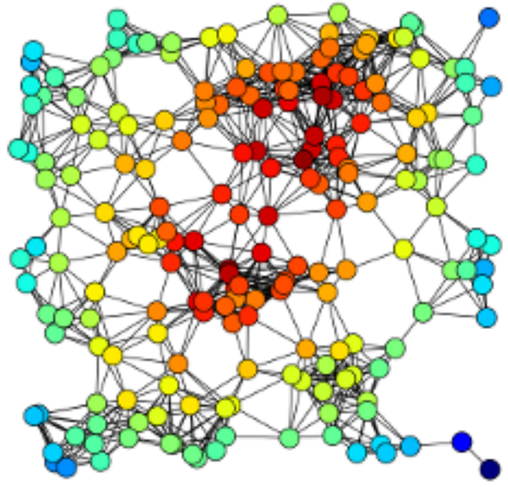
Betweenness Centrality



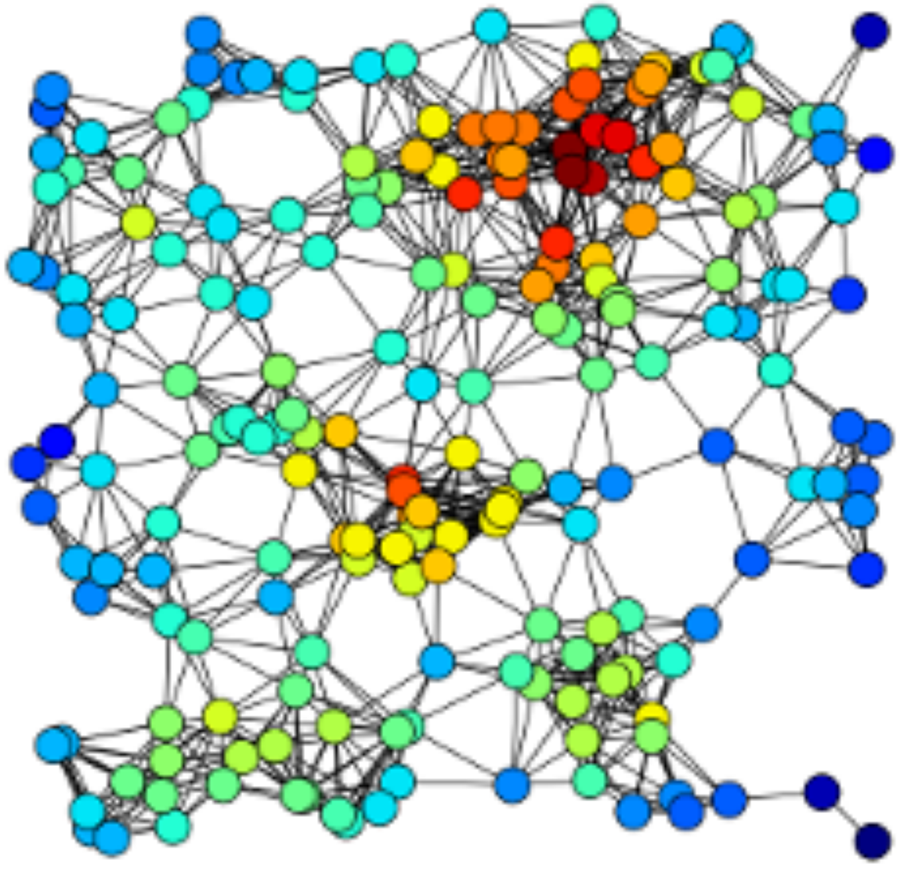
Eigenvector Centrality



Harmonic Centrality



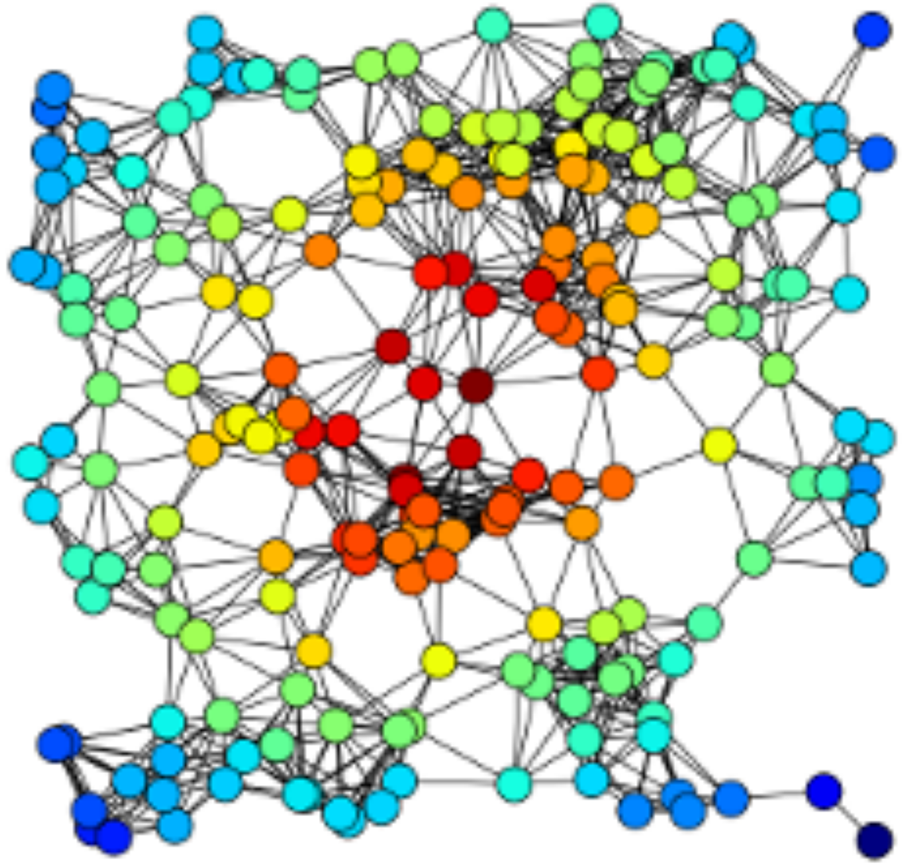
Degree Centrality



Number of relationships with each node.

