DATA PREPARATION

Dr. Aric LaBarr

Institute for Advanced Analytics

Course Layout

Data Preparation
- Transactional Data
- Recency vs. Frequency
- Network Features

Anomaly Models
- Univariate Analysis
- Clustering
- Isolation Forests
- CADE

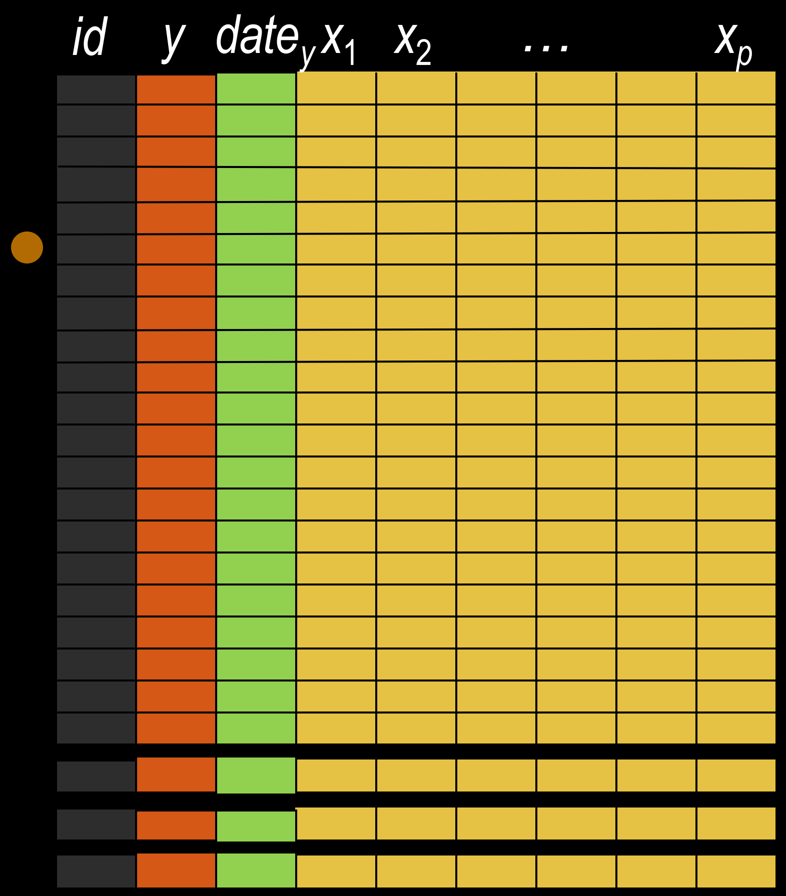
Fraud Supervised Models
- SMOTE
- Models
- Labeled vs. Unlabeled Bias
- Not Fraud Model
- Evaluation

Clusters of Not Goods
- Cluster Analysis
- Social Network Analysis

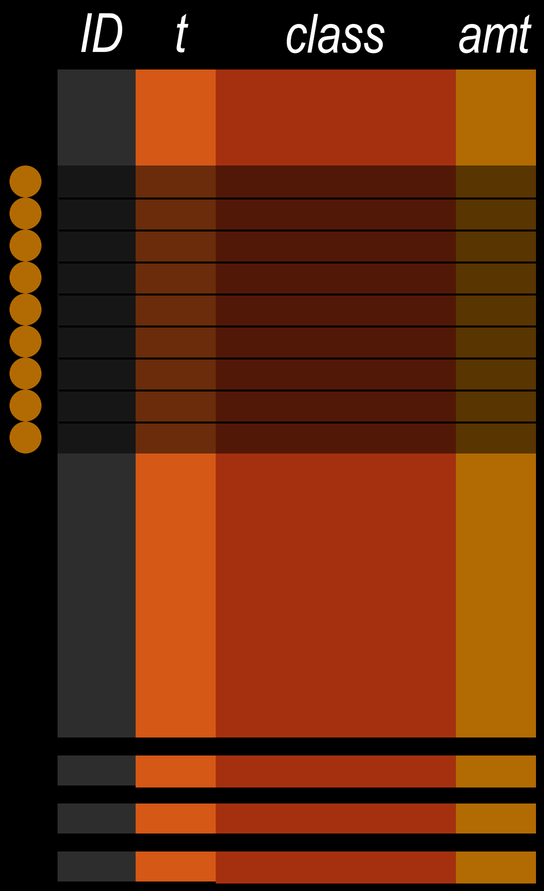
Implement
- Investigators
- Traffic Light Indicators
- Backtesting

FEATURE ENGINEERING

Transaction Data



Model Development Data



Transaction Data

Transaction Data Examples

- There are many different fields where transactional data plays an important role:
  - Credit card purchasing data
  - Medical claims data
  - Insurance claims data
  - Retail purchasing data
  - Etc.

Transaction Data Examples

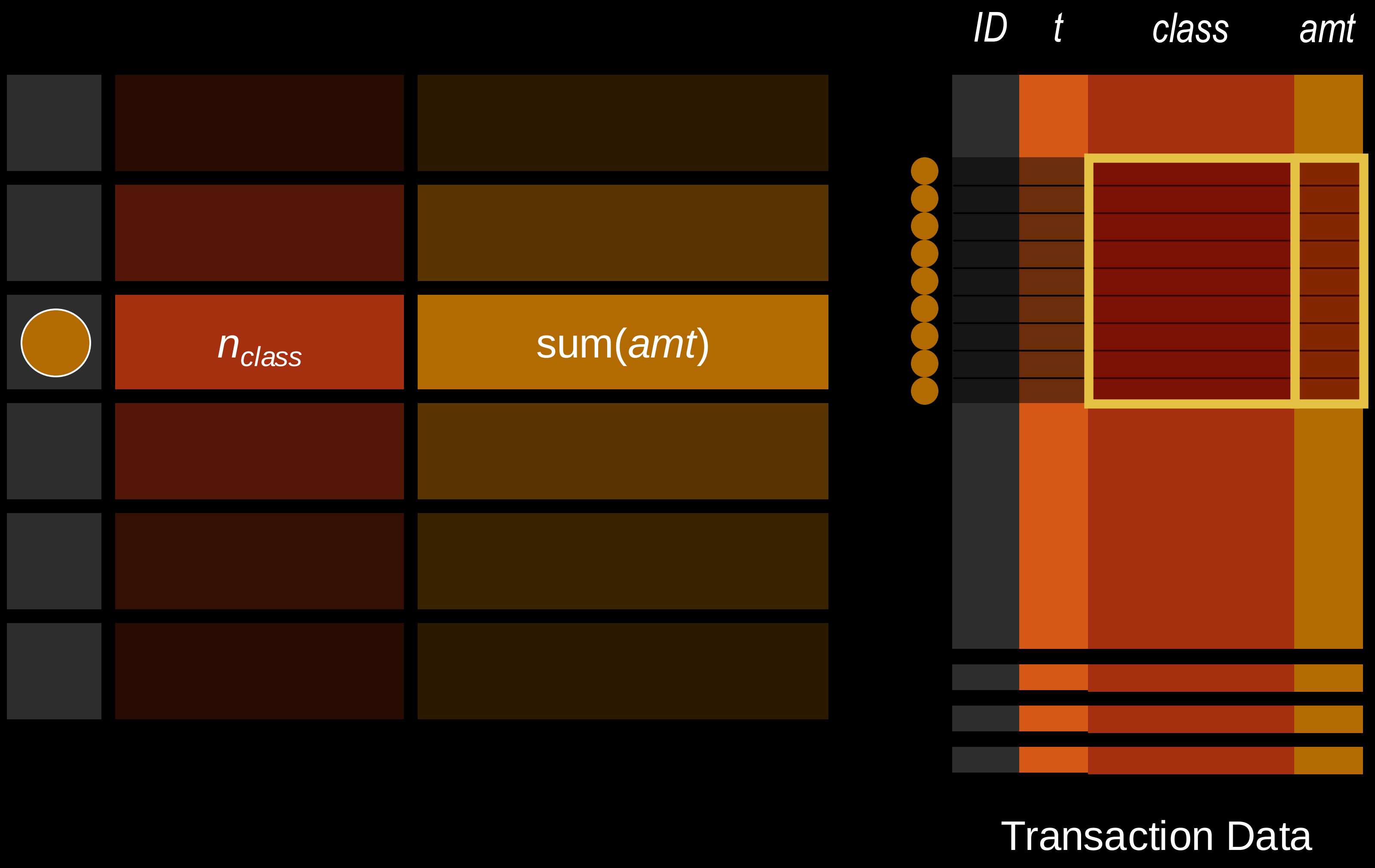
- There are many different fields where transactional data plays an important role:
  - Credit card purchasing data
  - Medical claims data
  - Insurance claims data
  - Retail purchasing data
  - Etc.

THINK OF YOUR DATA SPECIFICALLY!

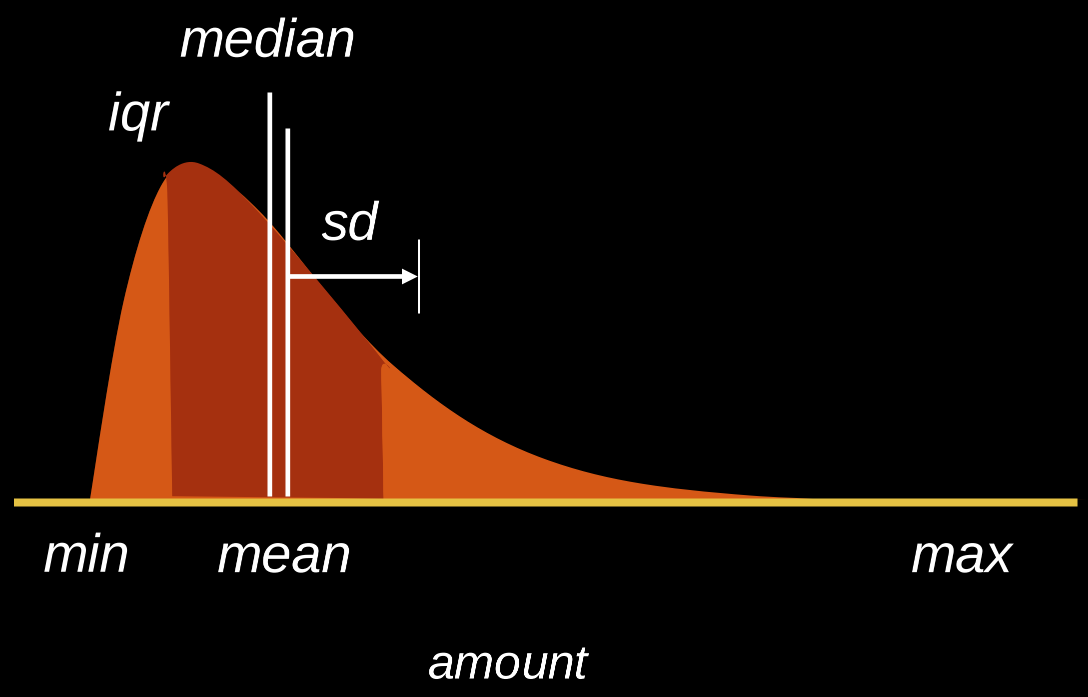
Transactions Data

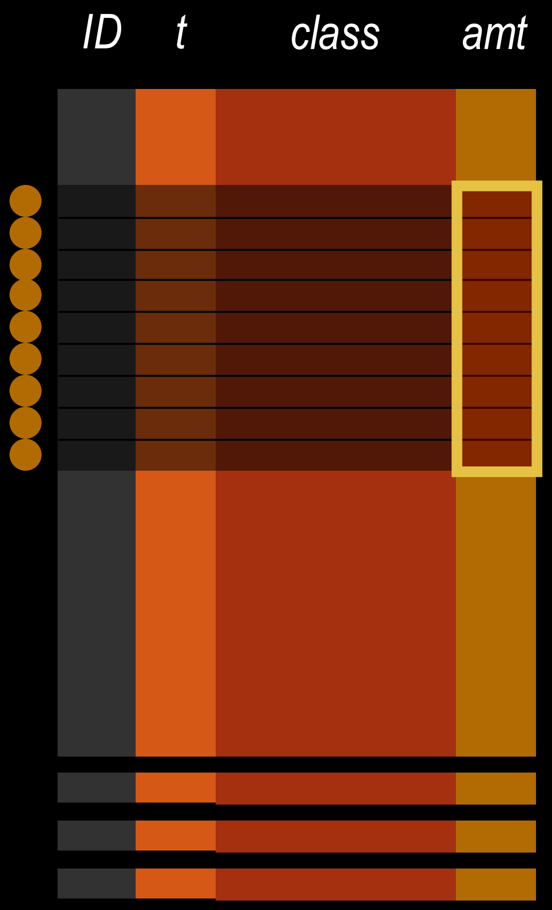
- Advantages
  - Highly Detailed
  - Captures Individual Behavior
  - Strong Target Correlation Possible
- Challenges
  - Highly Detailed
  - Difficult to Obtain
  - Difficult to Process

