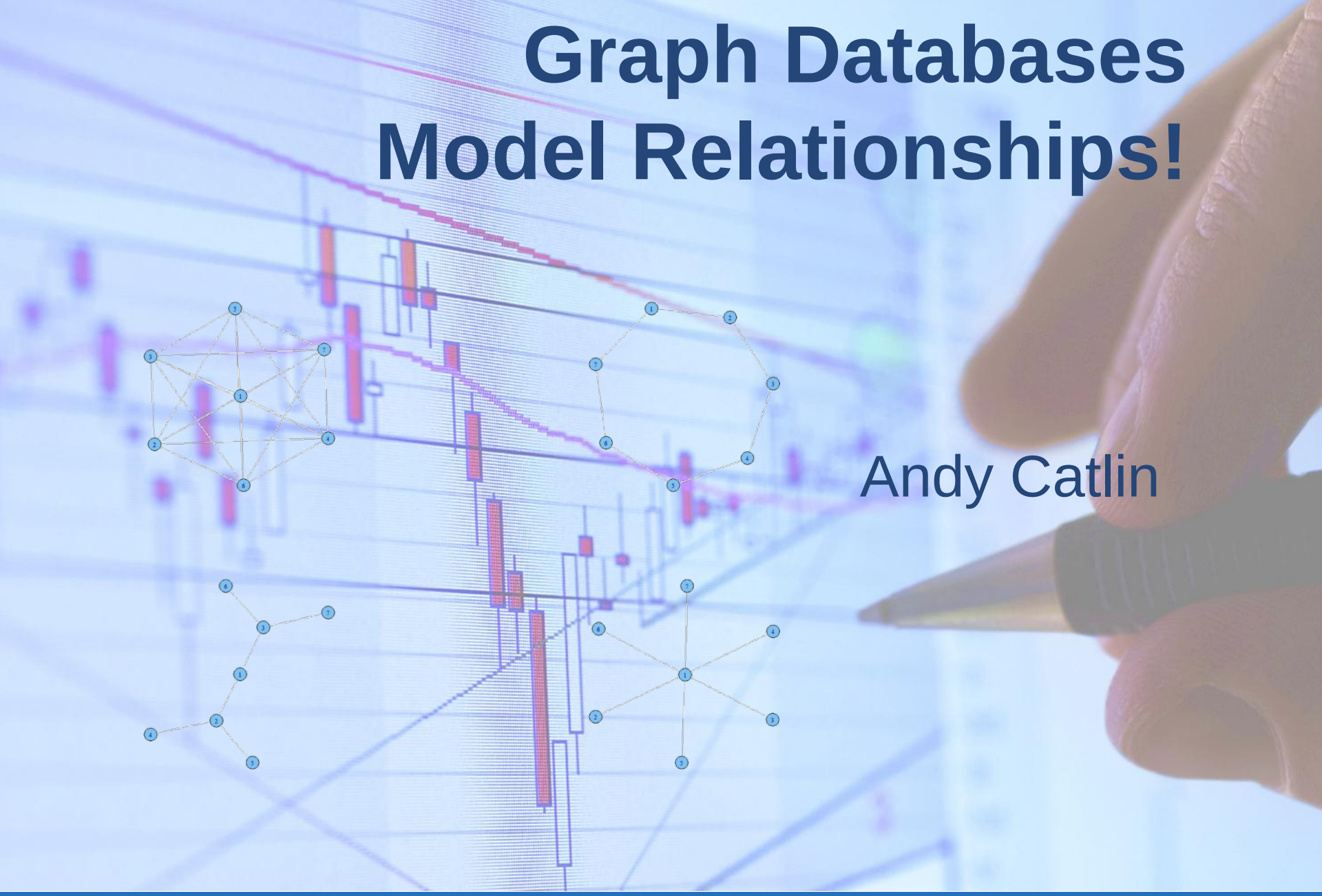
Graph Databases Model Relationships!