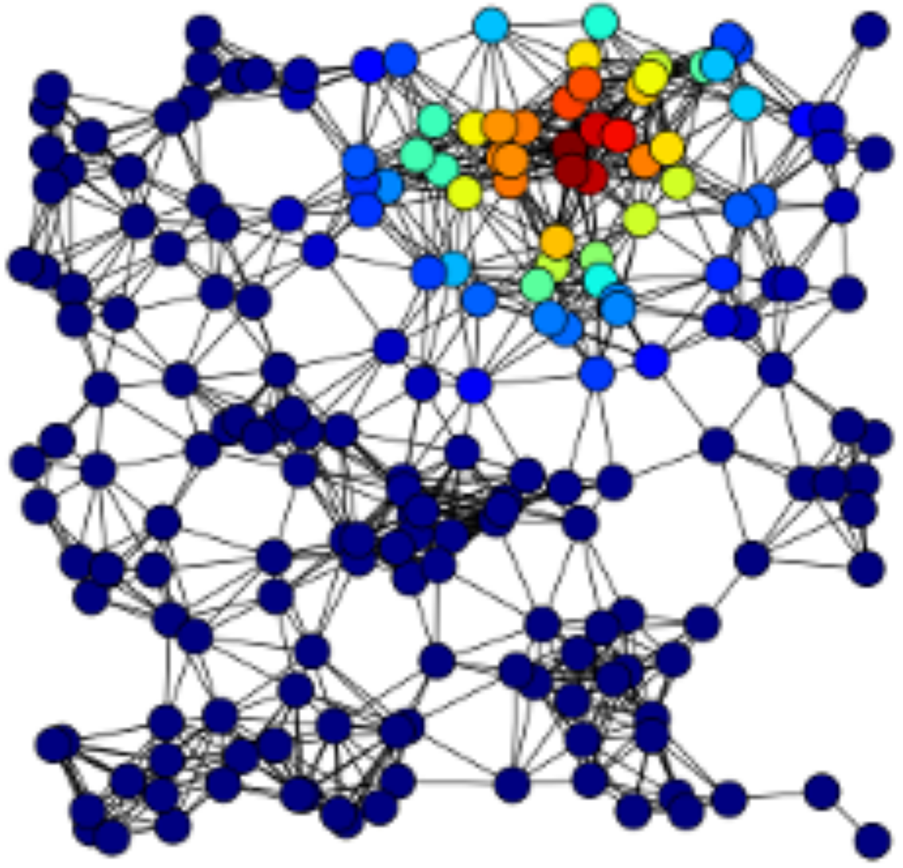
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Closeness Centrality



Average distance of node to all other nodes.

Eigenvector Centrality

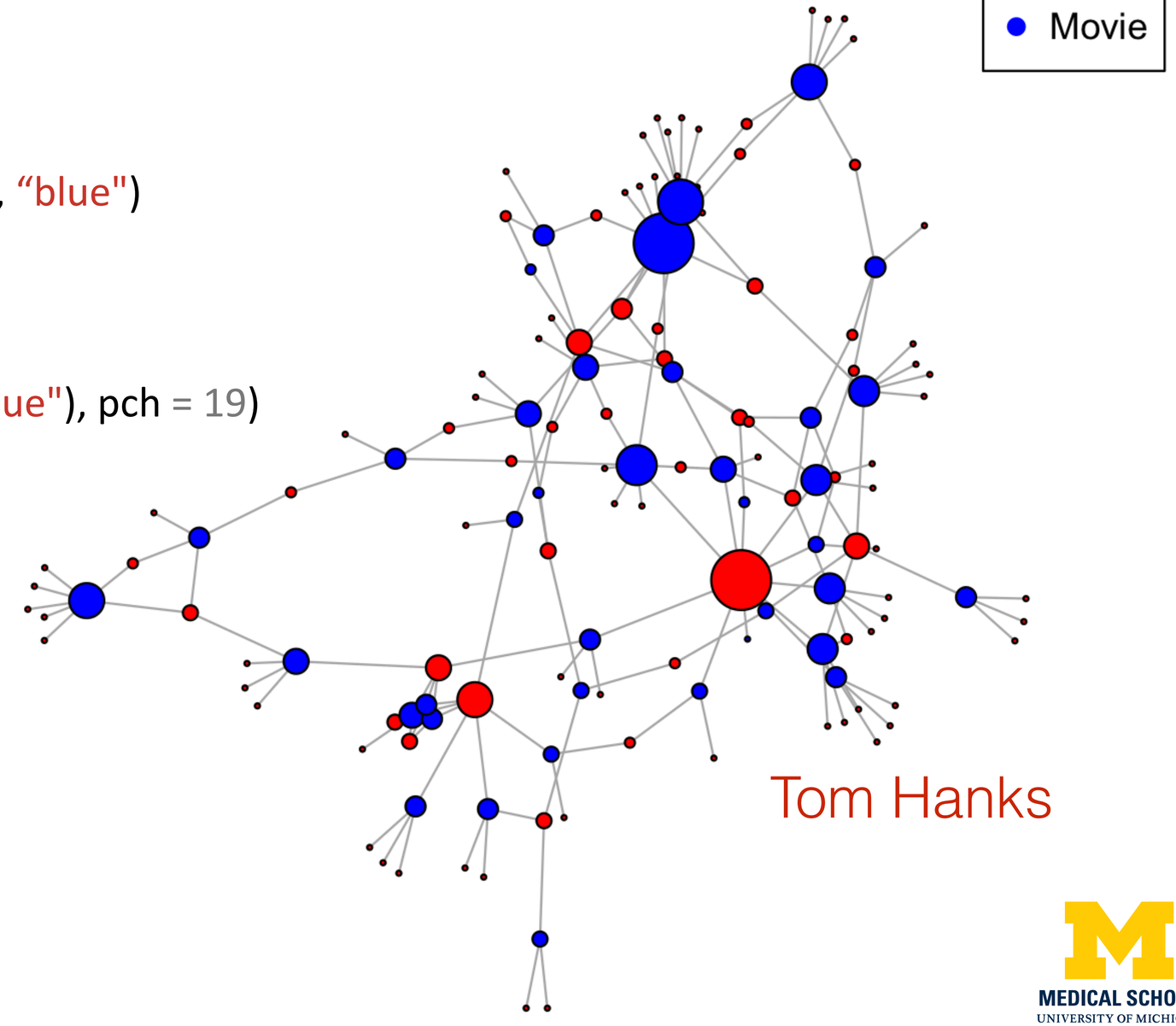


Function of number of relationships, and the centrality of connected nodes.

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Most Influential Actors (Degree Centrality)

```
# Get degree centrality
ig <- graph.data.frame(edges, directed=FALSE)
dig <- as.data.frame(degree(ig))
V(ig)$color <- ifelse(nodes$label %in% edges[,1], "red", “blue")
V(ig)$label <- NA
V(ig)$size <- dig[,1]
plot(ig)
legend('topright', c("Actor", "Movie"), col = c("red", "blue"), pch = 19)
```