Input Possibilities: Tabulations



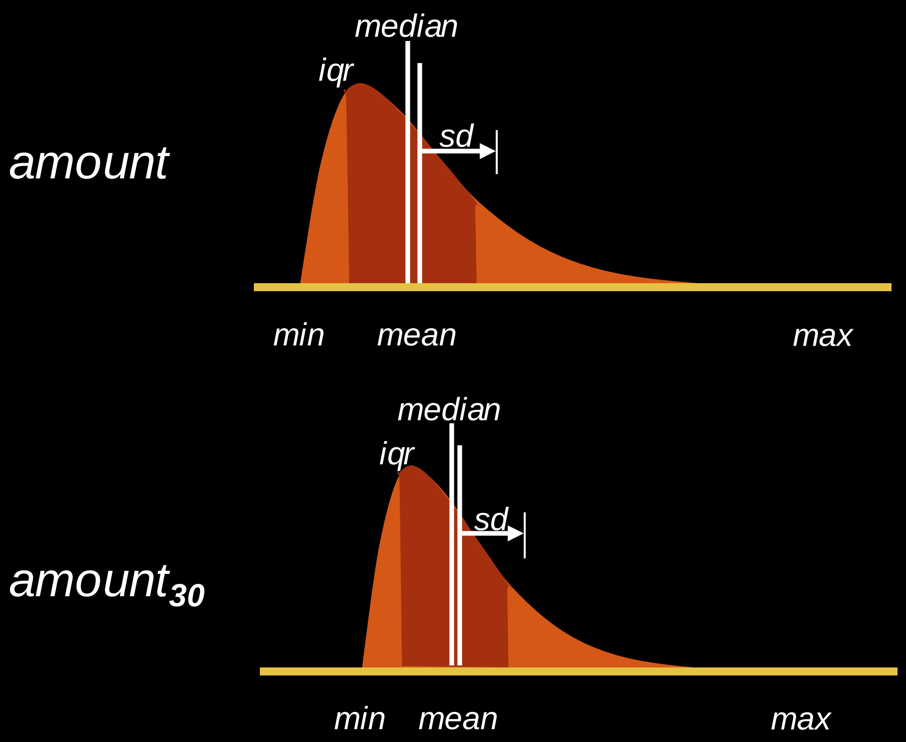
Input Possibilities: Distributions

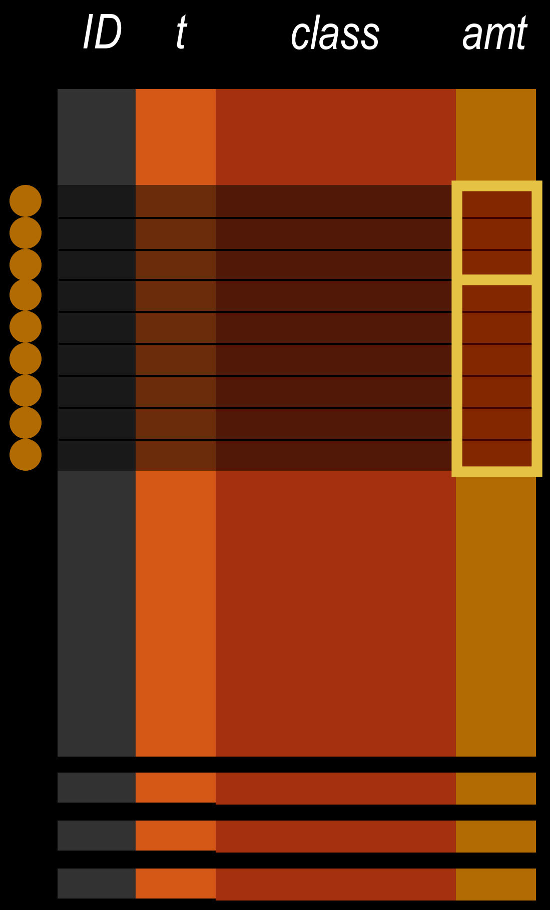




Transaction Data

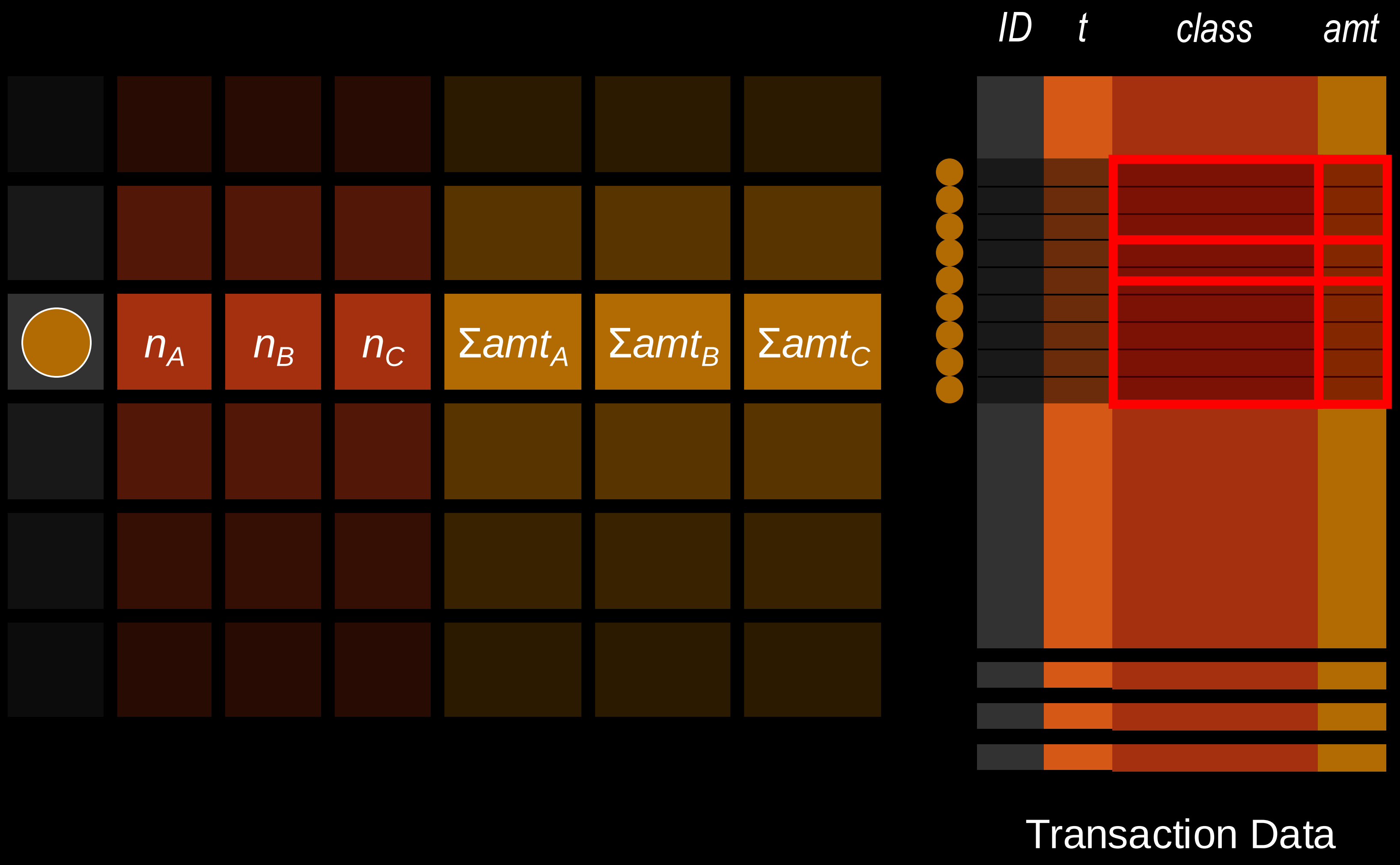
Input Possibilities: Stratifications



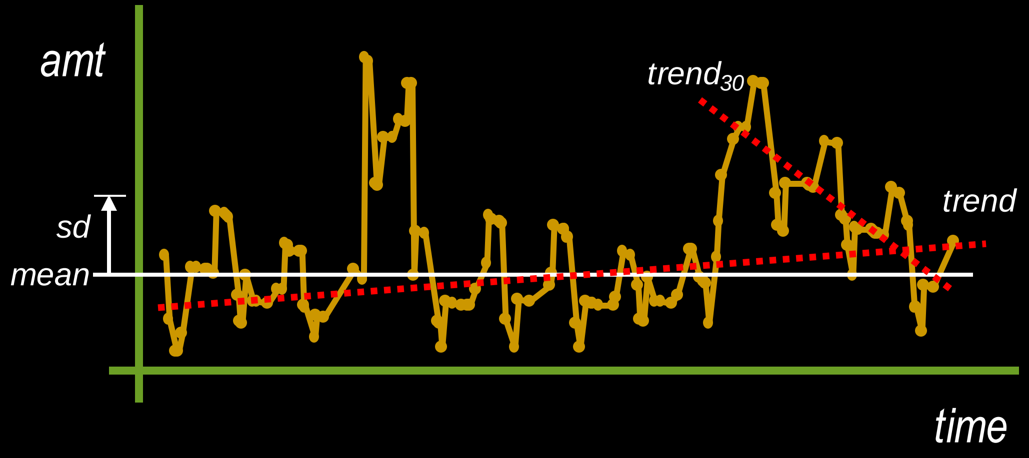


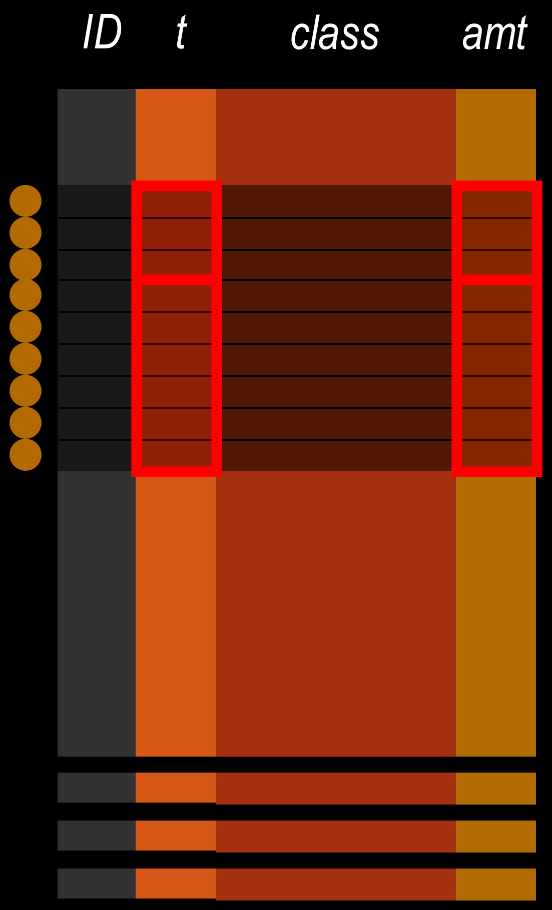
Transaction Data

Input Possibilities: Profiles



Input Possibilities: Time Series





Transaction Data

Process Transactions

- Here are the steps you need to take to process transactional data:
  1. Select your data.
  2. Sort your data.
  3. Augment your data.
  4. Process by ID.
  5. Finalize.

Grouping Transaction-Derived Inputs

- Examples
  - Mean of last five transactions
  - Standard deviation of transactions in last 14 days
  - Largest transaction per week
  - Slope of line fit to number of transactions per week (negative?)

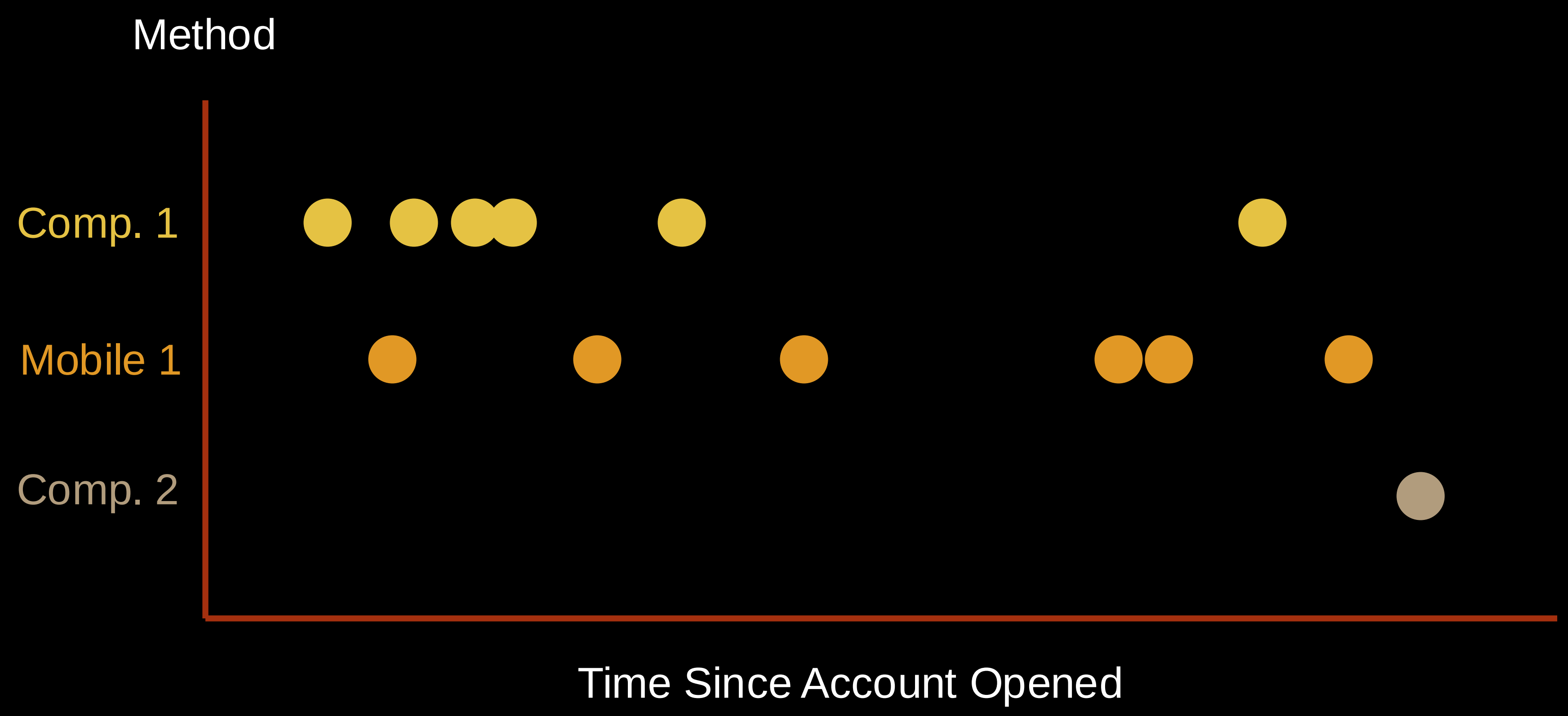
RECENCY & FREQUENCY

Recency & Frequency

- Transactional data provides extensive information.
- Two of the most important things in fraud detection (as well as other fields) are **recency** and **frequency** of transaction.

- **Recency** – time in between transactions
- **Frequency** – how often transactions occur

Online Account Access Example

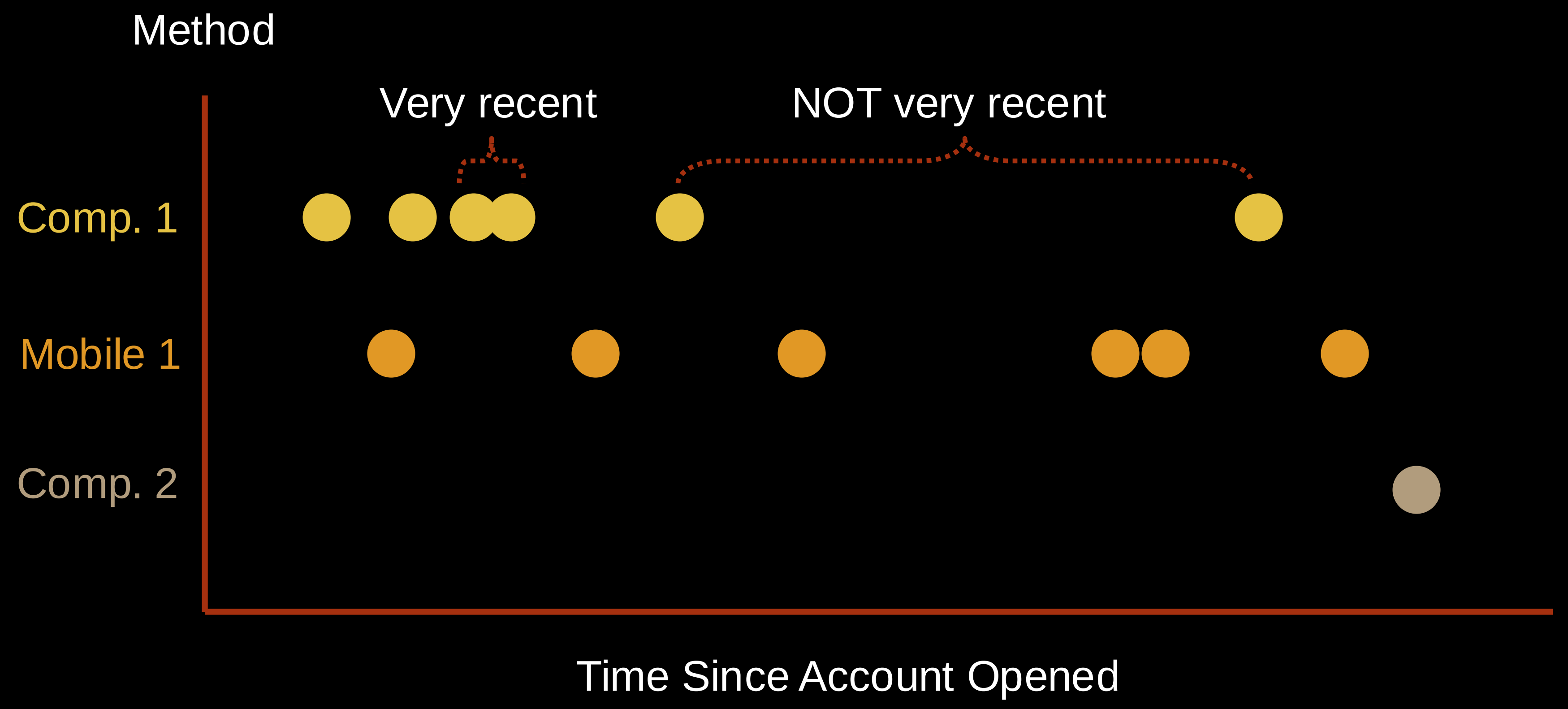


Recency

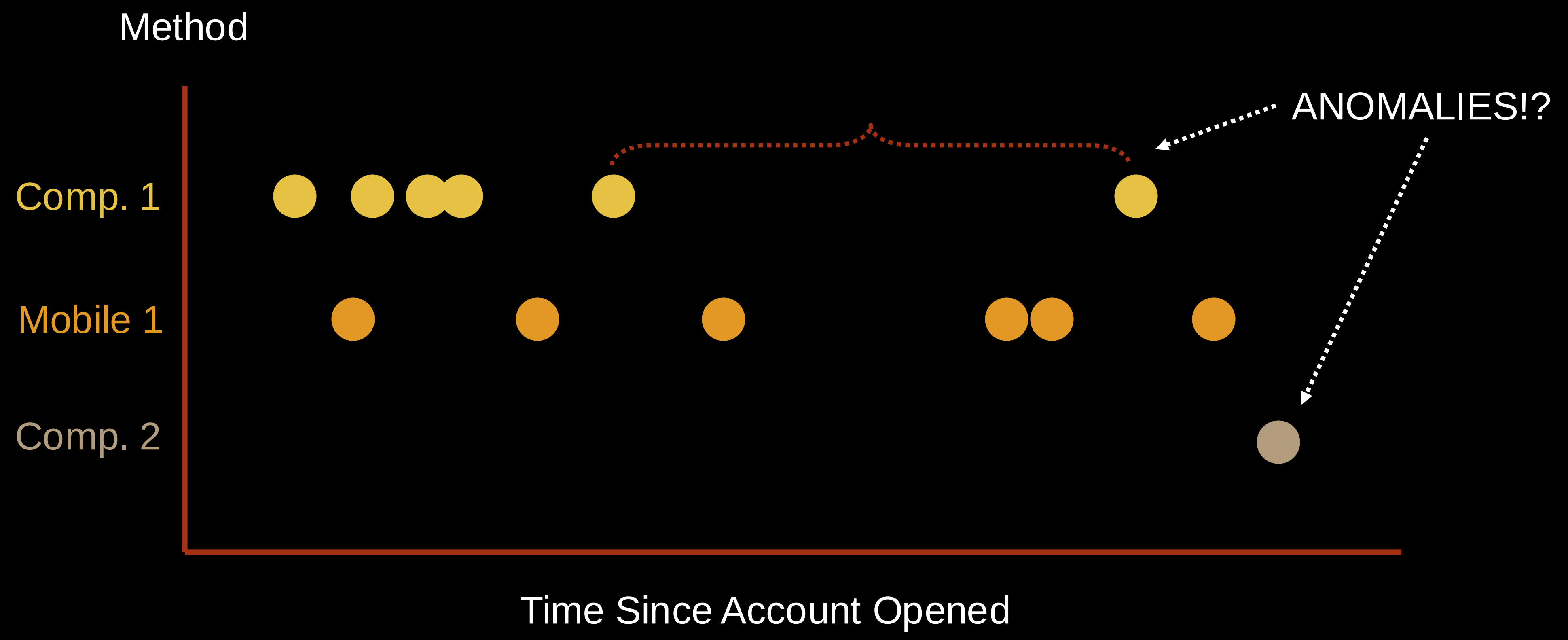
- **Recency** – time in between transactions

- Easy features:
  - Time in between transactions
  - Time since last transaction

Online Account Access Example



Online Account Access Example



Frequency

- **Frequency** – how often transactions occur

- Easy features:
  - How many transactions total
  - How many transactions per group
  - Ratio of frequency by group to days active

Online Account Access Example

Online Account Access Example

?