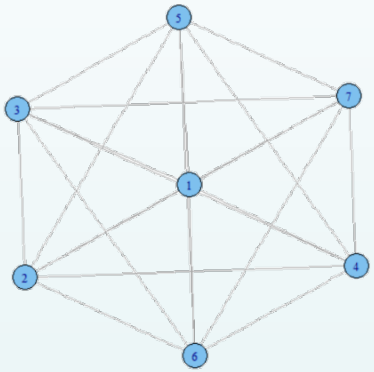


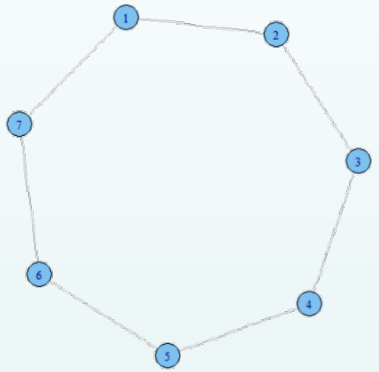
Andy Catlin

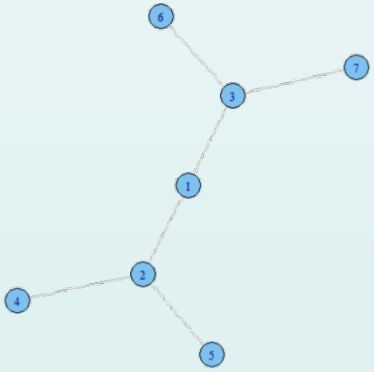
Session Goals

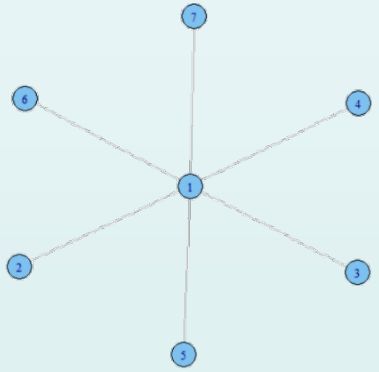
- **Map vertices and edges using the `igraph` package**
- Understand limits of the relational model
- Describe some use cases for network data, and some important network descriptive analytics
- Load network data from a CSV file into `igraph`
- Create an interactive web page from your network data

Vertices (nodes) and Edges (links)

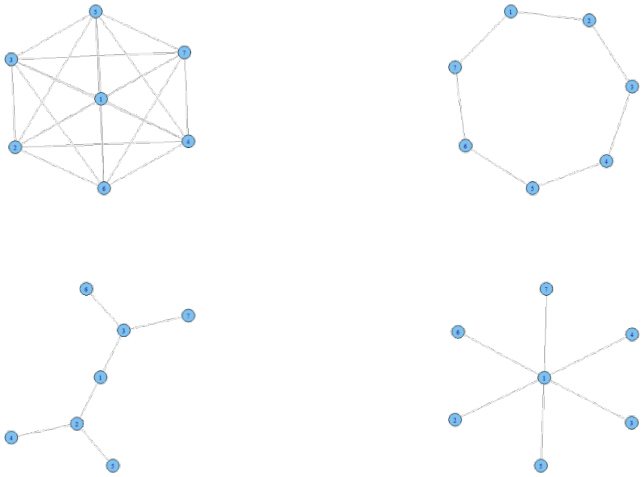








Vertices (nodes) and Edges (links)