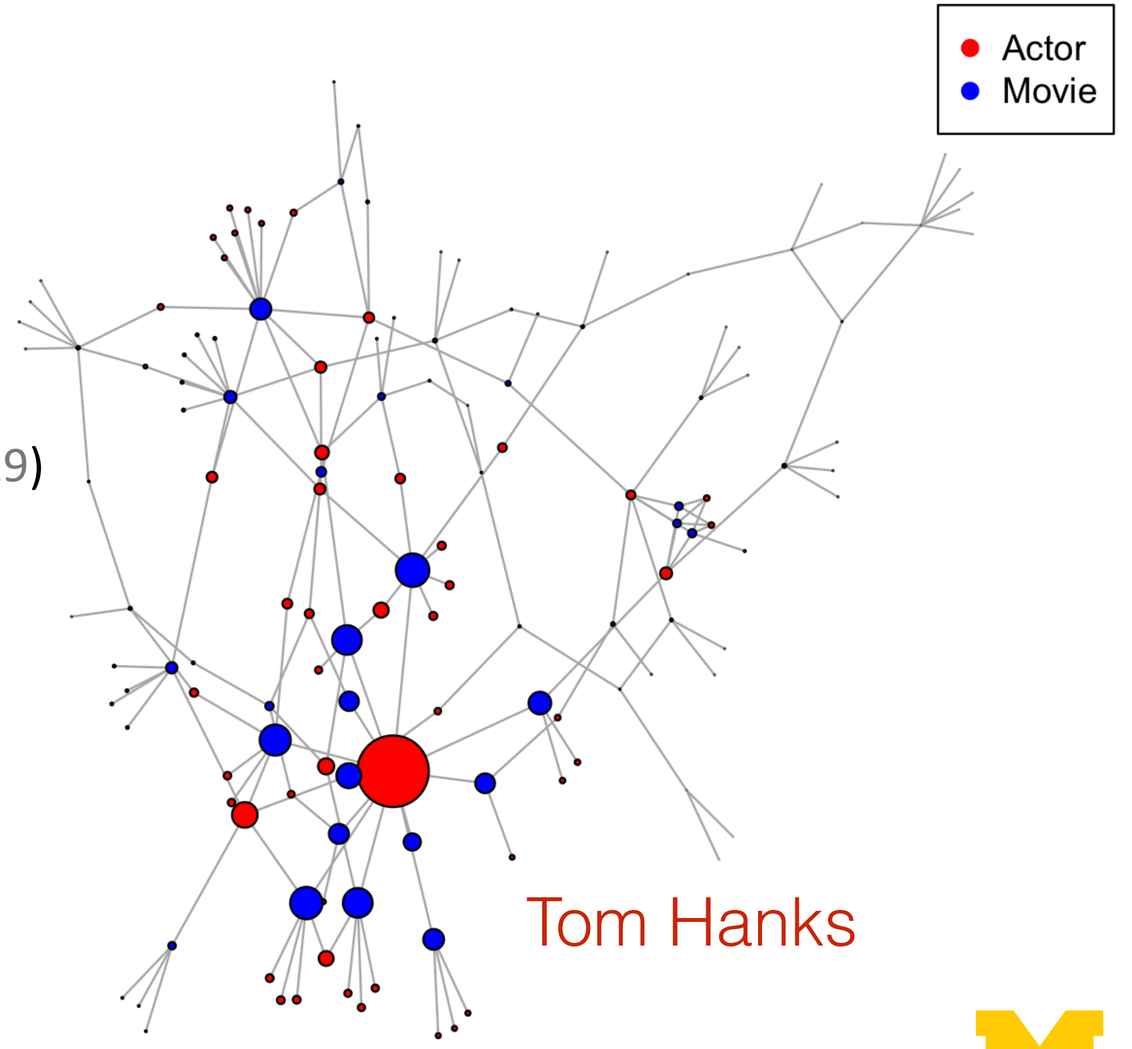


This means that Tom Hanks was in the most movies.

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Most Influential Actors (Eigenvector Centrality)

```
# Get eigenvector centrality
ig <- graph.data.frame(edges, directed=FALSE)
eig <- as.data.frame(eigen_centrality(ig)$vector)
V(ig)$color <- ifelse(nodes$label %in% edges[,1], "red", “blue")
V(ig)$label <- NA
V(ig)$size <- eig[,1]*15
plot(ig)
legend('topright', c("Actor", "Movie"), col = c("red", "blue"), pch = 19)
```



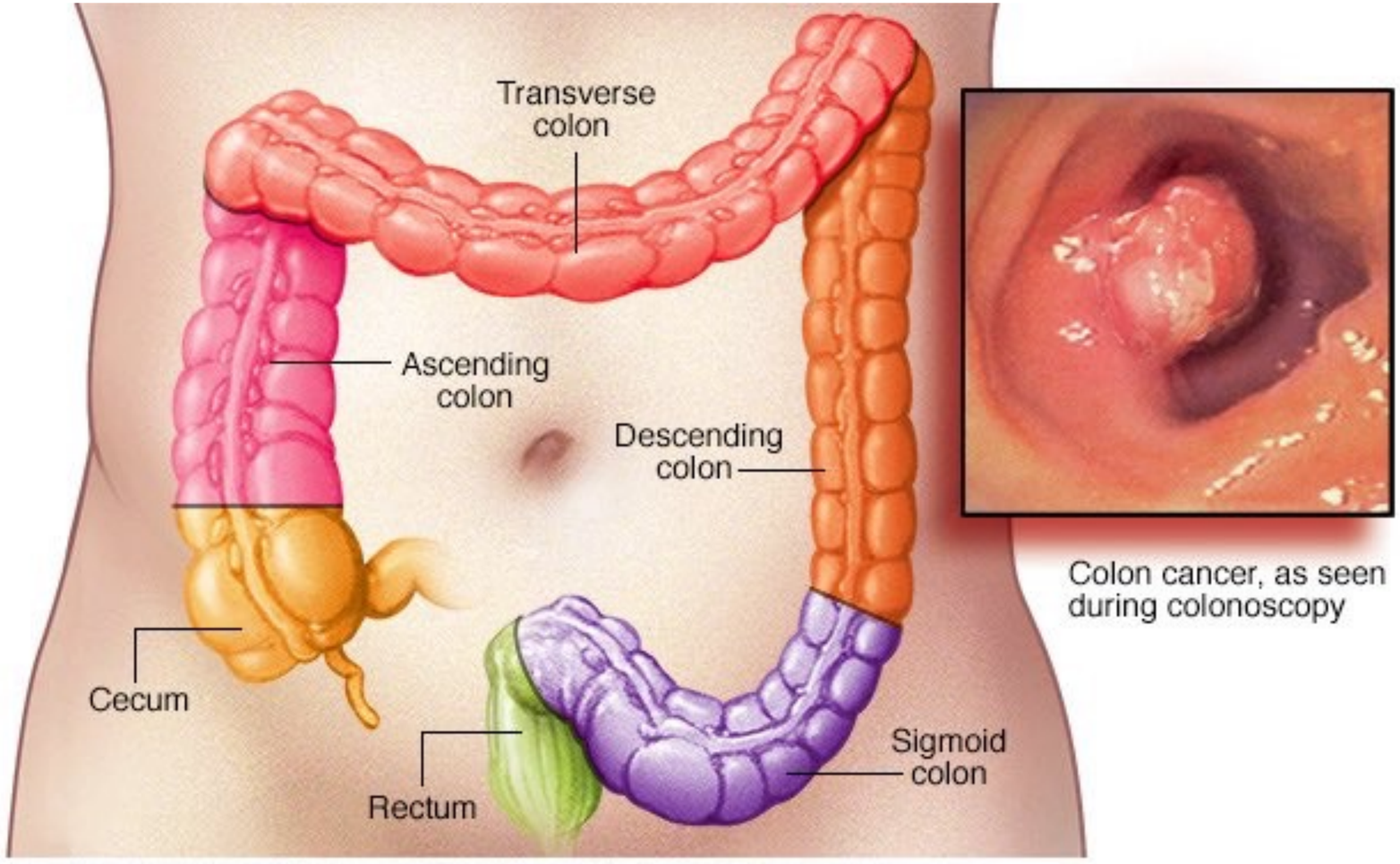
Tom Hanks was in the most movies that other actors also acted in. Tom Hanks is the most influential actor to this system.