TRANSFORMING CATEGORIES

Categorical Data

Physical characteristics

⋮

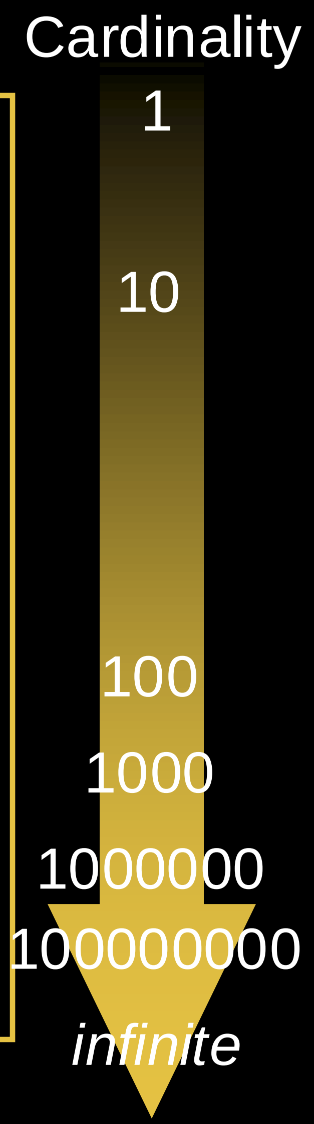
Education level

⋮

State

⋮

Postal code
Address
Social security number
Free-form text



Categorical Data

Physical characteristics

Education level

State

Postal code
Address
Social security number
Free-form text

Cardinality

1
10
100
1000
1000000
100000000
infinite

Number of elements in a set.

Strategies

Physical characteristics
⋮
Education level
⋮
State
⋮
Postal code
Address
Social security number
Free-form text

Cardinality

1
10
100
1000
1000000
100000000
infinite

Strategy

None / Model Based
Recoding
Transformations
Linking
Text Mining

Dummy Coding

X
D
B
C
C
A
A
D
C
A
⋮

D_A	D_B	D_C	D_D
0	0	0	1
0	1	0	0
0	0	1	0
0	0	1	0
1	0	0	0
1	0	0	0
0	0	0	1
0	0	1	0
1	0	0	0
⋮	⋮	⋮	⋮

Thresholding

Level	N_i
A	1562
B	970
C	223
D	111
E	85
F	50
G	23
H	17
I	12
J	5

Thresholding

Level	N_i
A	1562
B	970
C	223
D	111
E	85
F	50
G	23
H	17
I	12
J	5

Recombine to single new level, OTHER.

Target-Based Enumeration

Level	N_i	ΣY_i	p_i
A	1562	430	0.28
B	970	432	0.45
C	223	45	0.20
D	111	36	0.32
E	85	23	0.27
F	50	20	0.40
G	23	8	0.35
H	17	5	0.29
I	12	6	0.50
J	5	5	1.00

Target-Based Enumeration

Level	N_i	ΣY_i	p_i
J	5	5	1.00
I	12	6	0.50
B	970	432	0.45
F	50	20	0.40
G	23	8	0.35
D	111	36	0.32
H	17	5	0.29
A	1562	430	0.28
E	85	23	0.27
C	223	45	0.20

Target-Based Enumeration

X	N_i	ΣY_i	p_i
1	5	5	1.00
2	12	6	0.50
3	970	432	0.45
4	50	20	0.40
5	23	8	0.35
6	111	36	0.32
7	17	5	0.29
8	1562	430	0.28
9	85	23	0.27
10	223	45	0.20

New Ordinal Input

Weight of Evidence

Level	N_i	ΣY_i	p_i	WoE
J	5	5	1.00	.
I	12	6	0.50	-0.71
B	970	432	0.45	-0.49
F	50	20	0.40	-0.30
G	23	8	0.35	-0.08
D	111	36	0.32	0.03
H	17	5	0.29	0.17
A	1562	430	0.28	0.26
E	85	23	0.27	0.28
C	223	45	0.20	0.67

Old Categorical Input

New Numeric Input

Geocoding

ZIP	Longitude	Latitude
02713	-70.8017	41.45222
02840	-71.3114	41.49438
04848	-68.9096	44.30417
08739	-74.0549	40.02756
10927	-73.9604	41.19228
10960	-73.9187	41.08947
13640	-75.9098	44.33451
14555	-76.9867	43.27016
19939	-75.2052	38.57527
19944	-75.0509	38.46811
20004	-77.0275	38.89254

Transform zip code to location.

Derived Fields Specific to Insurance

- Approximate the distance from the claimant's address to the adjuster's location using only zip codes.

Zipcode ⇨ (Latitude,Longitude)

[(Lat_1,Long_1),(Lat_2,Long_2)] ⇨ Distance

⇨

(Claimant Zipcode, Adjuster Zipcode) ⇨ Distance

Derived Fields from Text

- Text mining can provide an immense amount of data when limited data may seem to exist.
- Mining the text data may reveal patterns that can be adapted into input variables.

NETWORK FEATURES

Occupational View

- “Everything is a nail to a kid with a hammer.”

- The view of the world around us is influenced by our experiences:
  - Economist: World is a supply/demand curve.
  - Chemist: World is a set of chemical equations.
  - Statistician: World is a collection of observations with dependent and independent variables.

Society – Statisticians

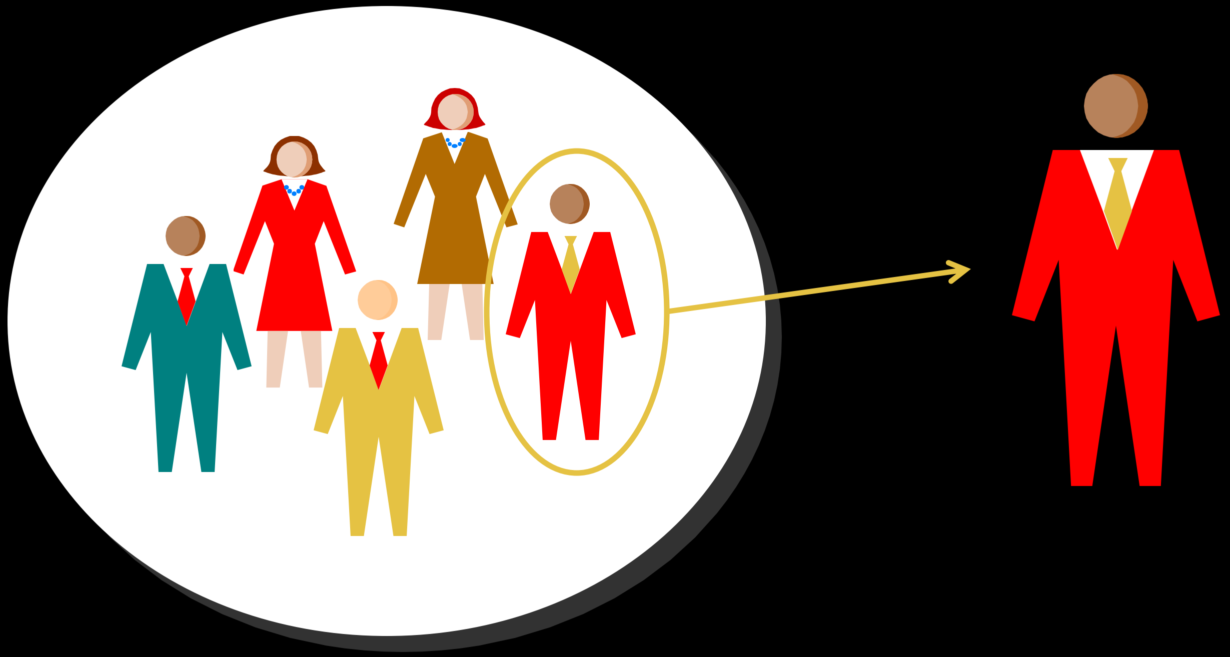
- Statisticians typically view society as a collection of individuals who have distinct, measureable characteristics.

Society – Statisticians

- Statisticians typically view society as a collection of individuals who have distinct, measureable characteristics.

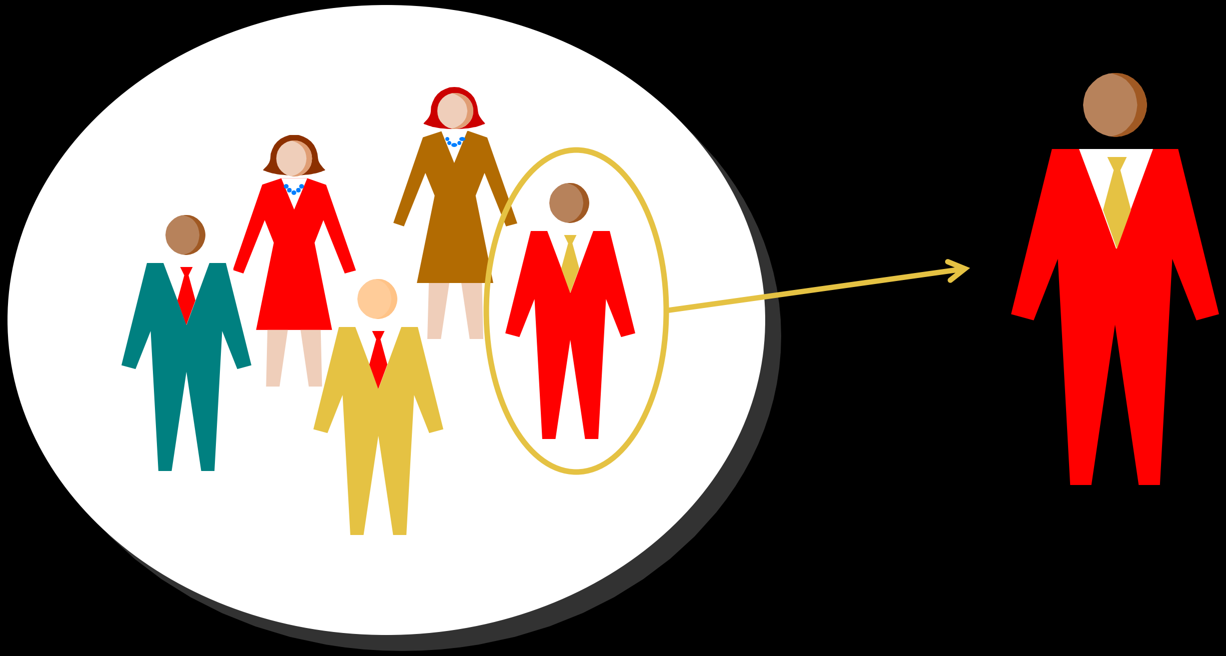
Society – Statisticians

- Statisticians typically view society as a collection of individuals who have distinct, measureable characteristics.



Society – Statisticians

- Statisticians typically view society as a collection of individuals who have distinct, measureable characteristics.



Characteristics:
- Age
- Weight
- Income
- Education
- Etc.

Statisticians' Data Structure

- Data structure is typically rectangular in nature.

Name	Age	Weight	Income	Years of College Education
Bill	54	190	$48,000	4
Tina	26	135	$95,000	4
Larry	39	215	$101,000	9
…	…	…	…	…

Statisticians' Data Structure

- Data structure is typically rectangular in nature.

Comparing Individuals

Name	Age	Weight	Income	Years of College Education
Bill	54	190	$48,000	4
Tina	26	135	$95,000	4
Larry	39	215	$101,000	9
…	…	…	…	…

Statisticians' Data Structure

- Data structure is typically rectangular in nature.

Comparing Variables

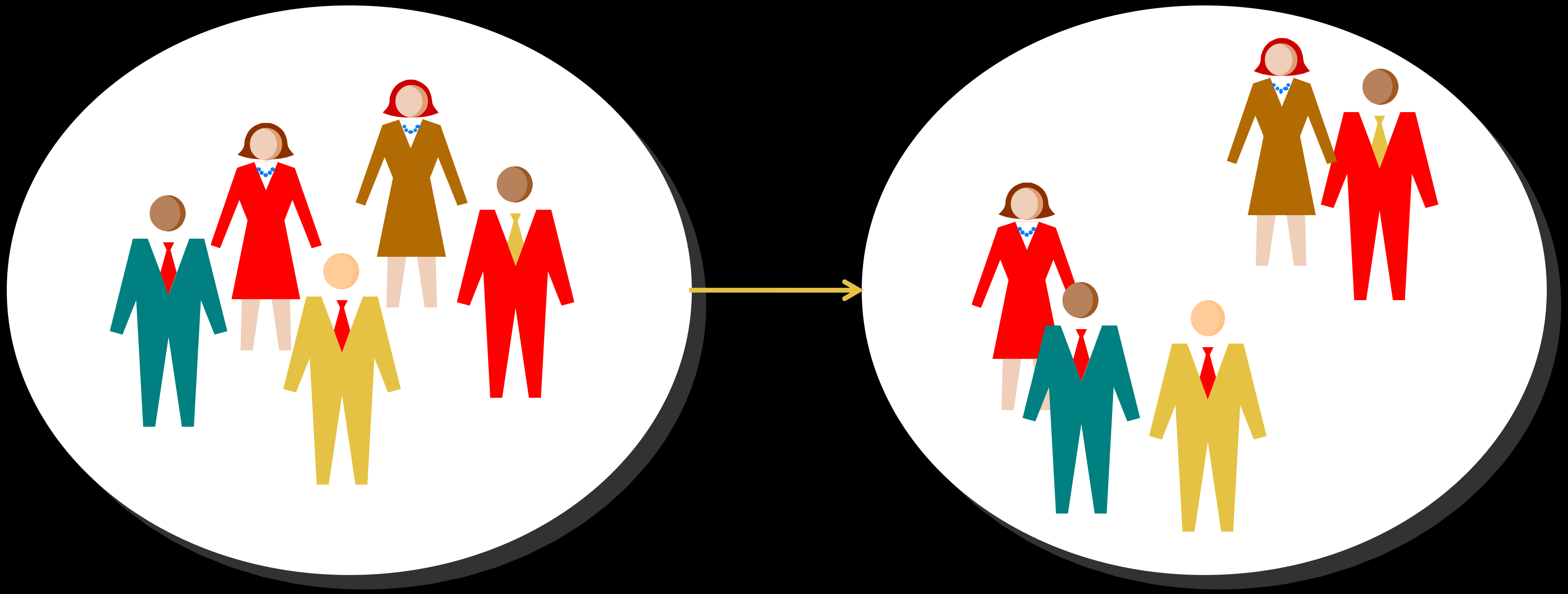
Name	Age	Weight	Income	Years of College Education
Bill	54	190	$48,000	4
Tina	26	135	$95,000	4
Larry	39	215	$101,000	9
…	…	…	…	…

Society – Sociometrists

- J L Moreno founded a social science called **sociometry**, where **sociometrists** believe that society is made up of individuals **and** their social, economic, or cultural ties.

Society – Sociometrists

- J L Moreno founded a social science called **sociometry**, where **sociometrists** believe that society is made up of individuals **and** their social, economic, or cultural ties.



Society – Sociometrists

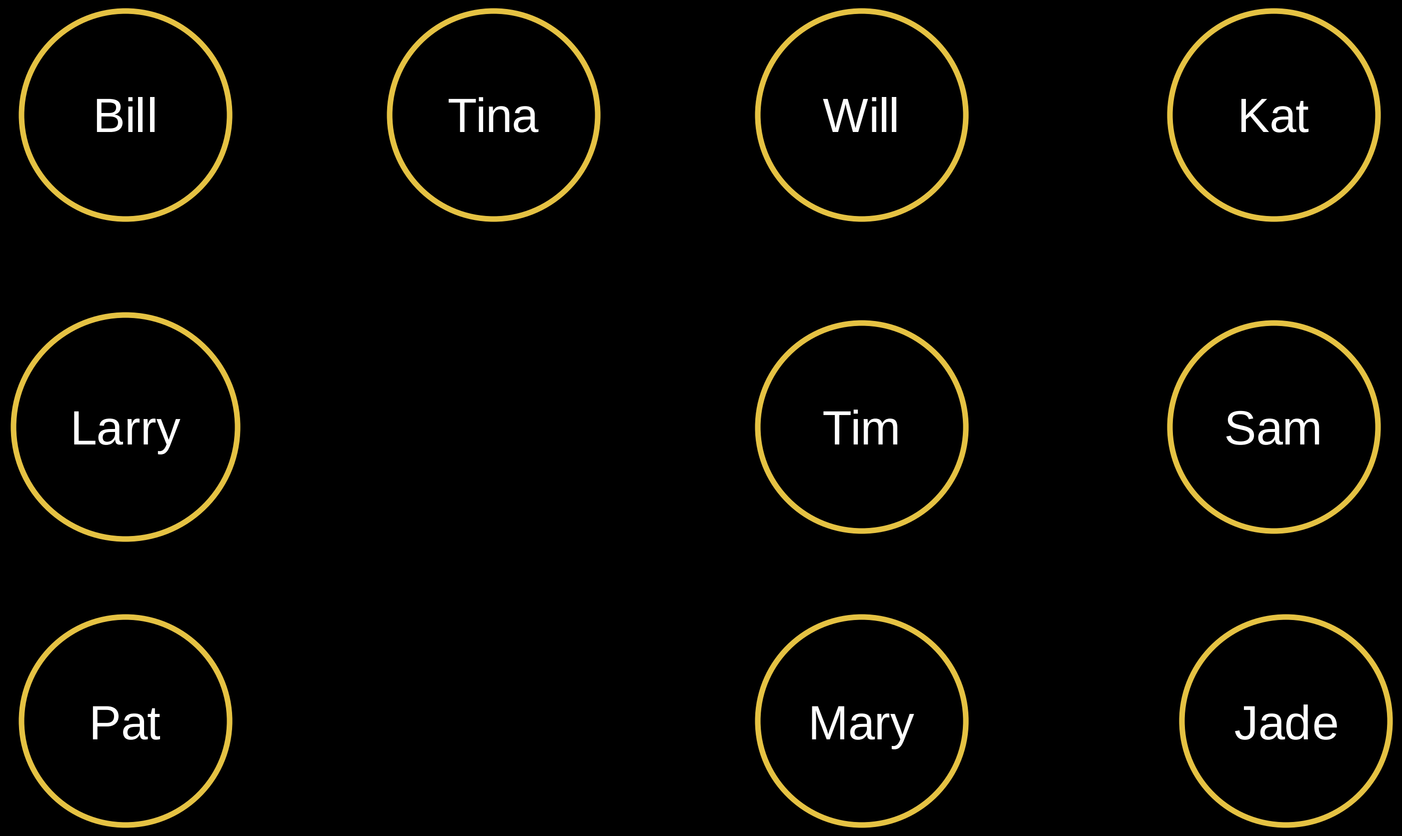
- J L Moreno founded a social science called **sociometry**, where **sociometrists** believe that society is made up of individuals **and** their social, economic, or cultural ties.