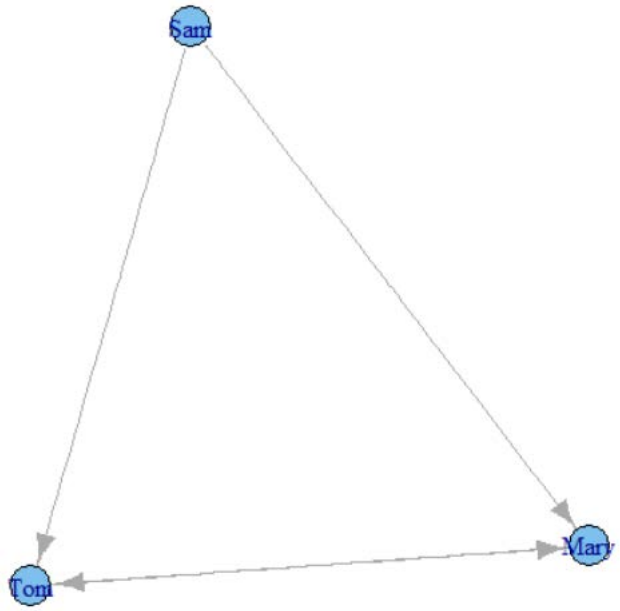
```
# source:
#    https://github.com/kolaczyk/sand

#install.packages("igraph")

library(igraph)

g.full <- graph.full(7)
g.ring <- graph.ring(7)
g.tree <- graph.tree(7, children=2,
            mode="undirected")
g.star <- graph.star(7,
            mode="undirected")
par(mfrow=c(2, 2))
plot(g.full)
plot(g.ring)
plot(g.tree)
plot(g.star)
```

Which of these graph types can be modeled in a relational database?




```
dg <- graph.formula(Sam-+Mary,
        Sam-+Tom, Mary++Tom)
plot(dg)
```

How would you model the following relationships?

There are flights from Honolulu to and from Los Angeles, from Los Angeles to New York, and from New York to Honolulu.

Session Goals

- Map vertices and edges using the `igraph` package
- **Understand limits of the relational model**
- Describe some use cases for network data, and some important network descriptive analytics
- Load network data from a CSV file into `igraph`
- Create an interactive web page from your network data

Limits of the Relational Model

- **Relational databases model data**. Unspecified single relationship provided by REFERENCES clause.
- **Graph database model relationships**. Relationships are first class objects.

RDBMS vs.Graph Modeling

Every path is of length 1:

Three column table (node, edge, node)

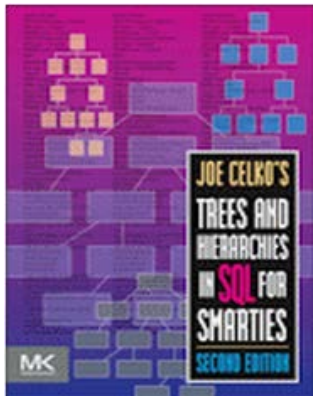
Increase length – add one self-join ("expanding search radius")

Degrees of Separation

- http://oracleofbacon.org/ ~2.6M actors

- Celko's tests (graph db up to 9,000x faster)
  - Find the degrees of separation with Kevin Bacon as "center of the universe"
  - Find a relationship between any two actors.
  - Change focus actor
  - Use only actress links
  - Add directors, etc.

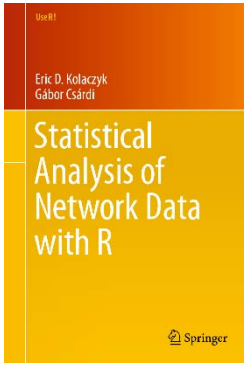
- “You quickly get into Cartesian explosions for longer paths, and can get lost in endless loops if there are cycles.” – Joe Celko



Session Goals

- Map vertices and edges using the `igraph` package
- Understand limits of the relational model
- **Describe some use cases for network data, and some important network descriptive analytics**
- Load network data from a CSV file into `igraph`
- Create an interactive web page from your network data

Source:



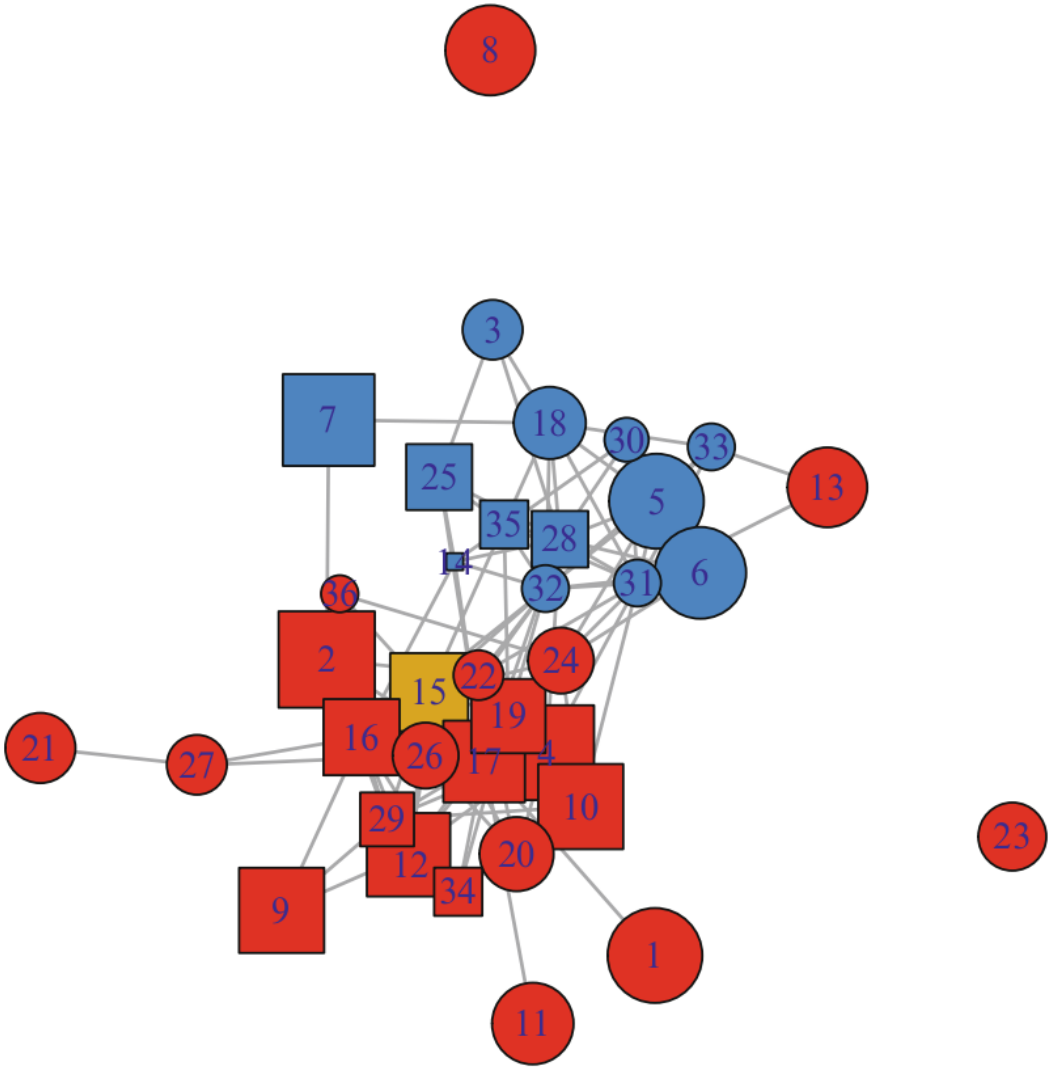


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

CUNY

To what extent are two lawyers that both work with a third lawyer likely to work with each other as well?

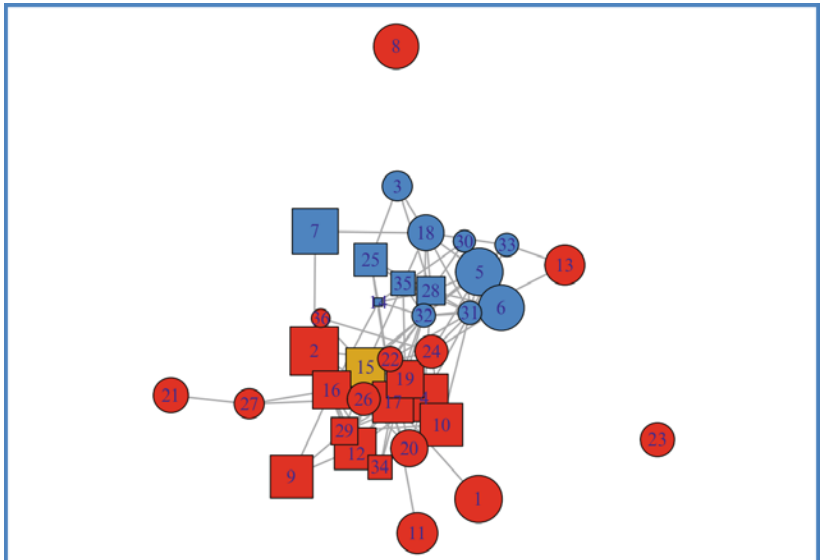


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

To what extent are two lawyers that both work with a third lawyer likely to work with each other as well?