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Colorectal Cancer:
Applying graphs to biological problems

Colorectal Cancer



MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Colorectal Cancer is Common

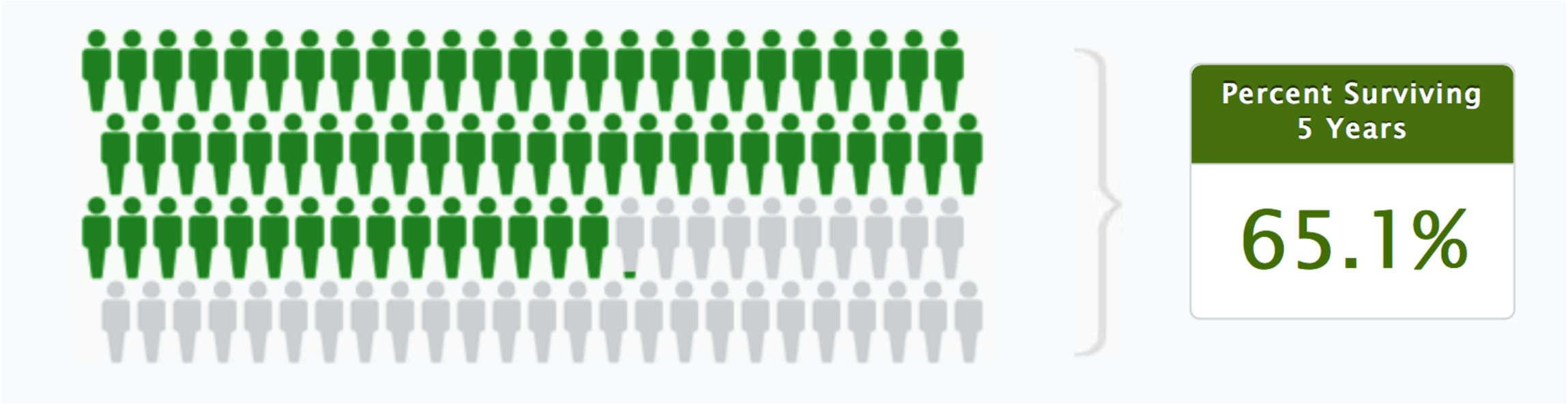
	Common Types of Cancer	Estimated New Cases 2016	Estimated Deaths 2016
1.	Breast Cancer (Female)	246,660	40,450
2.	Lung and Bronchus Cancer	224,390	158,080
3.	Prostate Cancer	180,890	26,120
4.	**Colon and Rectum Cancer**	**134,490 4th**	**49,190 2nd**
5.	Bladder Cancer	76,960	16,390
6.	Melanoma of the Skin	76,380	10,130
7.	Non-Hodgkin Lymphoma	72,580	20,150
8.	Thyroid Cancer	64,300	1,980
9.	Kidney and Renal Pelvis Cancer	62,700	14,240
10.	Leukemia	60,140	24,400

Colon and rectum cancer represents 8.0% of all new cancer cases in the U.S.

Howlader, N. *et al.* SEER Cancer Statistics Review, 1975-2013. *National Cancer Institute* (2016).

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Colorectal Cancer Has a Low Survival Rate

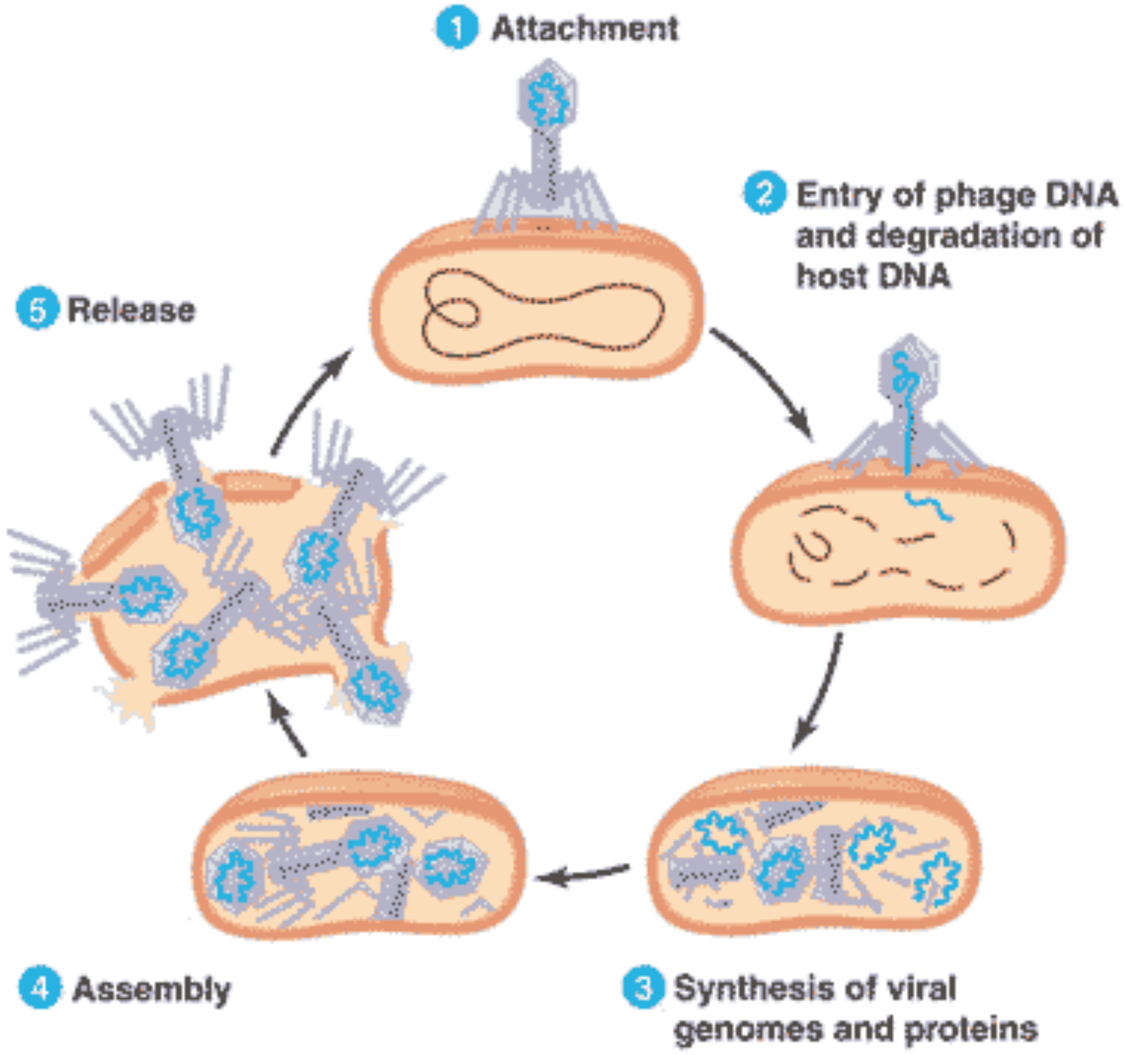


Howlader, N. *et al.* SEER Cancer Statistics Review, 1975-2013. *National Cancer Institute* (2016).