- The importance is not only on the individual’s characteristics, but also on the patterns of an individual’s interactions with other individuals.
- The interactions themselves are just as important as who the individual connects to.

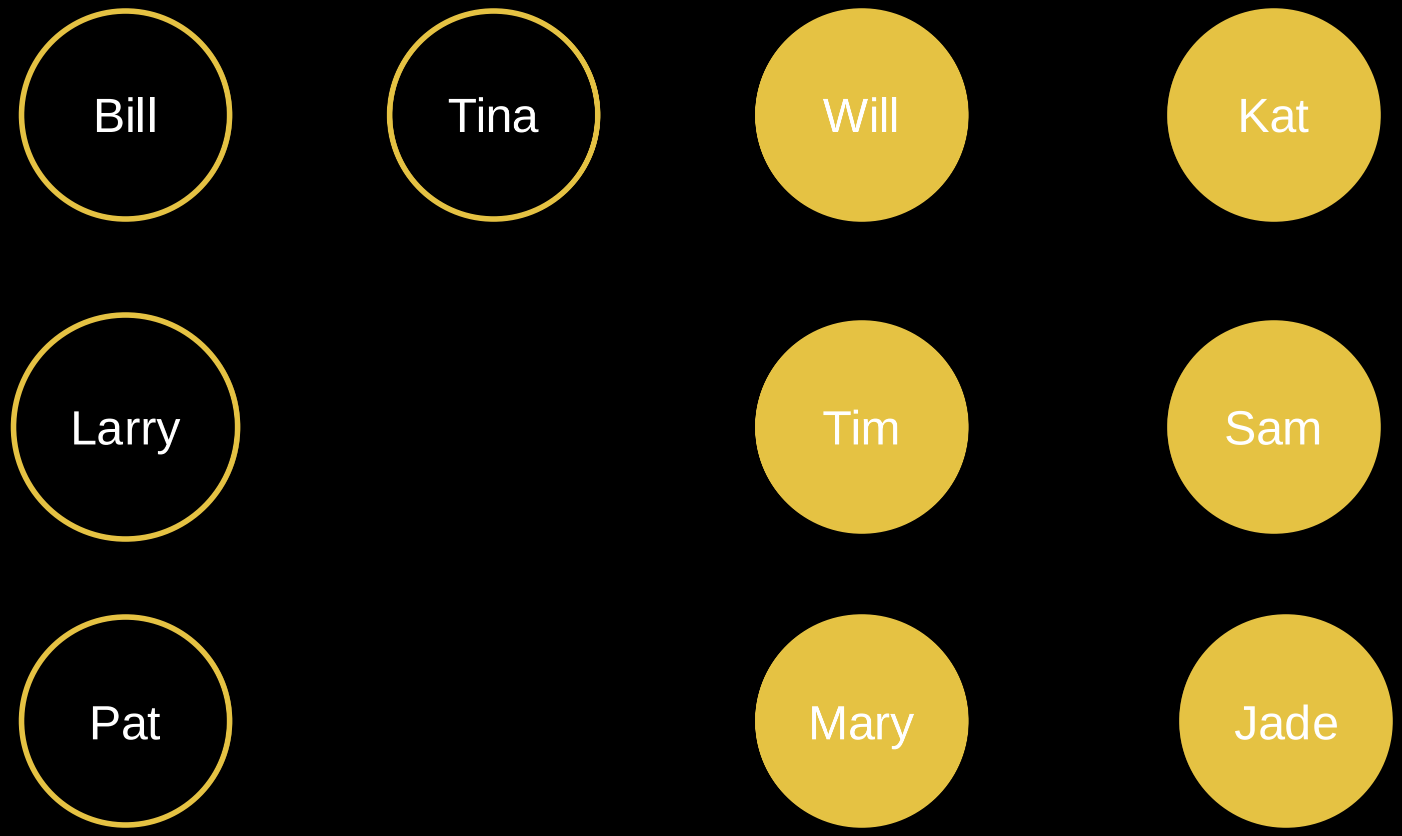
Exploring Social Networks

- Sociometrists use **graph networks (link graphs)** to visualize social networks.
- These graph networks reveal a structure to the data that can not be seen by basic summary statistics.
- Each of the circles are referred to as **nodes**.

Exploring Social Networks



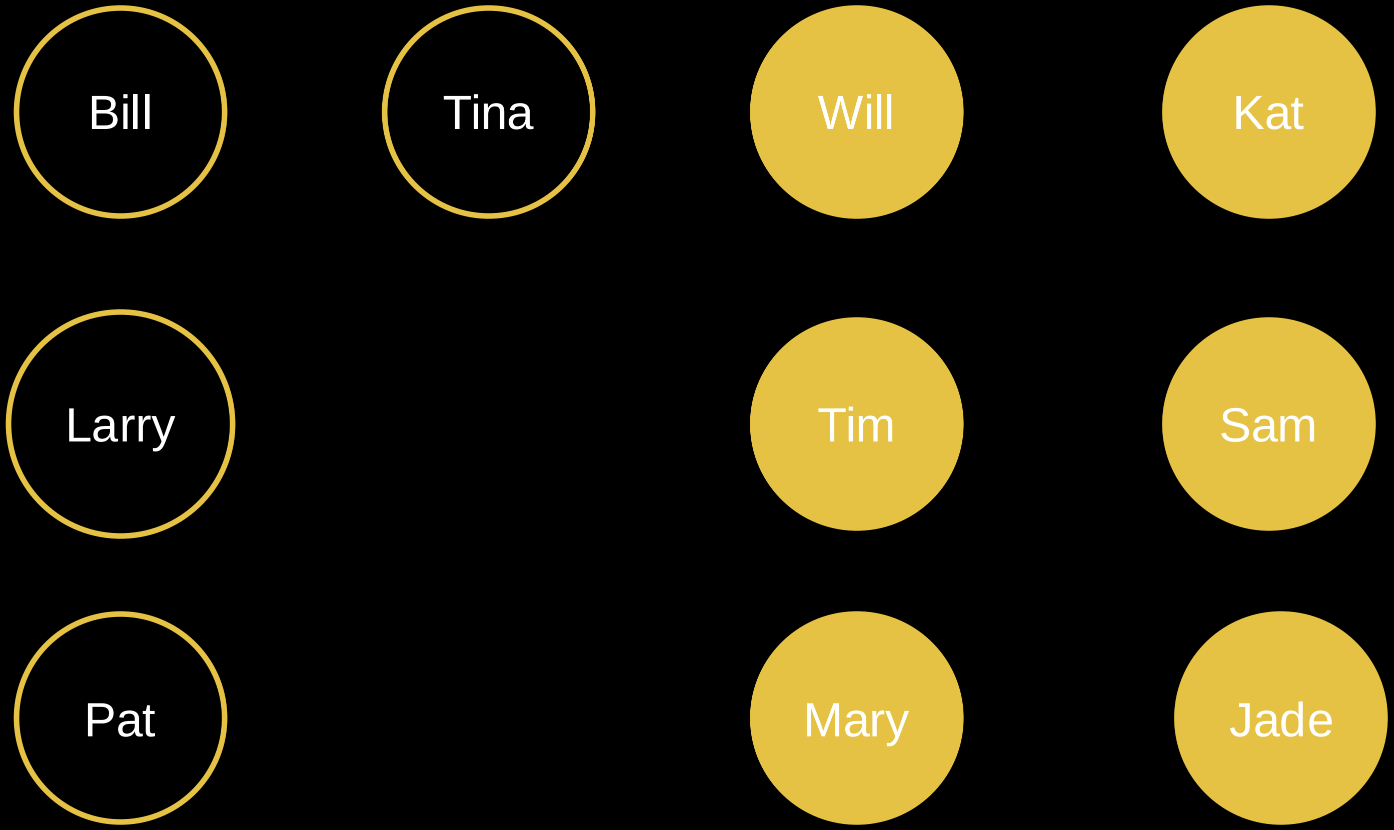
Exploring Social Networks



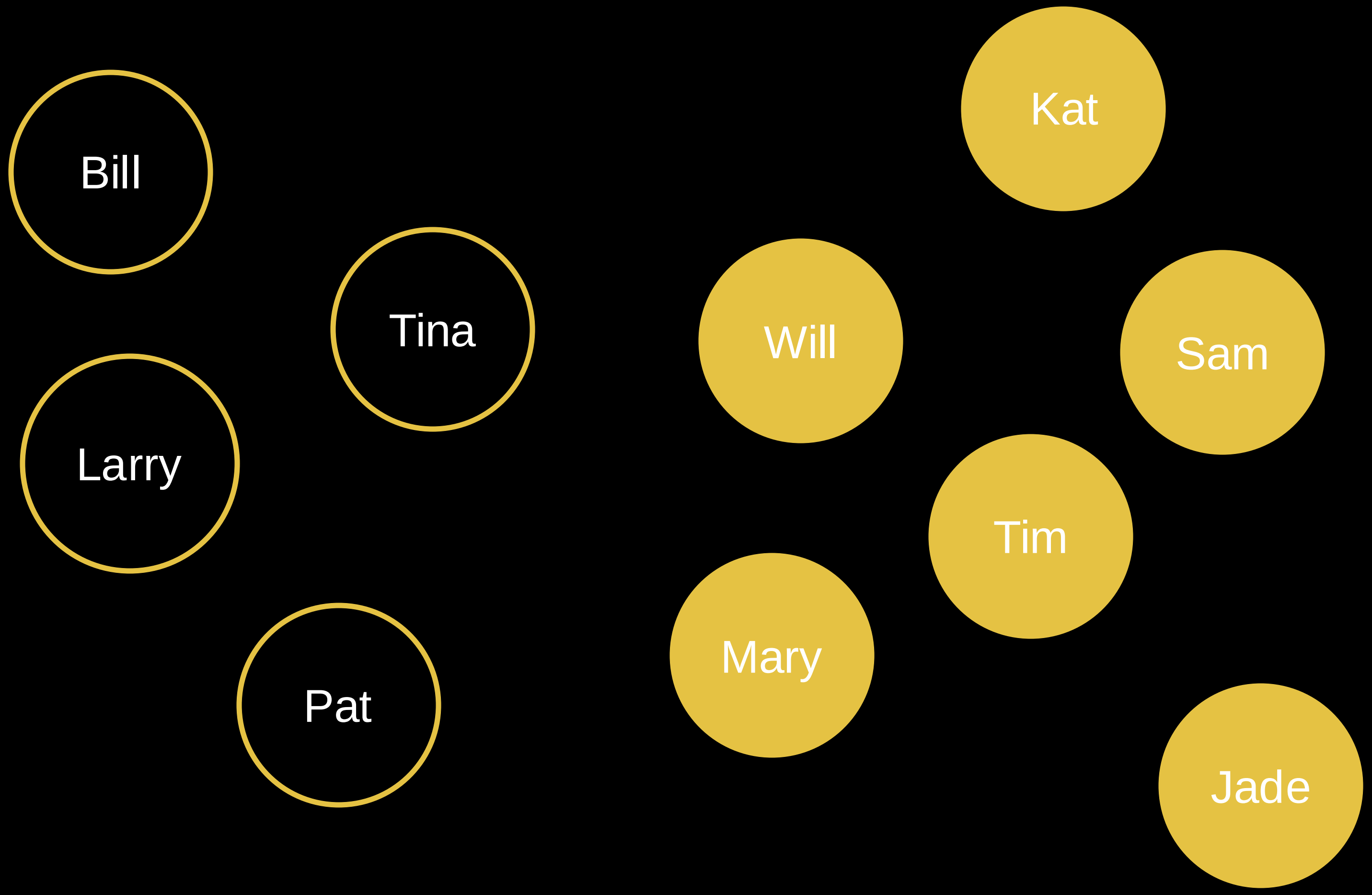
Exploring Social Networks

- Sociometrists use **graph networks** (**link graphs**) to visualize social networks.
- These sociograms reveal a structure to the data that can not be seen by basic summary statistics.
- Each of the circles are referred to as **nodes**.
- Each node could be connected by **links**.