That is, what proportion of vertex triples form triangles (i.e. all three vertexes are connected by edges)?

This measure of **transitivity** is summarized in a ***clustering coefficient***.

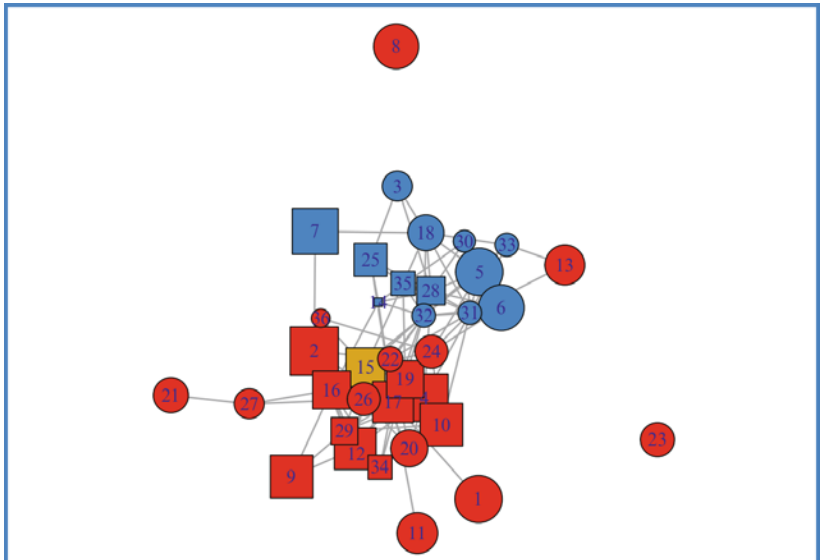


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

To what extent do lawyers of the same type (corporate, litigation) collaborate with those of the same type vs. other types?

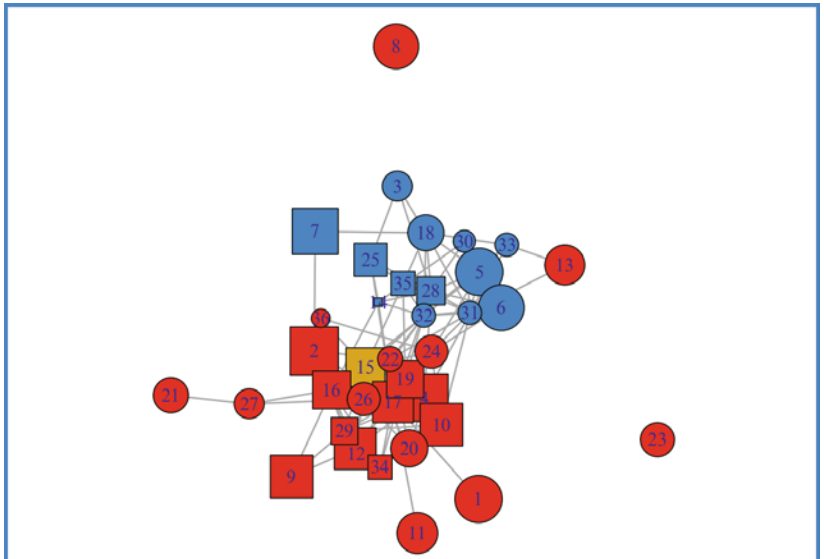


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

To what extent do lawyers of the same type (corporate, litigation) collaborate with those of the same type vs. other types?