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

The Role of Microbes in Colorectal Cancer

- Cancer patients have altered bacterial and **bacterial virus (bacteriophage)** communities.
- These two entities interact as an ecological community, with the viruses killing bacteria.
- Understanding which viruses are the most influential “hubs” on the community will allow us to develop targeted therapeutics and potentially engineer the system.

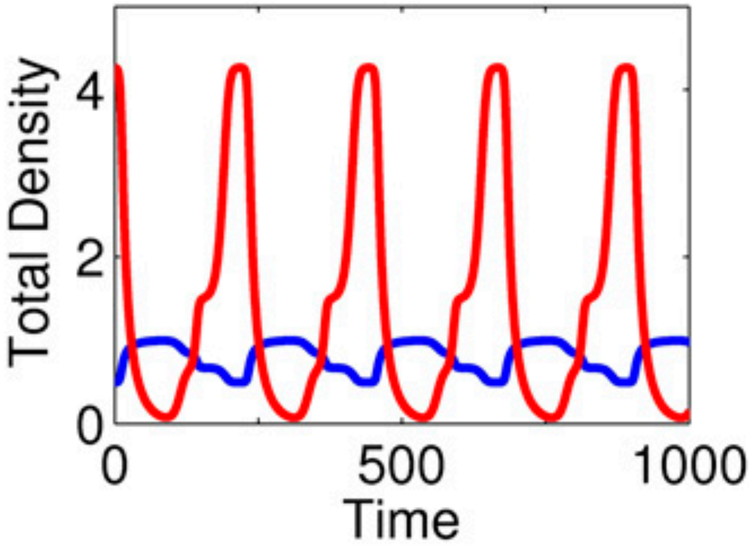


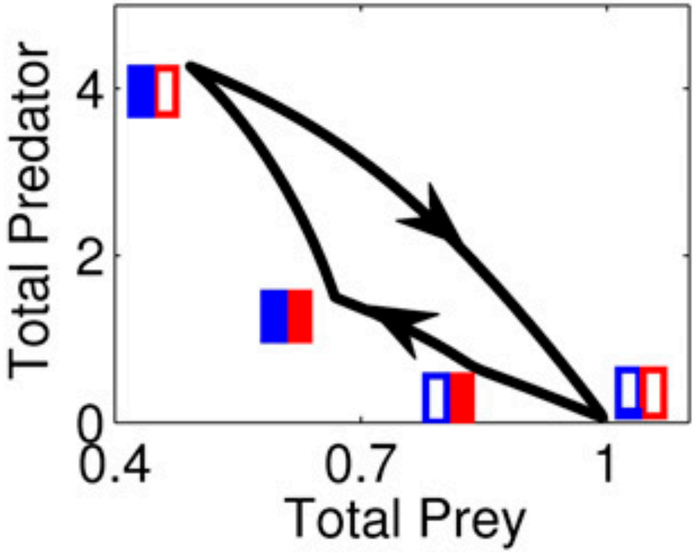
https://www.quia.com/files/quia/users/lmcgee/genetics/APchapter18-Viri/phage-lytic-cycle.gif

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

The Role of Microbes in Colorectal Cancer

- Cancer patients have altered bacterial and **bacterial virus (bacteriophage)** communities.
- These two entities interact as an ecological community, with the viruses killing bacteria.
- Understanding which viruses are the most influential “hubs” on the community will allow us to develop targeted therapeutics and potentially engineer the system.

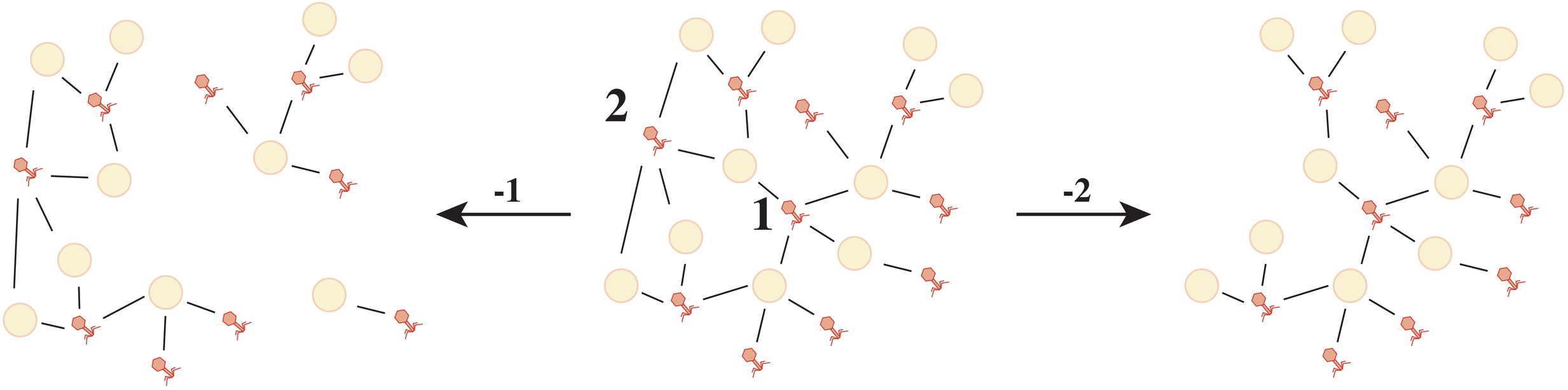




Cortez, M. H. & Weitz, J. S. Coevolution can reverse predator-prey cycles. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 7486–7491 (2014).