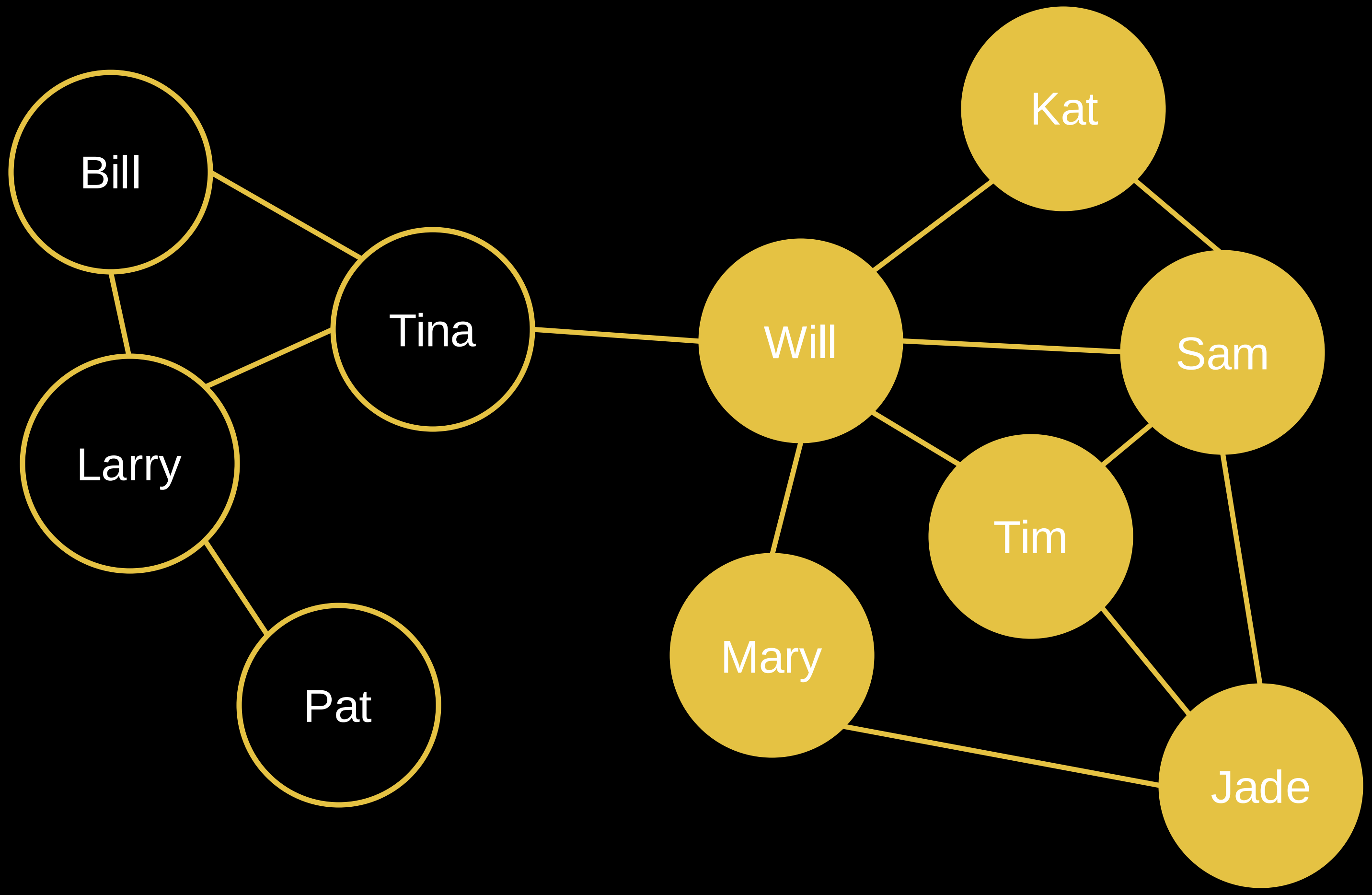
Exploring Social Networks



Exploring Social Networks



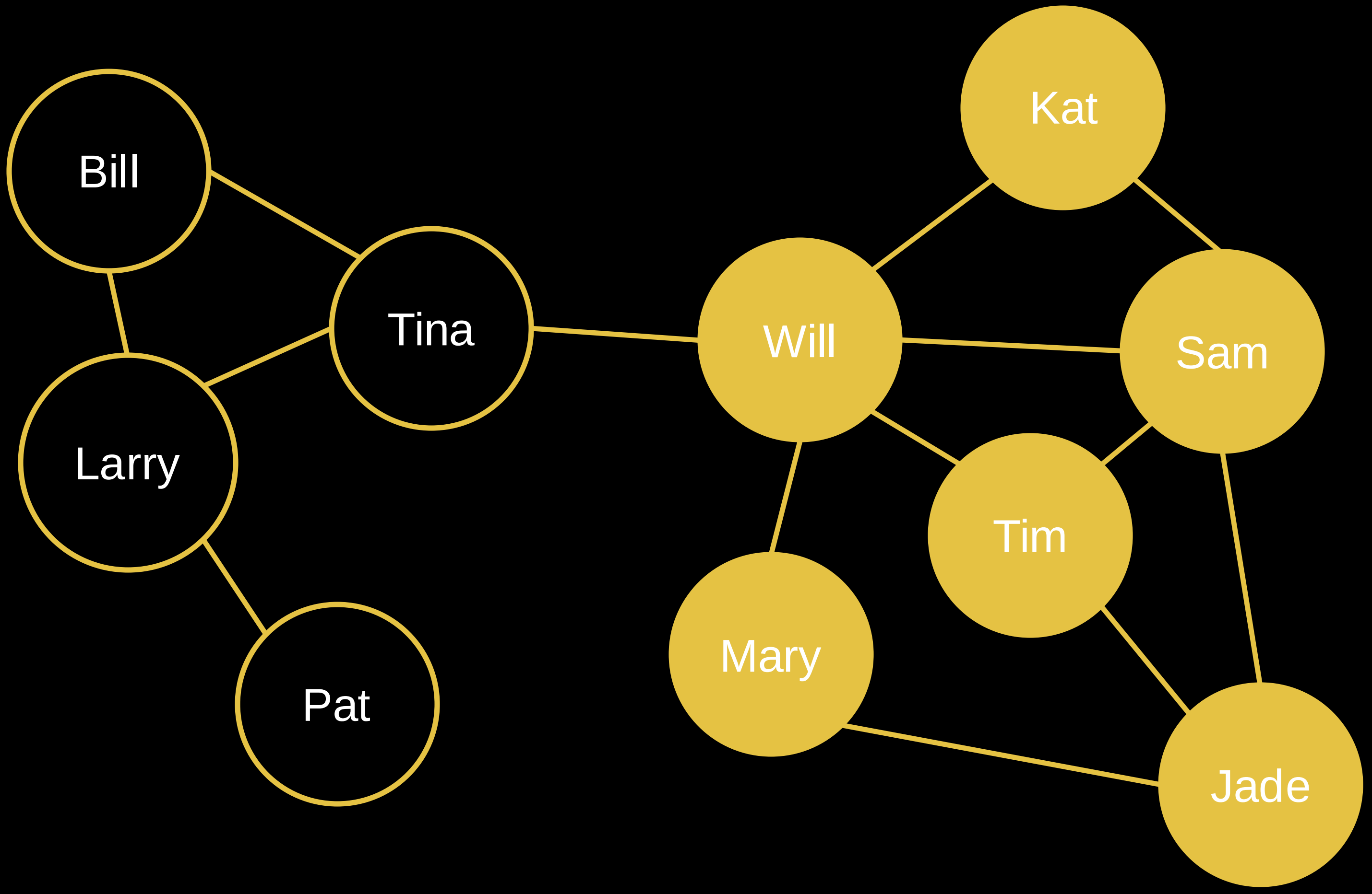
Exploring Social Networks



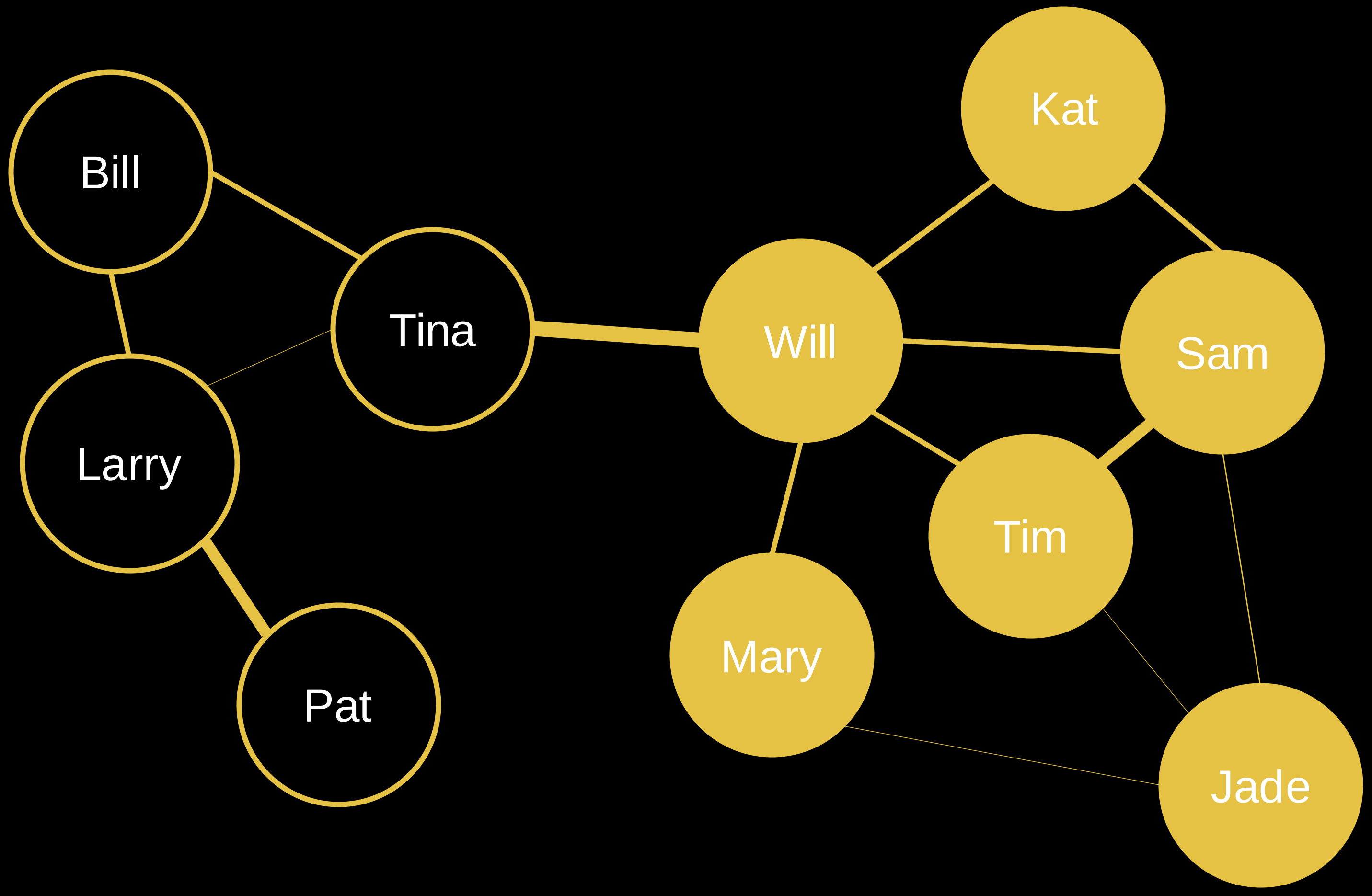
Exploring Social Networks

- Sociometrists use **graph networks** (**link graphs**) to visualize social networks.
- These sociograms reveal a structure to the data that can not be seen by basic summary statistics.
- Each of the circles are referred to as **nodes**.
- Each node could be connected by **links**.
  - Links can be of different sizes to summarize **strength** of connection.

Exploring Social Networks



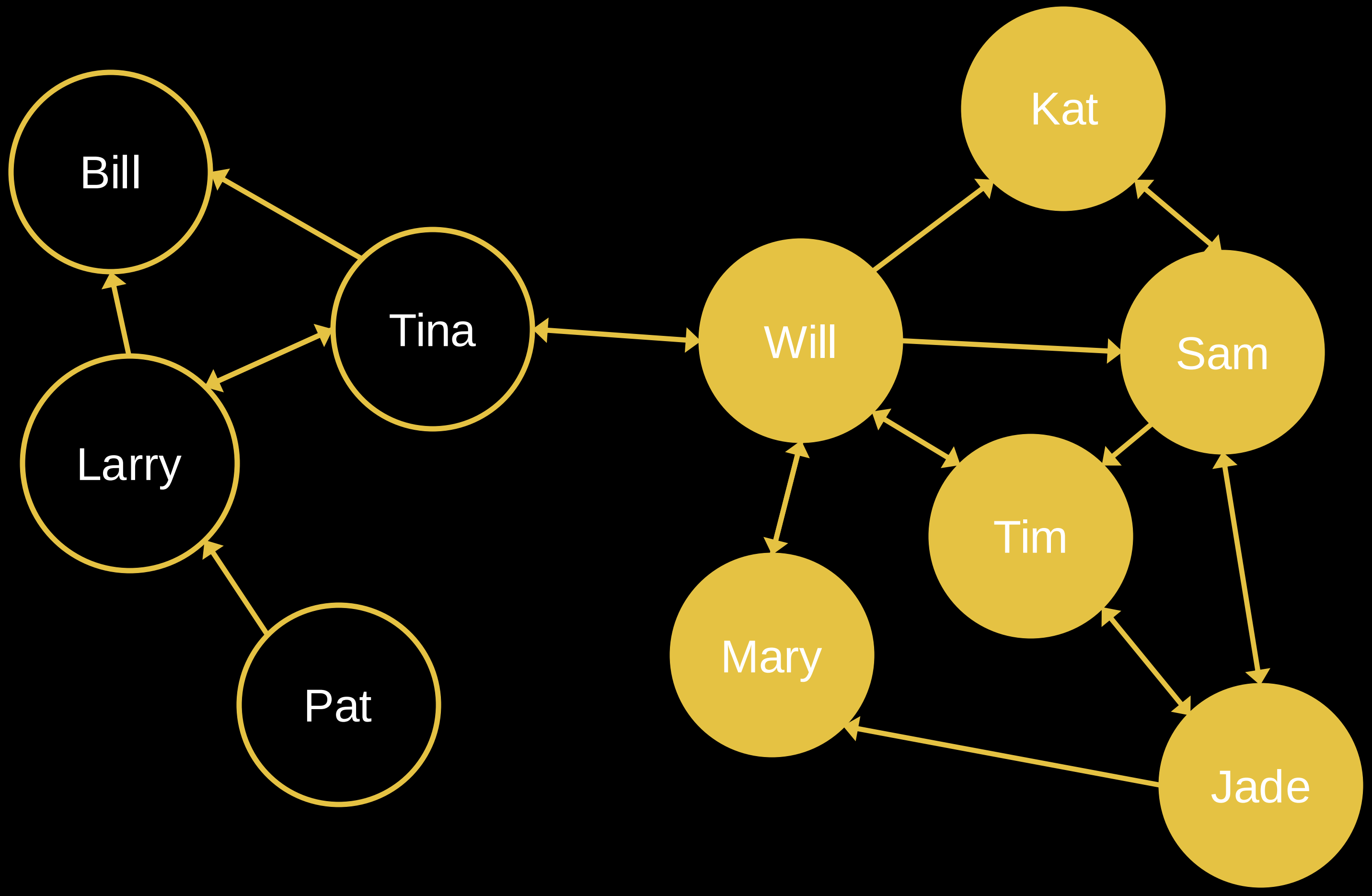
Exploring Social Networks



Exploring Social Networks

- Sociometrists use **graph networks** (**link graphs**) to visualize social networks.
- These sociograms reveal a structure to the data that can not be seen by basic summary statistics.
- Each of the circles are referred to as **nodes**.
- Each node could be connected by **links**.
  - Links can be of different sizes to summarize **strength** of connection.
  - **Reciprocity** can also be represented by links.

Exploring Social Networks



Graph Network Data Structure

- Data structure is typically square in nature.

Who Reports Liking Whom?				
	Choice:			
Chooser:	**Bill**	**Tina**	**Larry**	**…**
Bill	–	1	1	…
Tina	0	–	1	…
Larry	0	0	–	…
…	…	…	…	…

Graph Network Data Structure

- Data structure doesn't have to be limited to binary.

How Does Someone Know Someone (0 = Don't Know, 1 = Work, 2 = Family)				
	Mark	**Anthony**	**April**	**Tim**
Mark	–	1	0	2
Anthony	1	–	2	0
April	1	2	–	1
Tim	2	0	1	–

Graph Network Data Structure

- Other differences:
  - **No** independence of observations
  - Samples are rarely desired – try for population of a known network
  - Individuals don’t only have to be linked through other individuals
    - Example – schools in a school district

Modern Adaptations

- Several problems have been addressed by these methods:
  - Spread of disease
  - Marketing of products
  - Fraud detection
- There are also popular cultural themes that have arisen from these methods:
  - Facebook
  - “Six degrees of separation”
  - The Oracle of Bacon

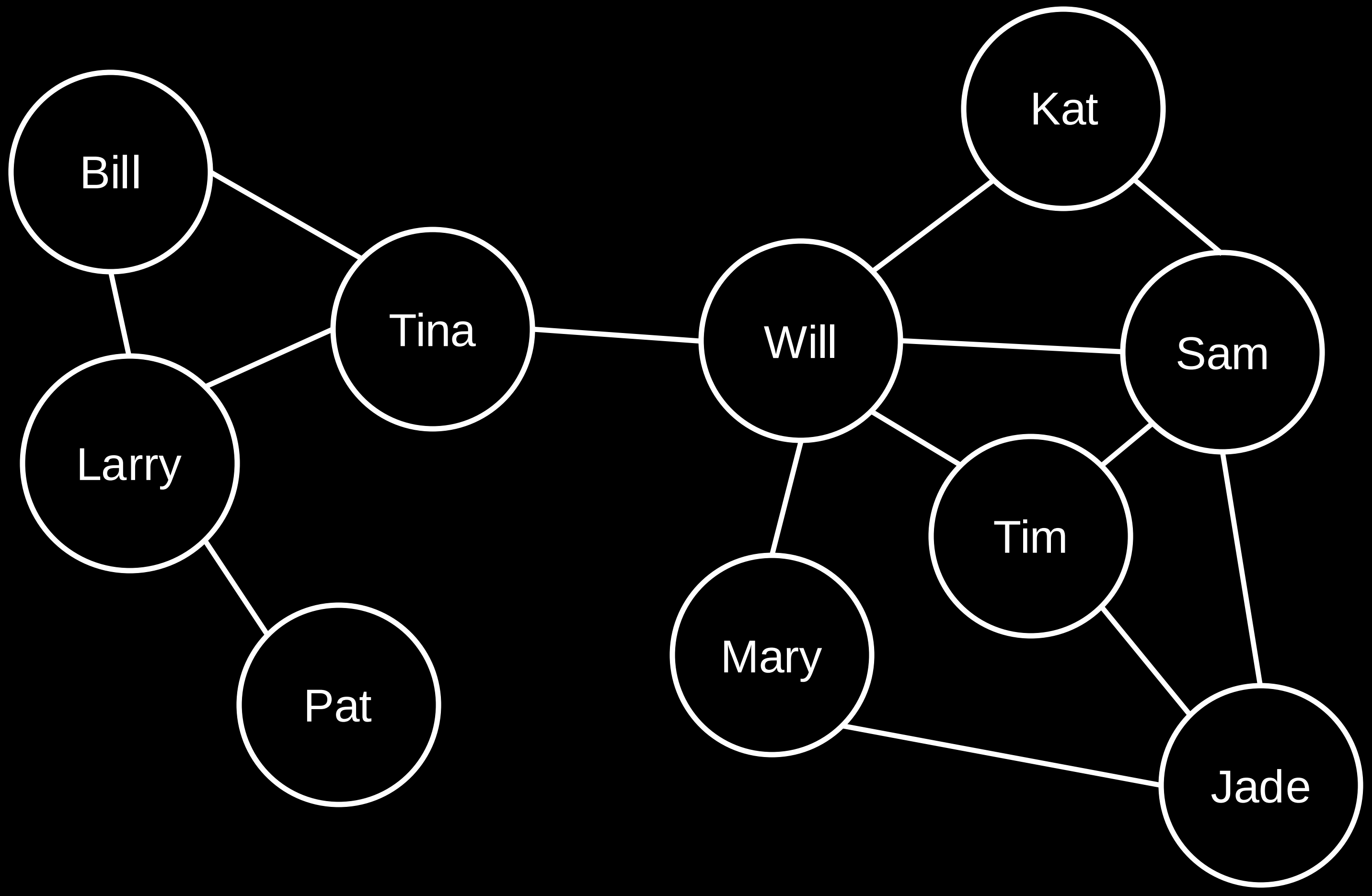
Social Structure

- There are many different summaries and important calculations obtained from sociograms.
- Here are a few we will focus on:
  - Subgroups
  - Centers and Closeness
  - Brokers and Bridges
  - Diffusion and Adoption