Measured by ***assortativity coefficient*** (a type of correlation statistic, where labels of connected pairs of vertices are compared). Here, the focus is on an attribute associated with network vertices (i.e. lawyer practice) and the network structure plays a comparatively implicit role.

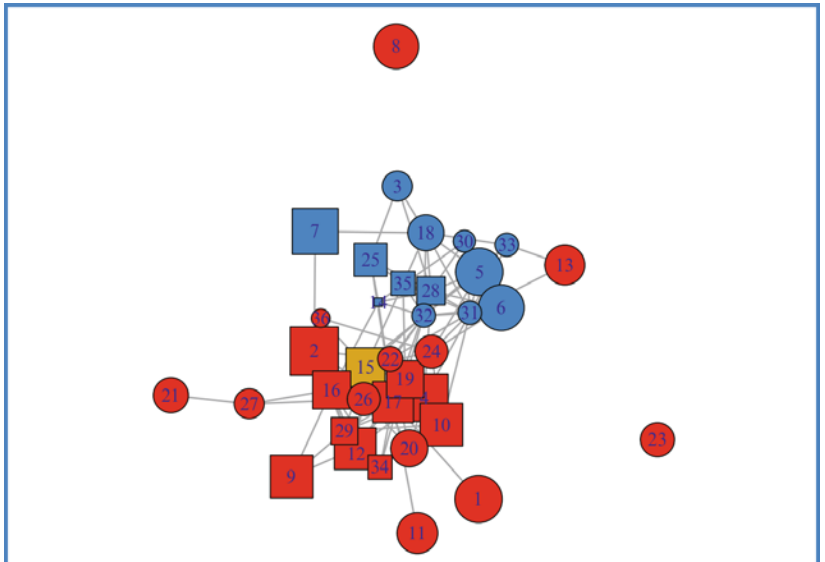


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

Could we predict with some accuracy the practice of a lawyer in a network if we know (1) the vertices of the neighbors of that lawyer in the network graph, and (2) the practice of those neighbors?

What are kinds of predictions might we want to do on network data?

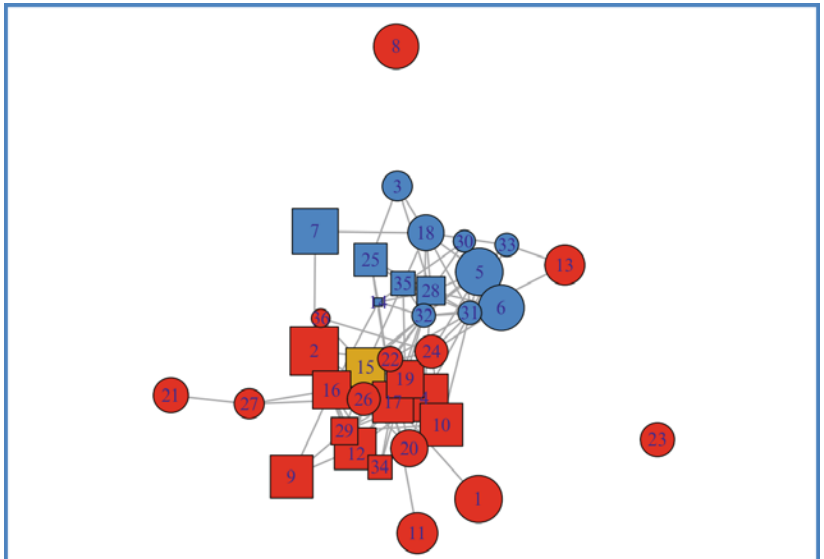


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

Could we predict with some accuracy the practice of a lawyer in a network if we know (1) the vertices of the neighbors of that lawyer in the network graph, and (2) the practice of those neighbors?

Are my Facebook friends' consumer preferences indicative of my own?