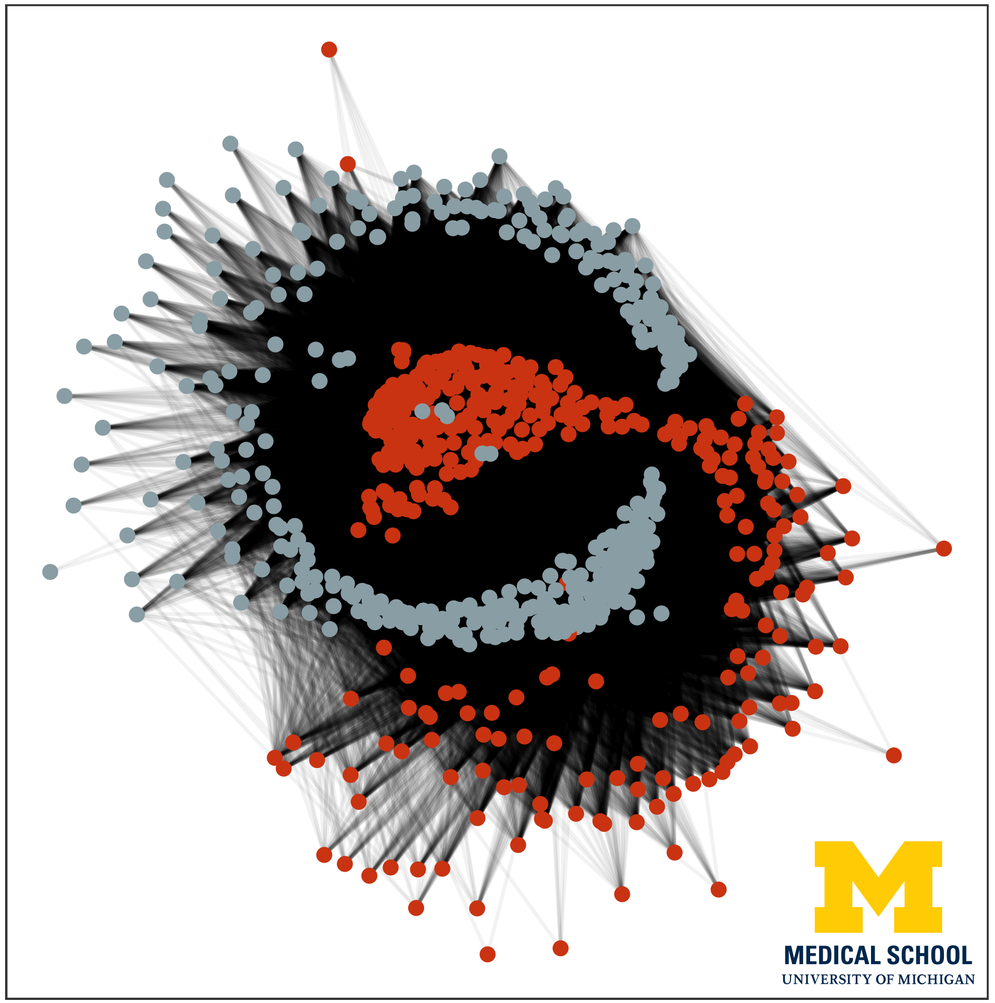
MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Network Centrality Identifies Keystone Phages

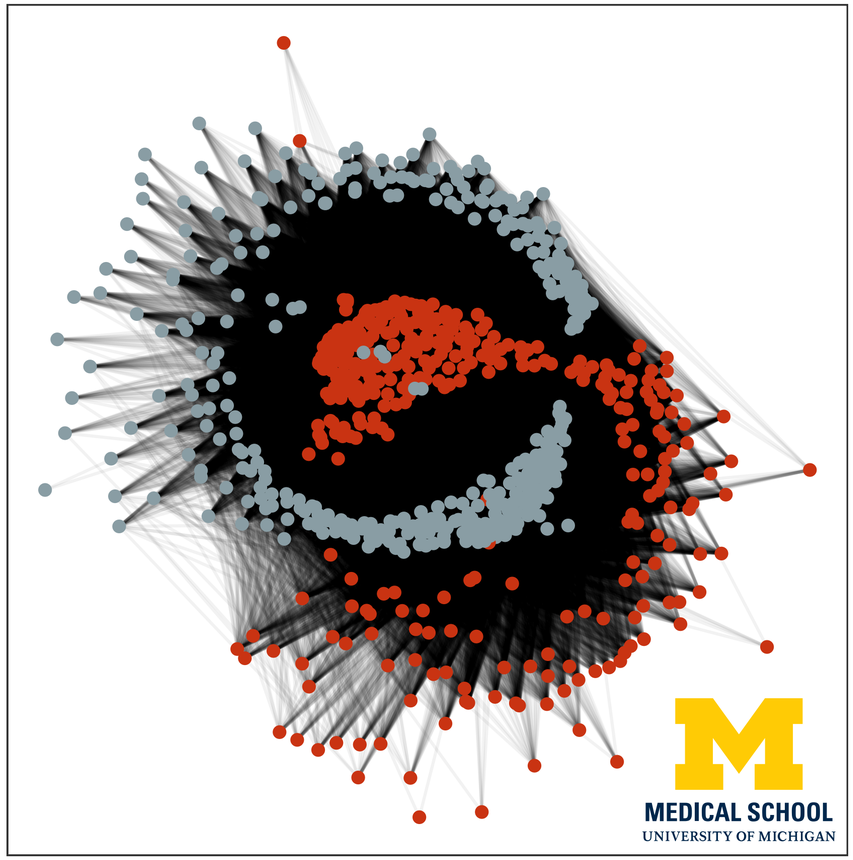


- Centrality is a measurement of importance and influence within a community.
- Alpha centrality detects phages highly connected to bacteria which are highly connected to other phages (central in community).
- High centrality indicates keystone phages of community. Important for structure.

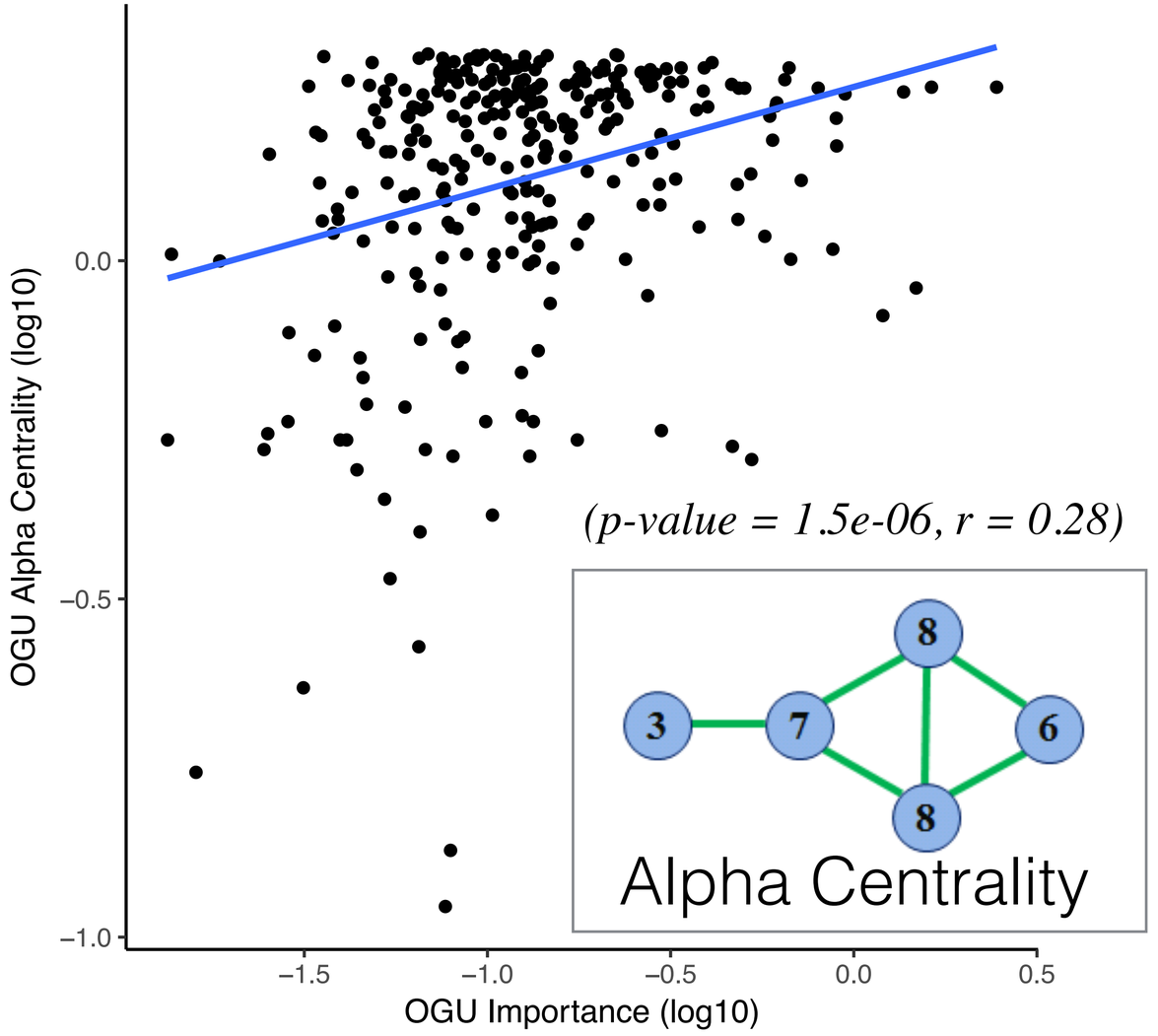
We Built a Network From Predicted Infections