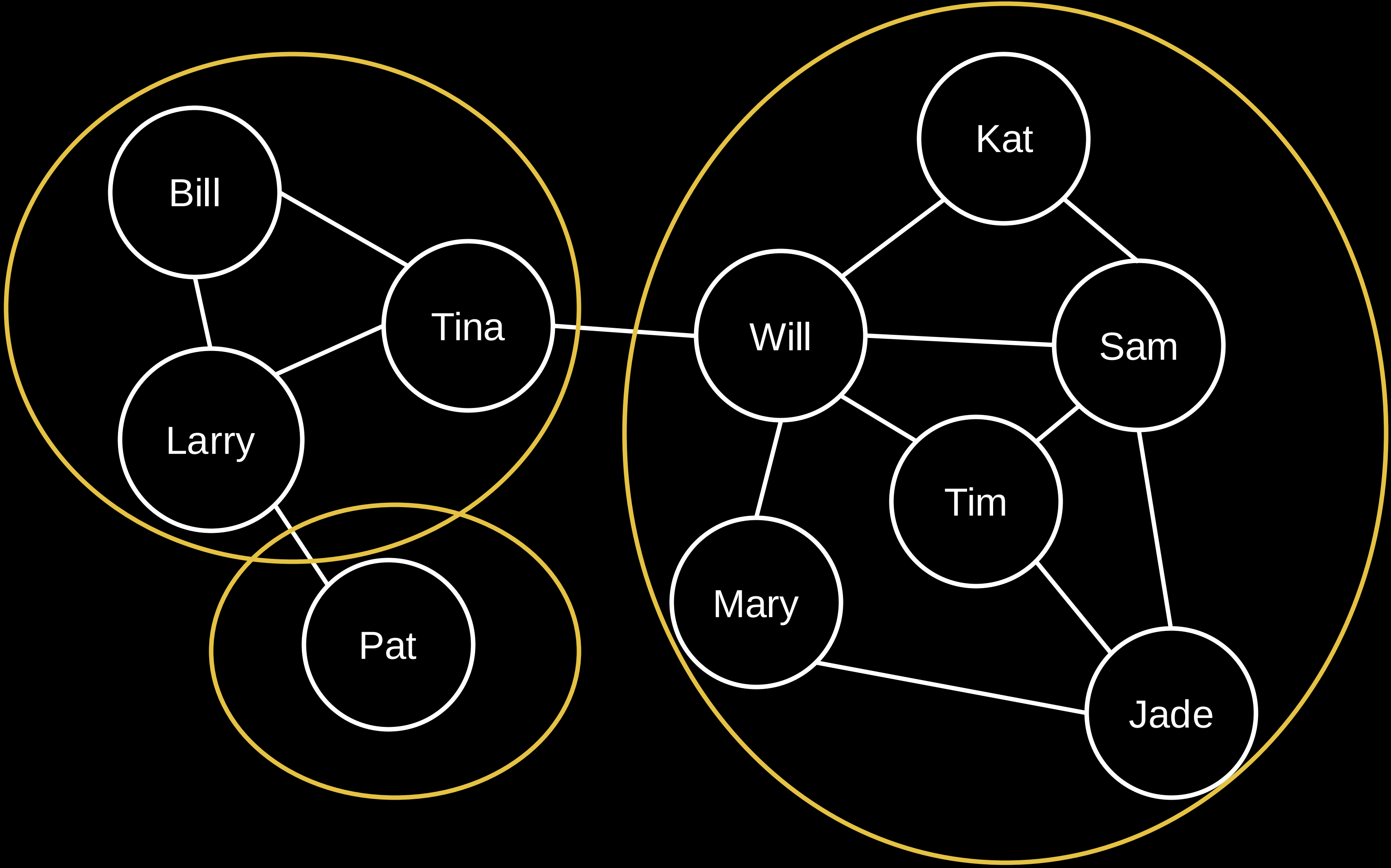
Subgroups

- Social networks typically contain dense pockets of individuals.
- These dense pockets are sometimes called **subgroups**.
- If a subgroup is completely separated from the rest of the network, then it is a **cohesive subgroup**.
- Homophily: “Birds of a feather flock together.”
- This can help in the identification of individuals with similar characteristics.
  - Marketing campaigns
  - Fraud detection

Subgroups



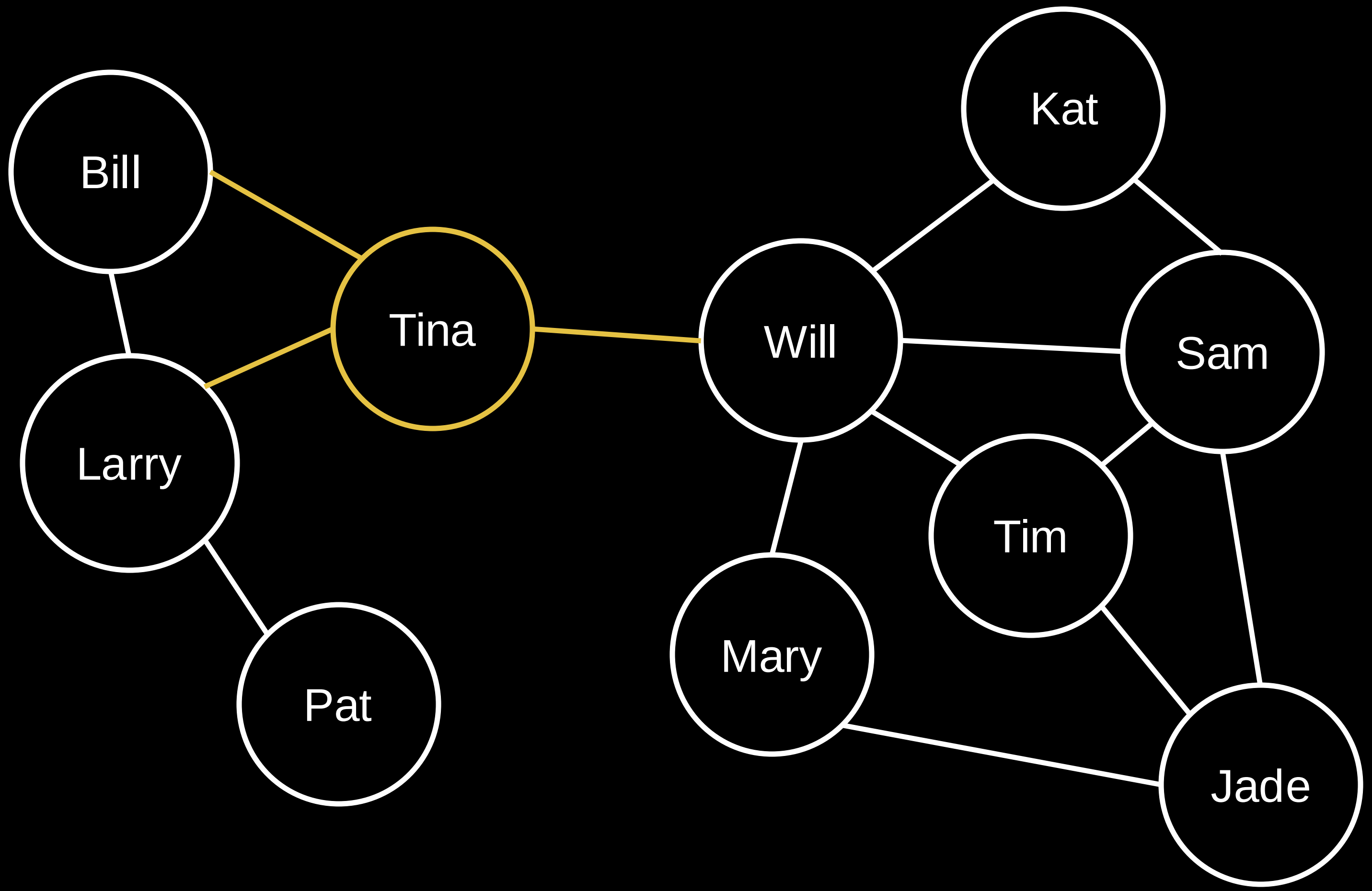
Subgroups



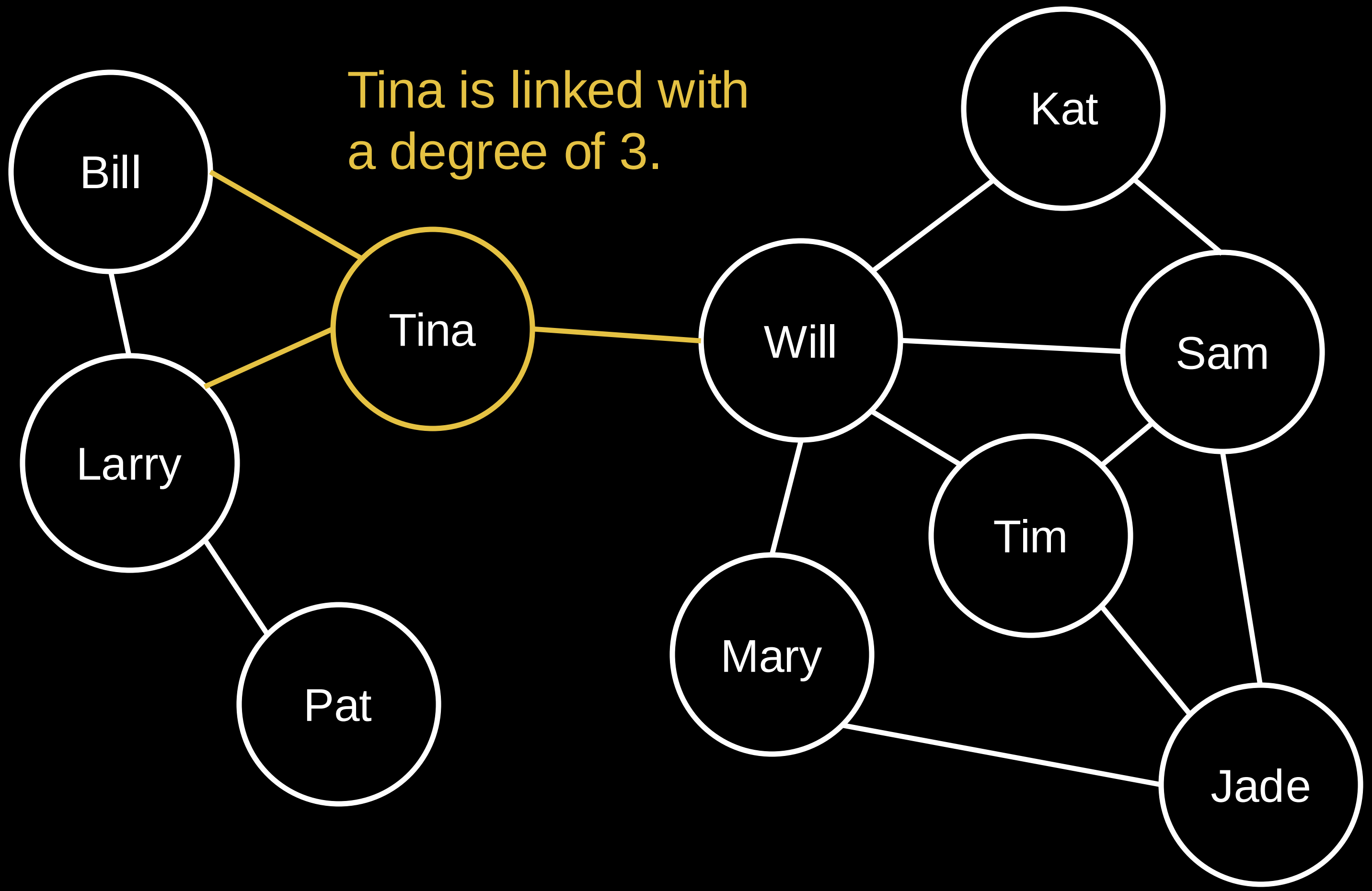
Social Structure

- There are many different summarizes and important calculations obtained from sociograms.
- Here are a few we will focus on:
  - Subgroups
  - Centers and Closeness
  - Brokers and Bridges
  - Diffusion and Adoption

Degree of Connection

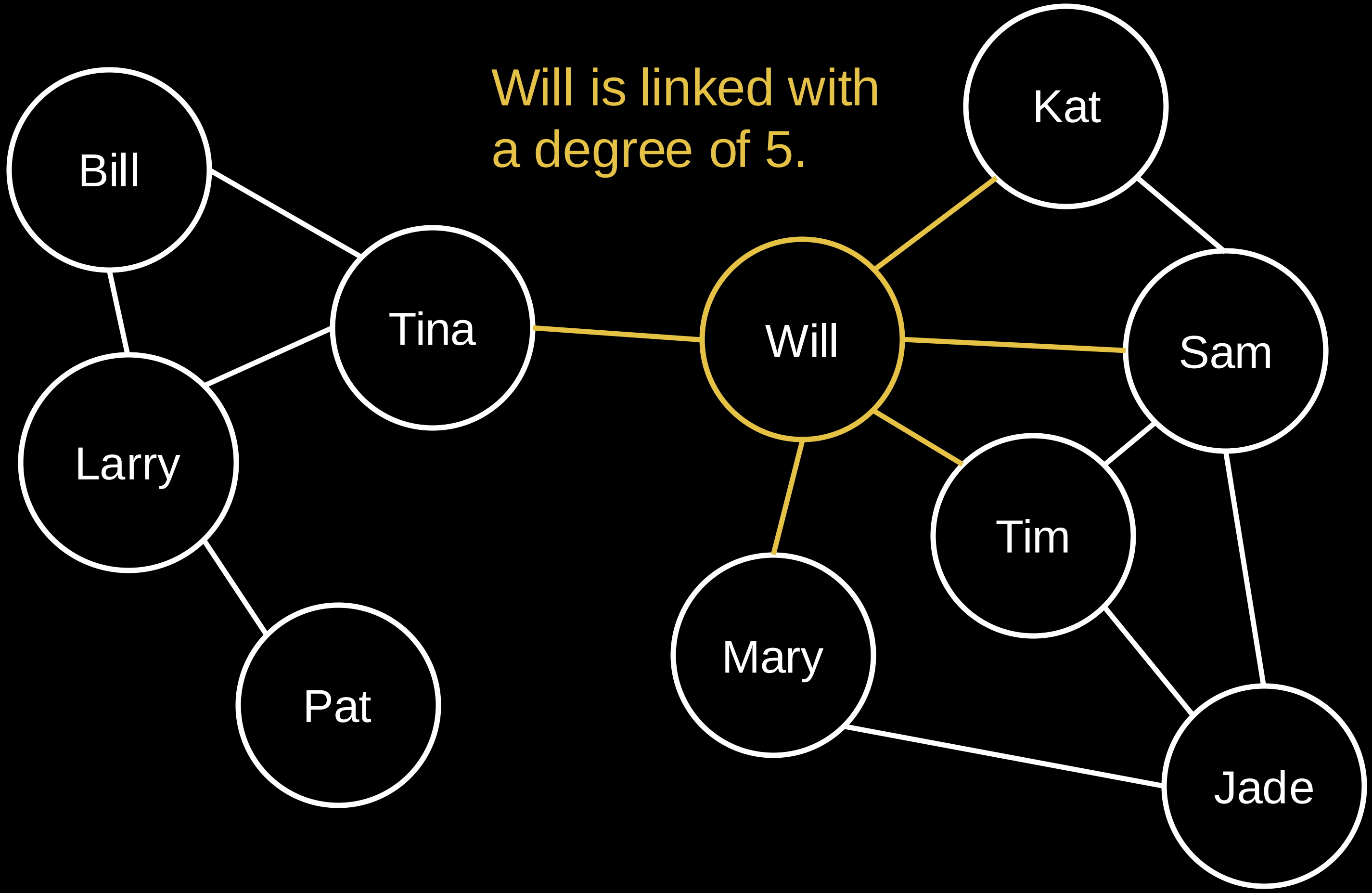


Degree of Connection



Degree of Connection

Will is linked with a degree of 5.