…

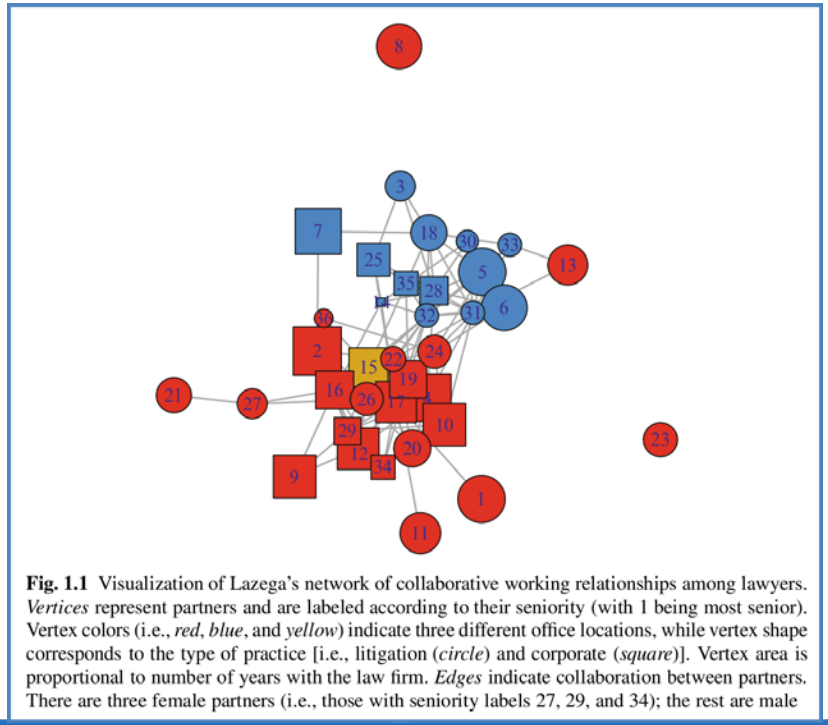


Fig. 1.1 Visualization of Lazega's network of collaborative working relationships among lawyers. *Vertices* represent partners and are labeled according to their seniority (with 1 being most senior). Vertex colors (i.e., *red*, *blue*, and *yellow*) indicate three different office locations, while vertex shape corresponds to the type of practice [i.e., litigation (*circle*) and corporate (*square*)]. Vertex area is proportional to number of years with the law firm. *Edges* indicate collaboration between partners. There are three female partners (i.e., those with seniority labels 27, 29, and 34); the rest are male

Session Goals

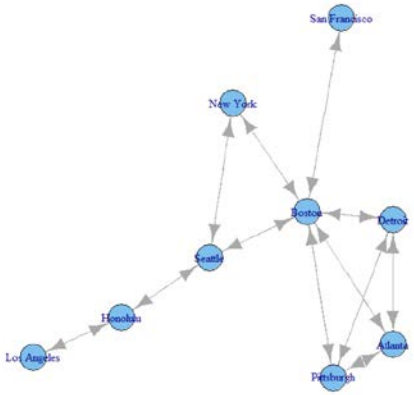
- Map vertices and edges using the `igraph` package
- Understand limits of the relational model
- Describe some use cases for network data, and some important network descriptive analytics
- **Load network data from a CSV file into `igraph`**
- Create an interactive web page from your network data

depart	arrive	flighttime
Boston	San Francisco	399
San Francisco	Boston	319
Los Angeles	Honolulu	362
Honolulu	Los Angeles	336
Boston	New York	74
New York	Boston	74
New York	Seattle	495
Seattle	New York	495
Seattle	Honolulu	355
Honolulu	Seattle	355

A few rows from **flights.csv**

```
df.flights <-
  read.table("c://data//flights.csv",
                        sep=",", header=TRUE,
                        stringsAsFactors=FALSE)

flights <- graph.data.frame(df.flights,
                        directed=TRUE)
plot(flights)
```



Session Goals

- Map vertices and edges using the `igraph` package
- Understand limits of the relational model
- Describe some use cases for network data, and some important network descriptive analytics
- Load network data from a CSV file into `igraph`
- **Create an interactive web page from your network data**

```
#install.packages("d3Network")
library(d3Network)

d3SimpleNetwork(df.flights,
Source="depart", Target = "arrive",
width=1200, height=800,
file="RouteMap.html")
```