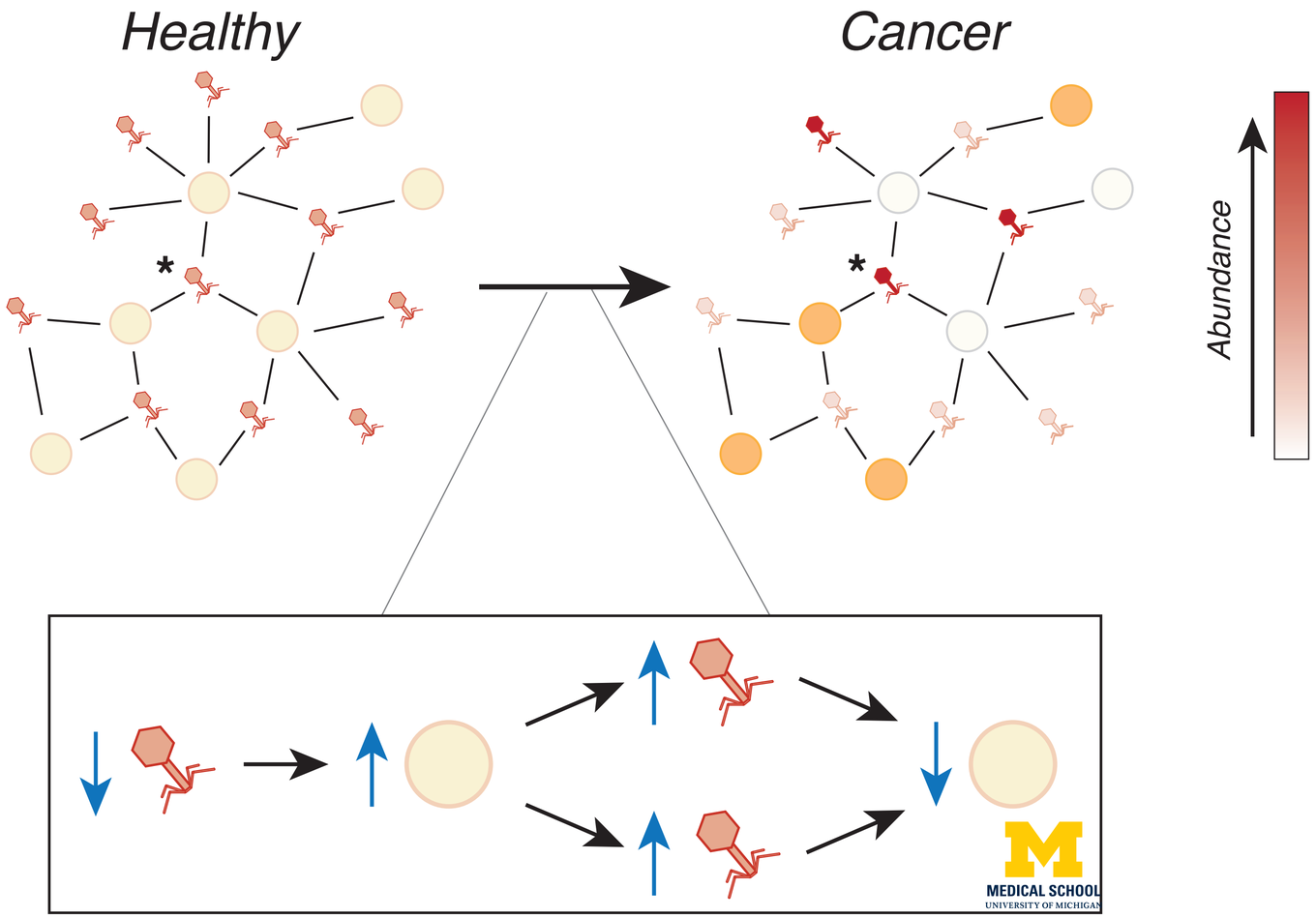
Correlation of Host Range and Cancer Importance