Degree of Connection

Name	Degree of Connection
Will	5
Sam	4
Jade	3
Tim	3
Tina	3
Larry	3
Mary	2
Kat	2
Bill	2
Pat	1

Degree Centrality

- Networks consist of N nodes and n links.
- The maximum degree of each node is N-1.
- Degree centrality ”standardizes” the degree of a node.

$$C_D = \frac{degree}{N-1}$$

Degree Centrality

Name	Degree of Connection	Degree Centrality
Will	5	0.555
Sam	4	0.444
Jade	3	0.333
Tim	3	0.333
Tina	3	0.333
Larry	3	0.333
Mary	2	0.222
Kat	2	0.222
Bill	2	0.222
Pat	1	0.111

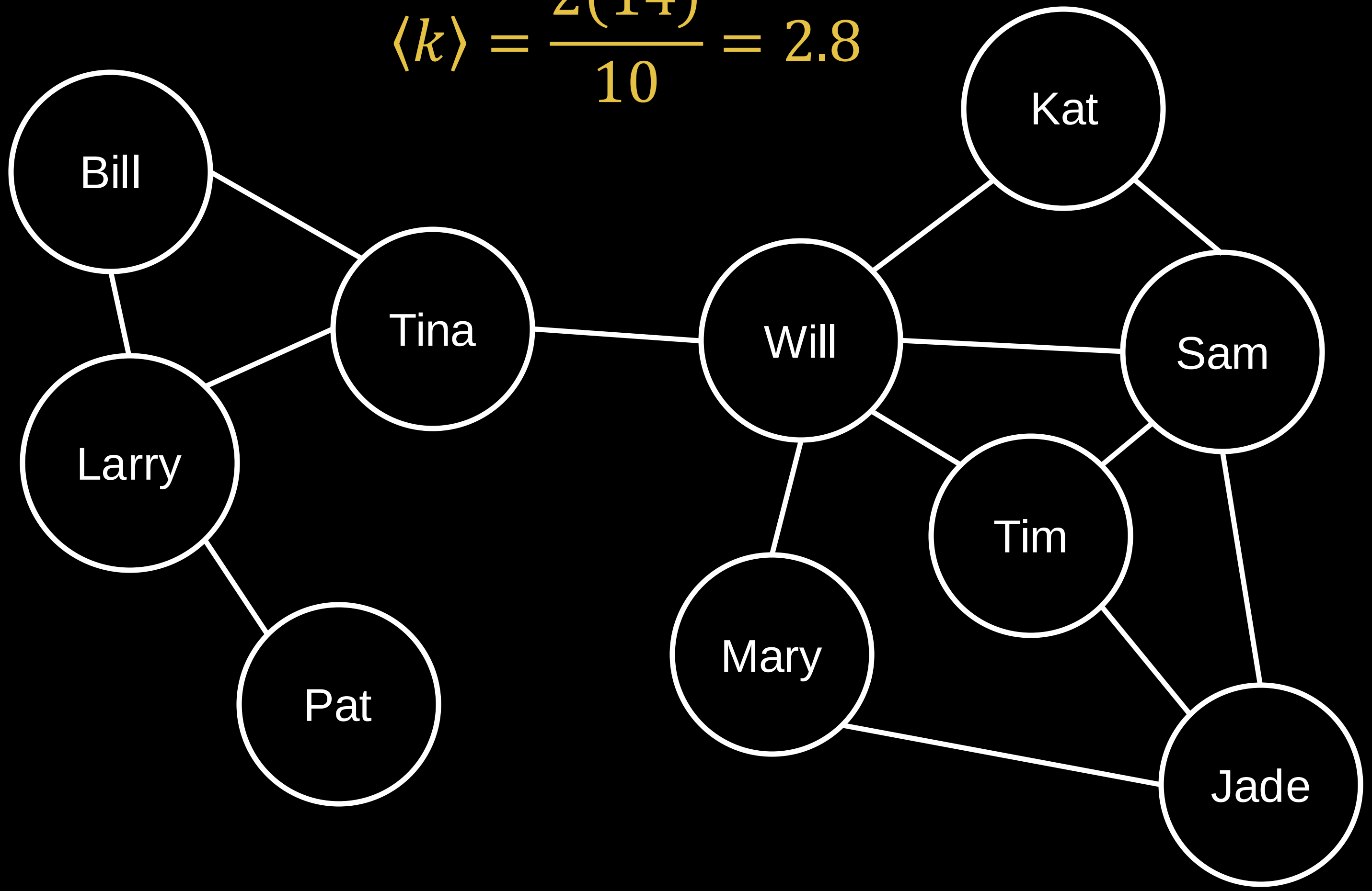
Average Degree of Graph

- Networks consist of N nodes and n links.
- The average degree of the graph, $\langle k \rangle$, is the following:

$$\langle k \rangle = \frac{2n}{N}$$

Average Degree of Graph

$$\langle k \rangle = \frac{2(14)}{10} = 2.8$$



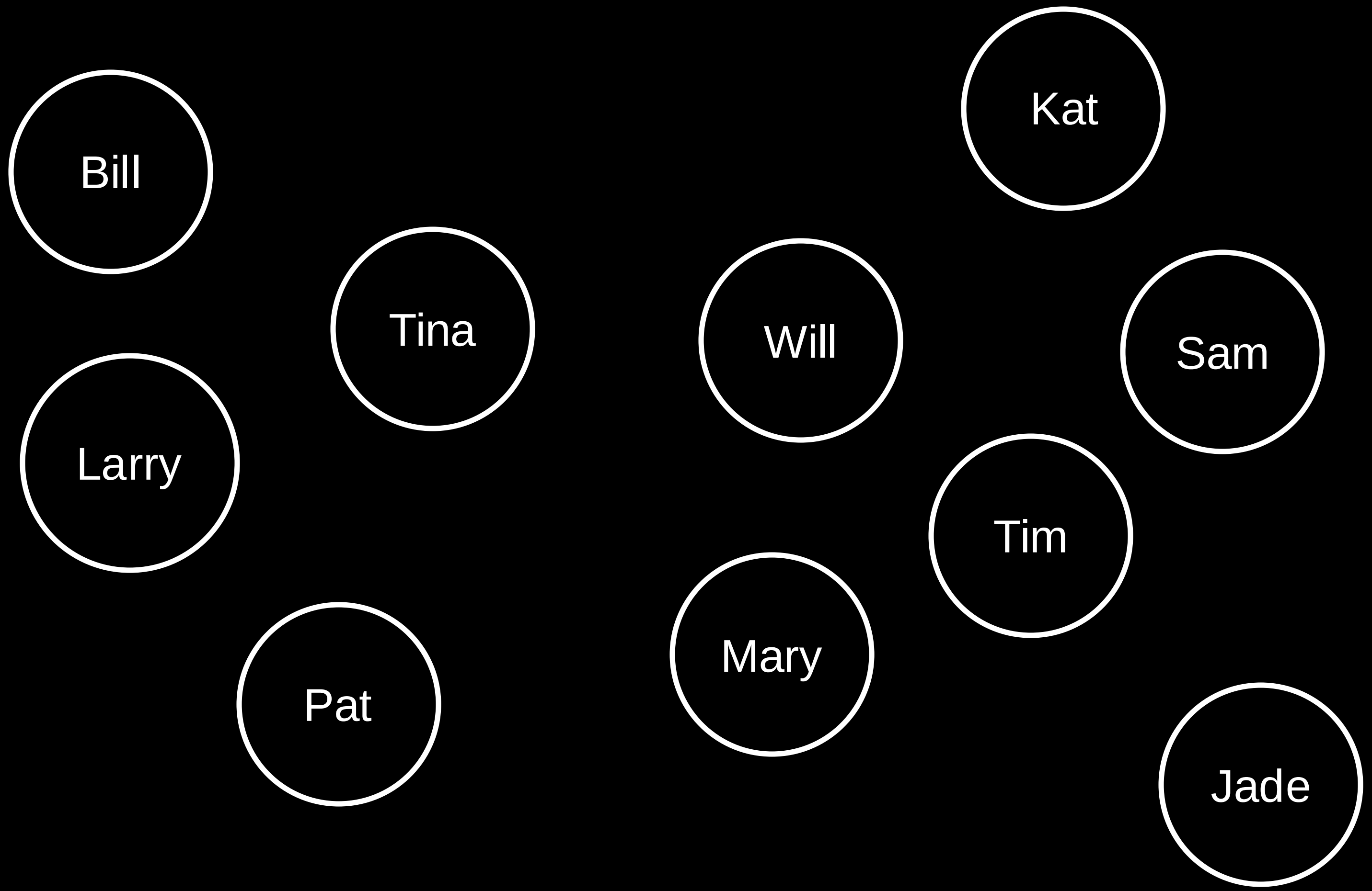
Density of Graph

- Networks consist of N nodes and n links.
- The density of the graph is the proportion of the number of links actually in the graph compared to the maximum number of links possible in the graph.
- The density of the graph, Δ, is the following:

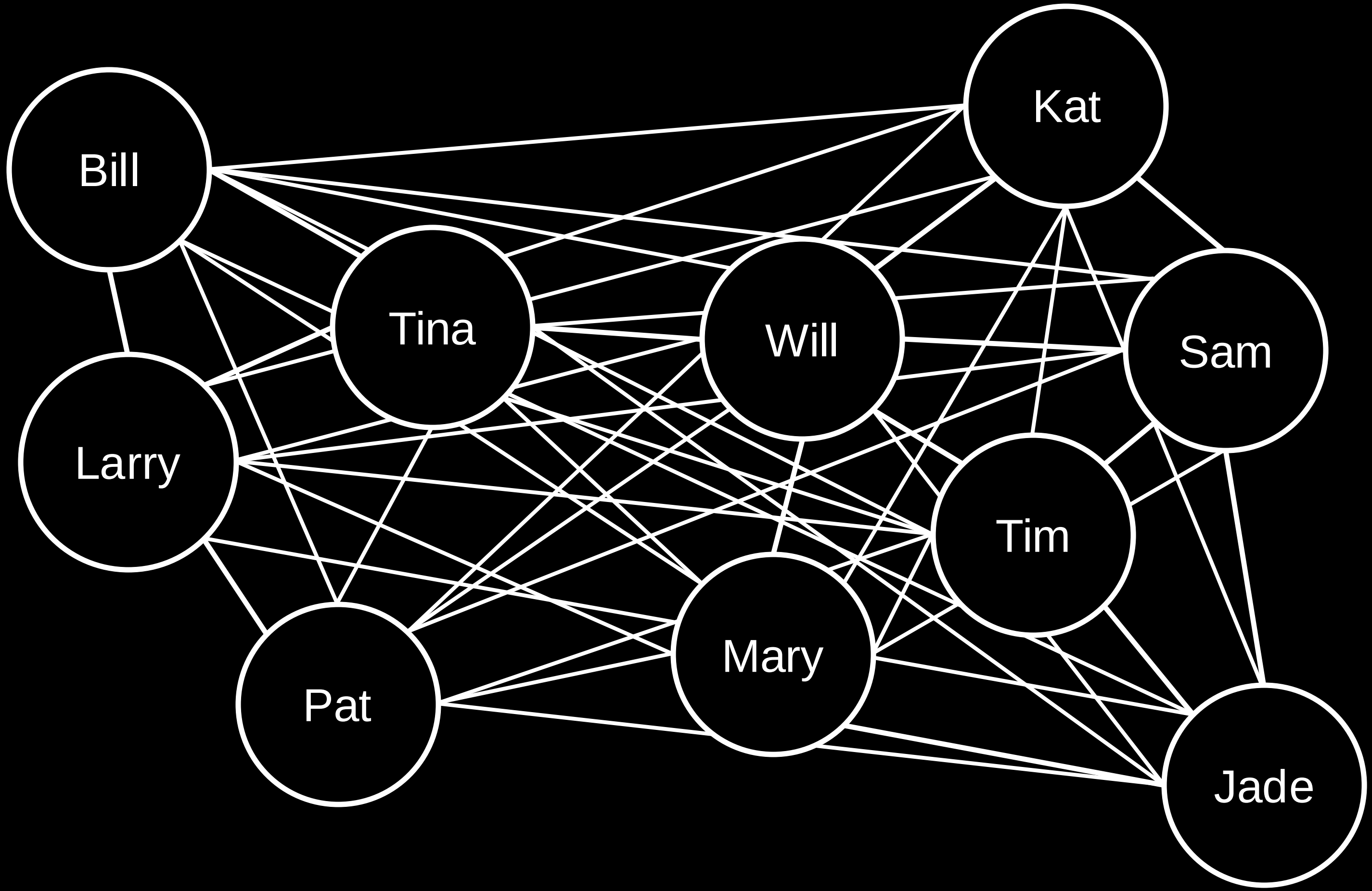
$$\Delta = \frac{2n}{N(N-1)}$$

- This is also called the **connection probability**.

Density of Graph

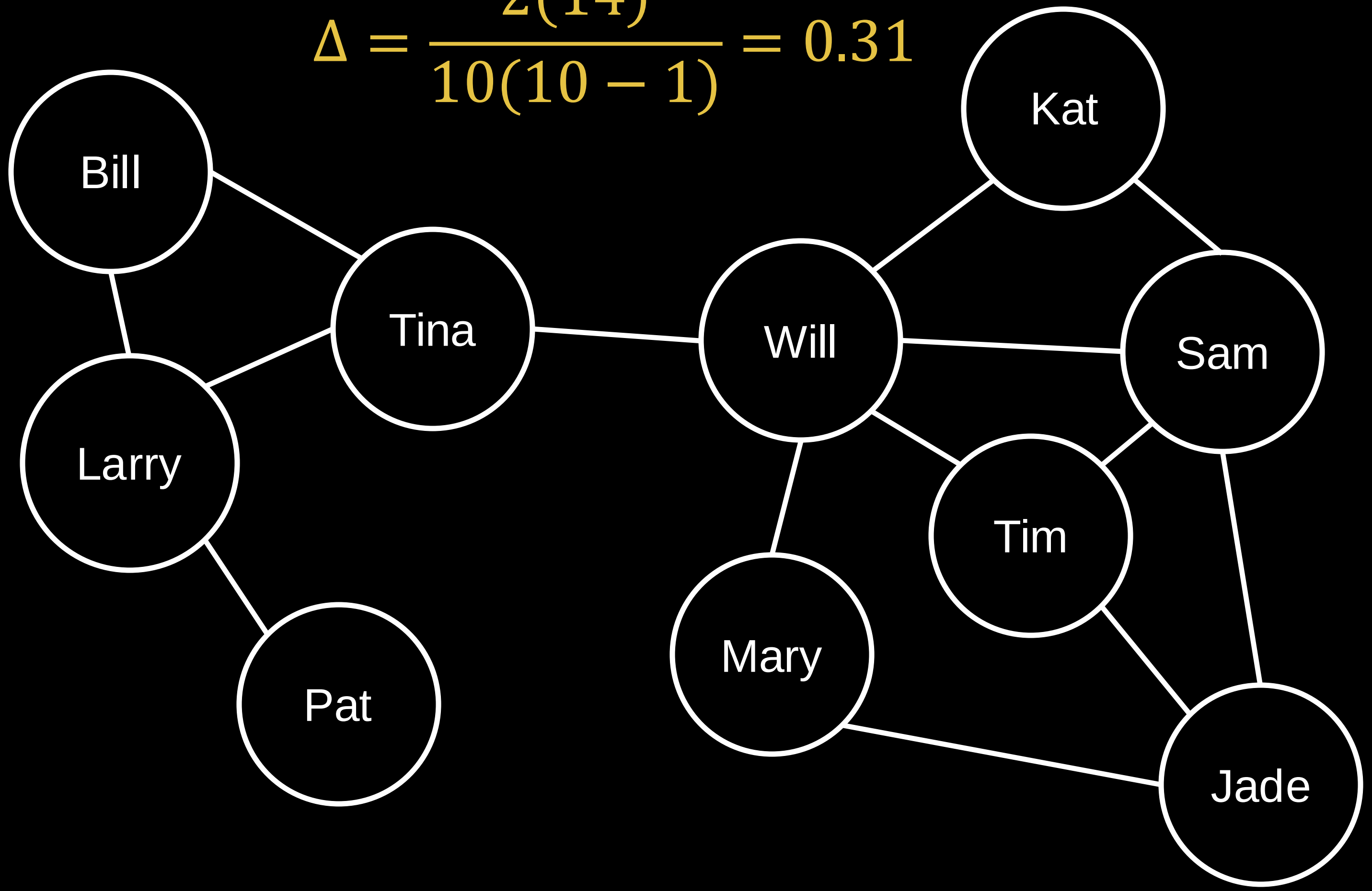


Density of Graph



Density of Graph

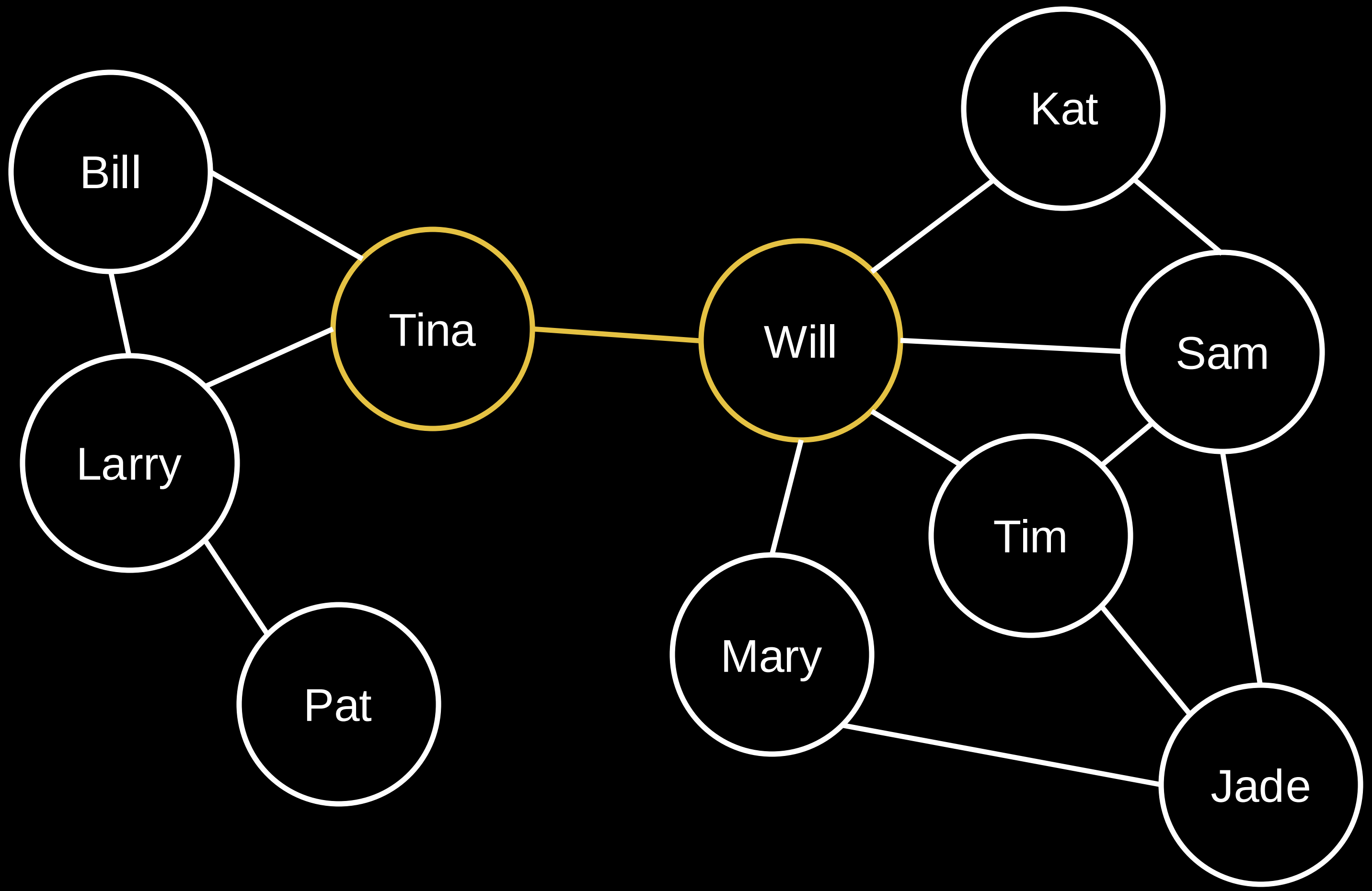
$$\Delta = \frac{2(14)}{10(10-1)} = 0.31$$