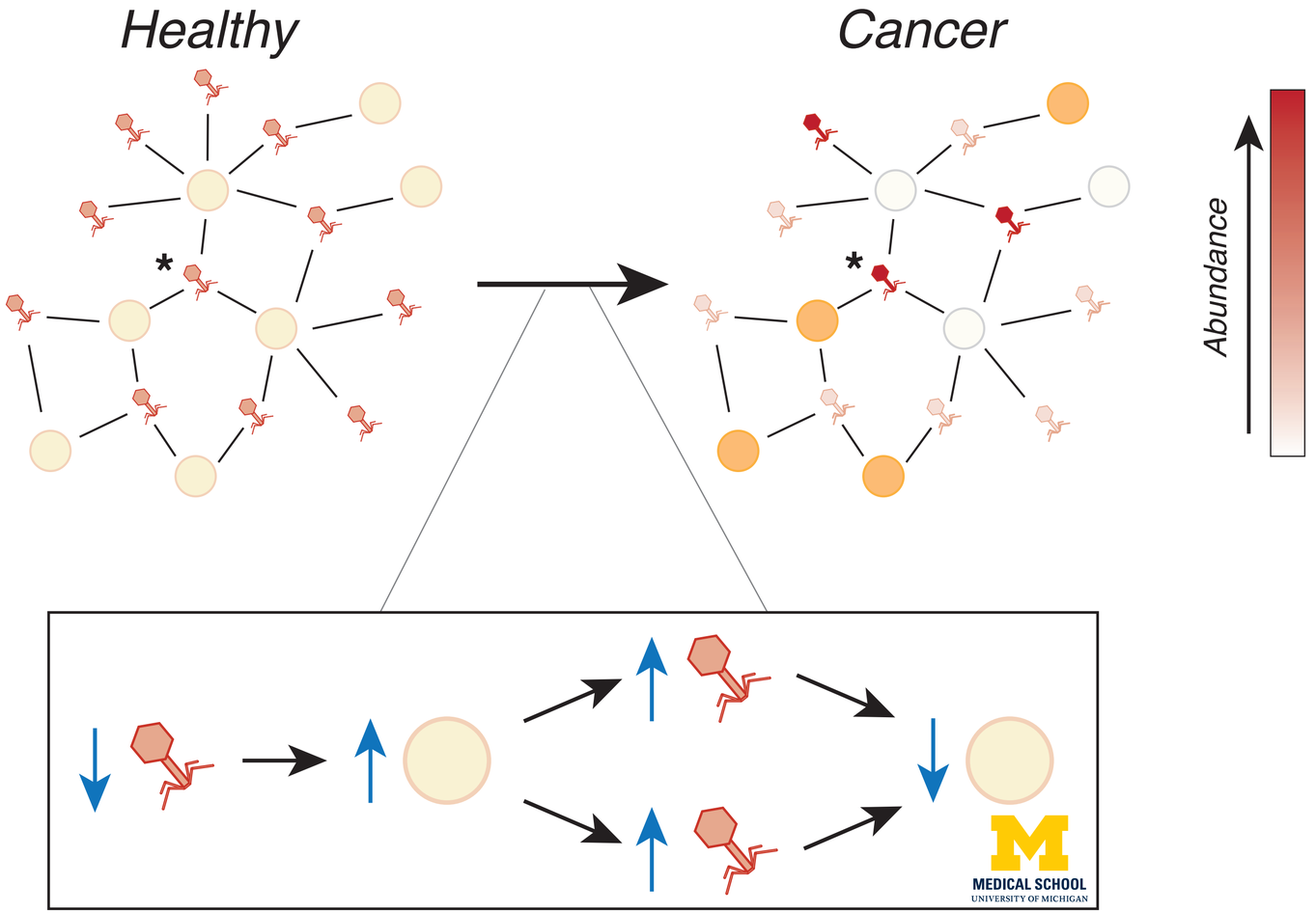
label
- Bacteria
- Phage



Modulation of Influential Phages Alters the Bacterial Community



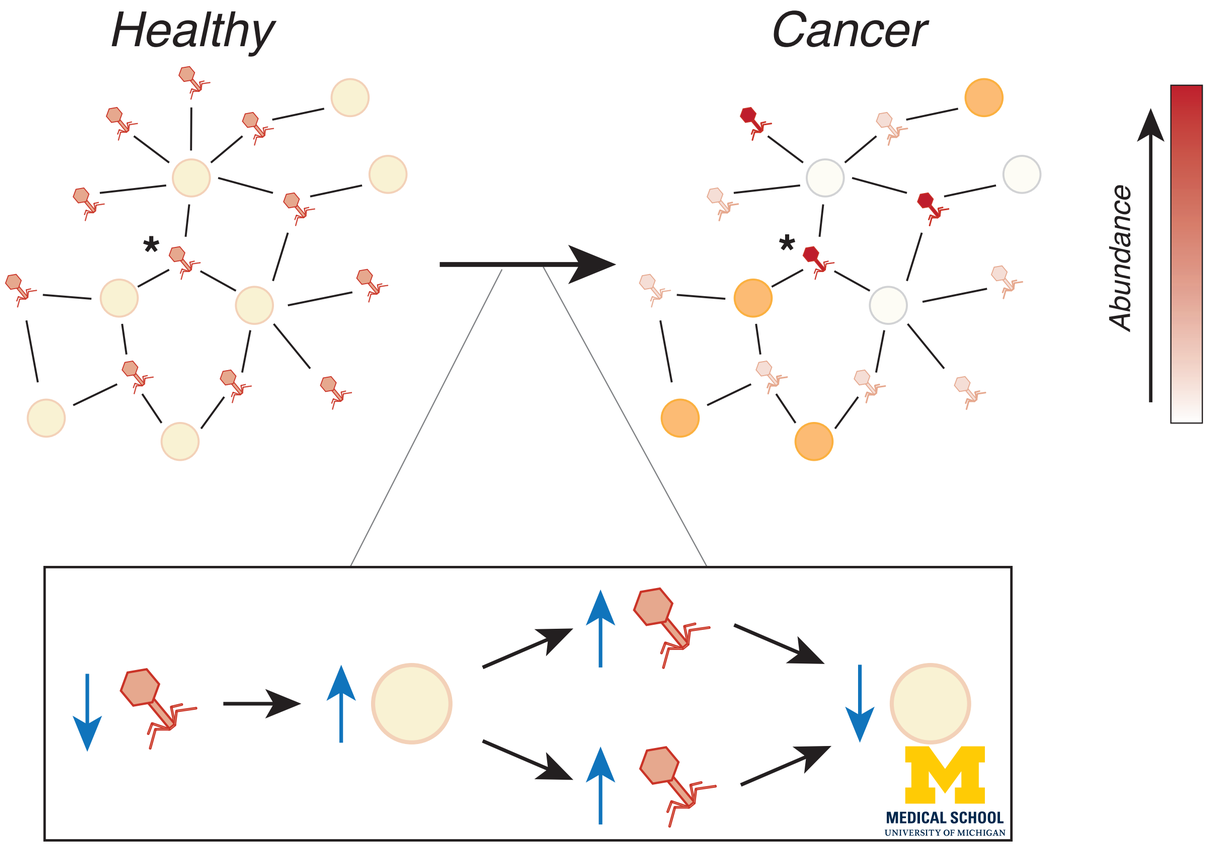
Modulation of Influential Phages Alters the Bacterial Community



Zackular,J.P. *et al.* (2016) Manipulation of the Gut Microbiota Reveals Role in Colon Tumorigenesis. *mSphere*, **1**, e00001–15.

Conclusions

- Just like we identified the most influential actors in our movie network, we are able to identify the most influential phages in the intestines.
- The phage hubs (most influential) were also the most important in colorectal cancer.
- By identifying the most influential phages for cancer and the ecosystem, we can begin developing targeted approaches to engineering these intestinal ecosystems.



MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN

Wrapping it up.

Conclusions

- Graphs allow for a new level of analysis by focusing on relationships.
- Neo4j allows for efficient and scalable graph databasing.
- RNeo4j and igraph provide robust analysis tools for neo4j graphs.
- Graphs provide new insight into medicine and mirobiology.
- Highly connected phages are associated with colorectal cancer.

How to Contact Me

@iprophage

Microbiology

prophage.blogspot.com

microbiology.github.io

ghannig@umich.edu

MEDICAL SCHOOL
UNIVERSITY OF MICHIGAN