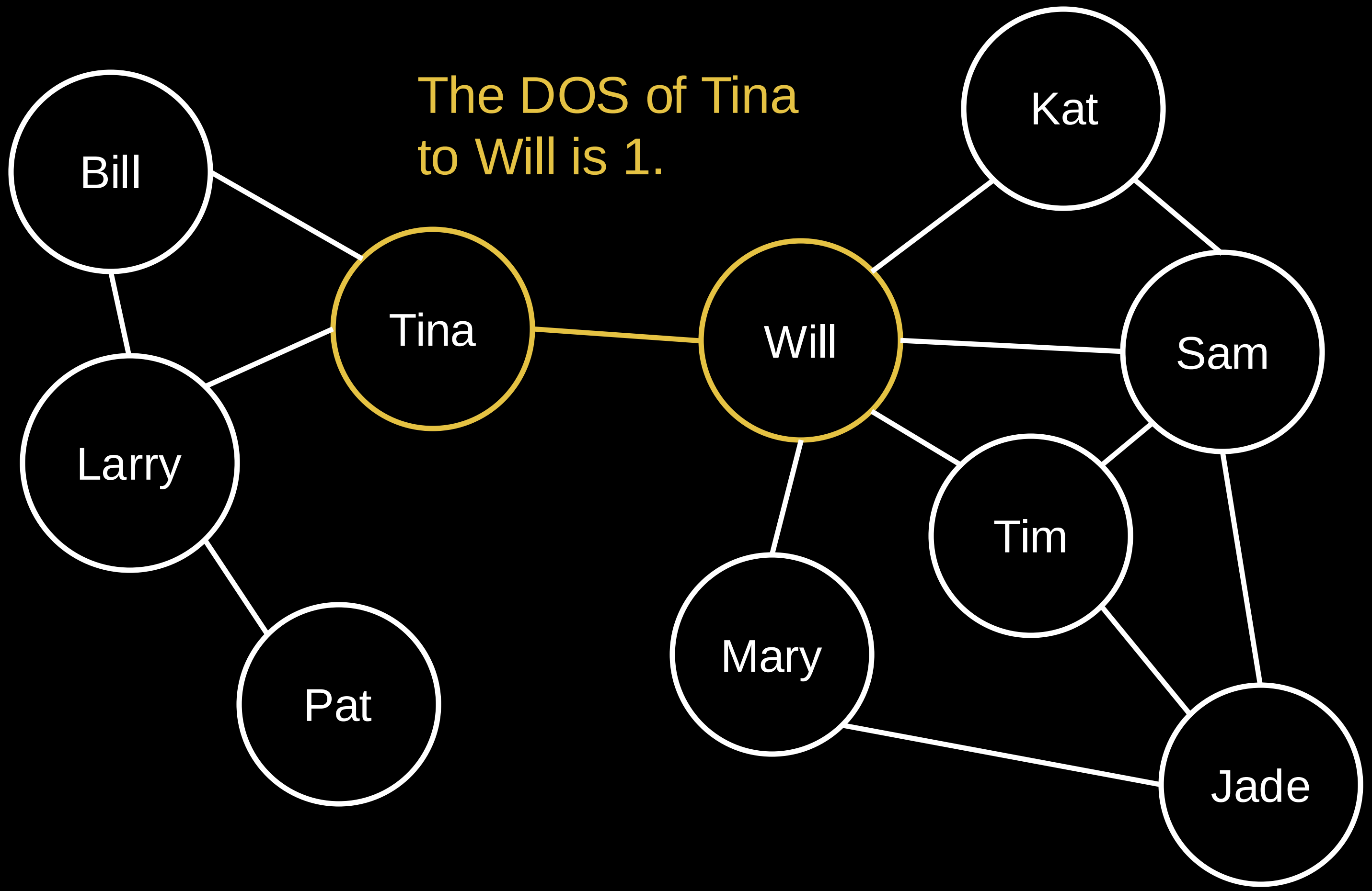
Degree of Separation

- The degree of connection is one way to measure the center of a network.
- The degree of separation is another way to measure center.
- The degree of connection only focuses on the links for a certain individual, while degree of separation focuses on the value of those links.

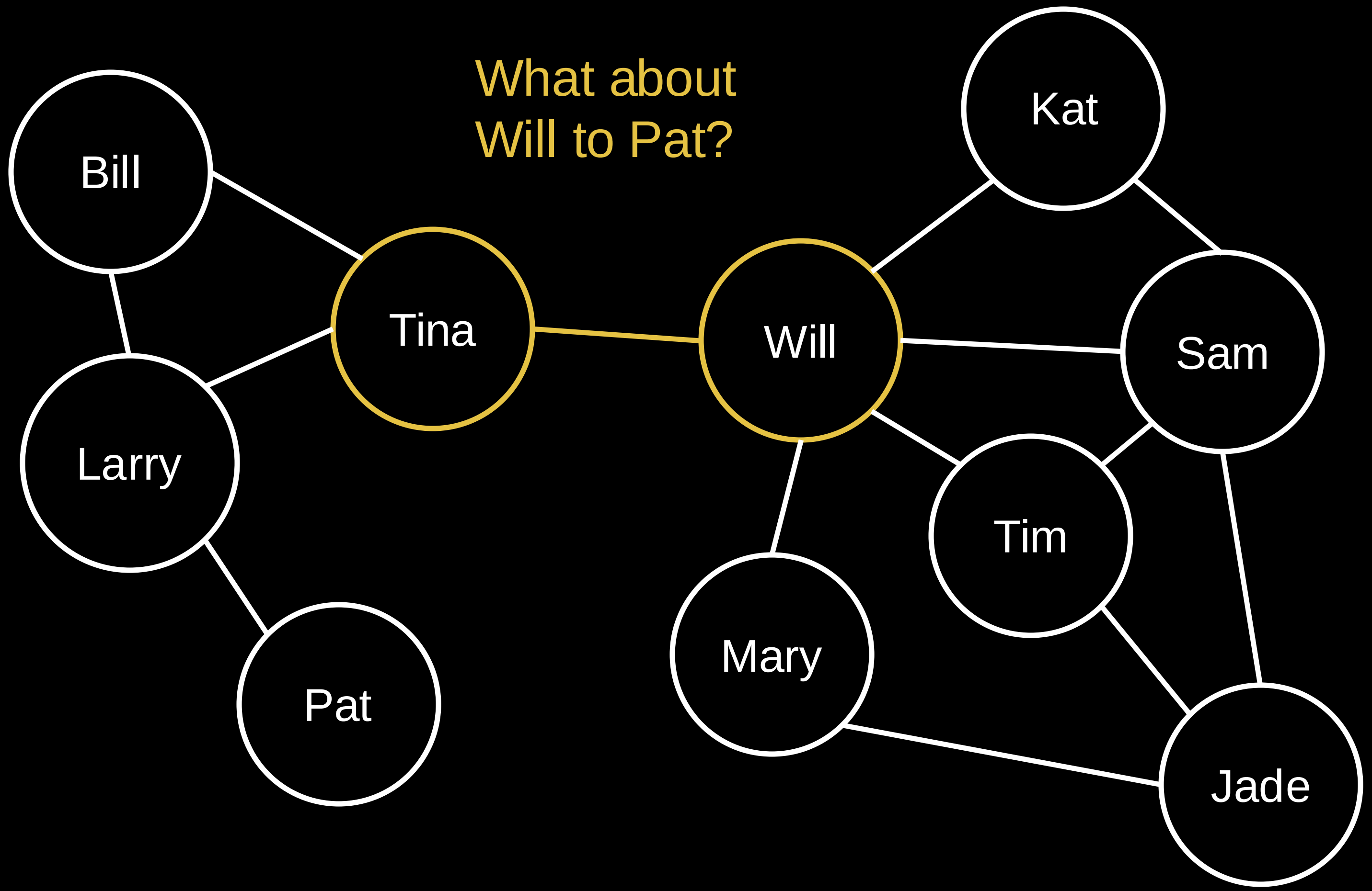
Degree of Separation (DOS)



Degree of Separation (DOS)

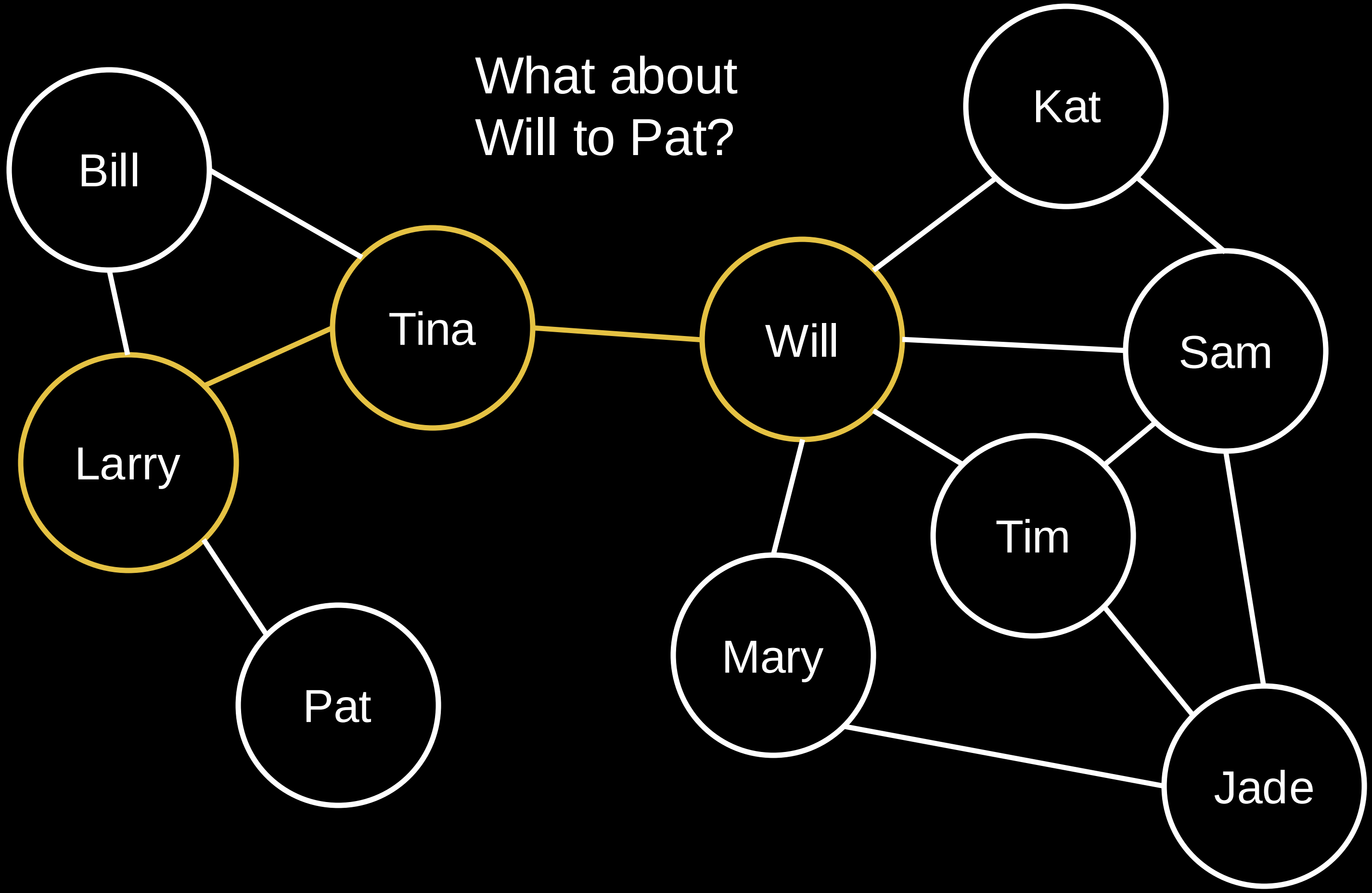


Degree of Separation (DOS)

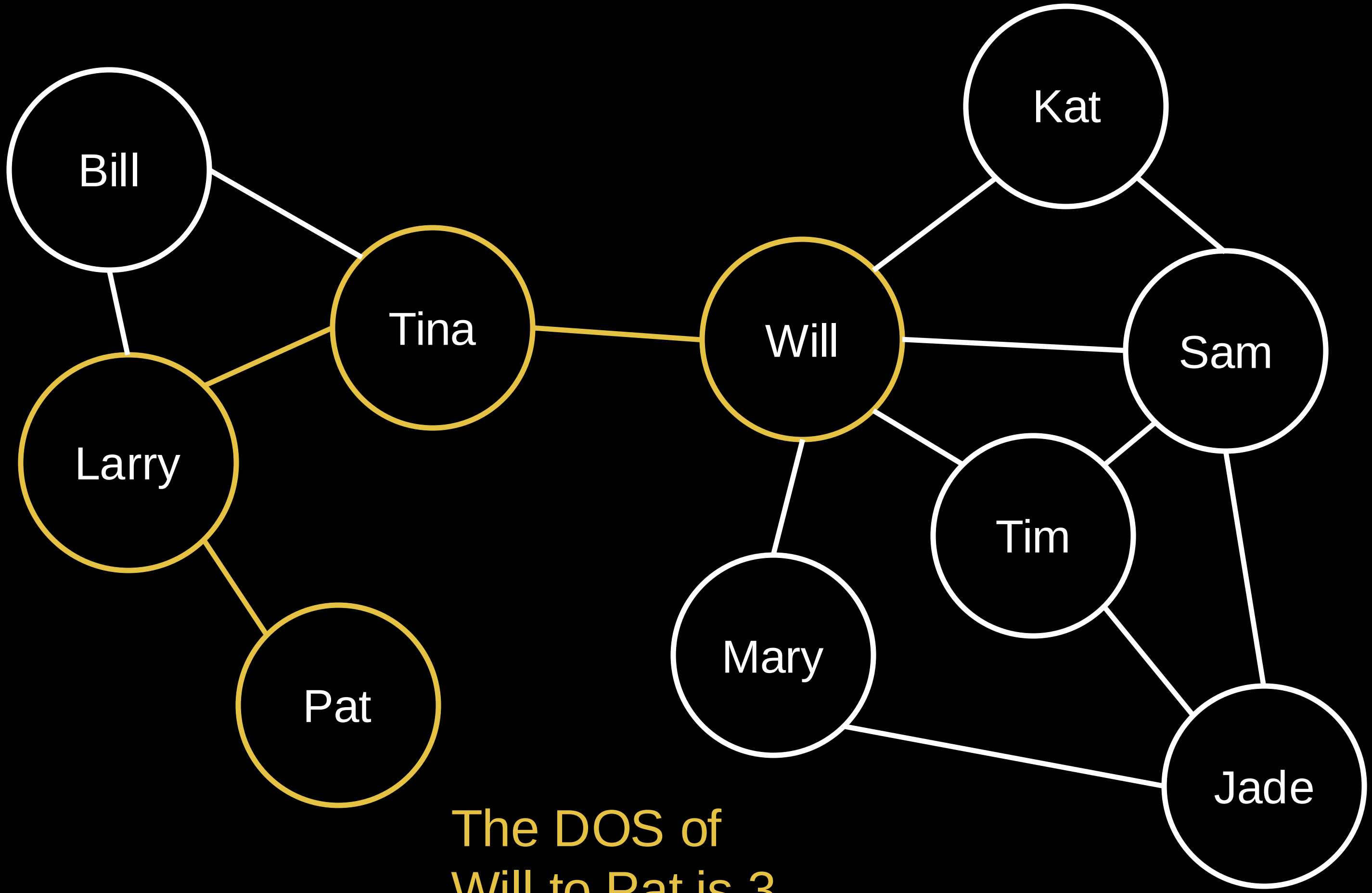


Degree of Separation (DOS)

What about
Will to Pat?



Degree of Separation (DOS)



The DOS of
Will to Pat is 3.

Closeness Centrality

- Closeness centrality is a measure of how well everyone in a network can connect to every other member of the network.
- It is calculated as follows:

$$C_C = \frac{N-1}{\sum_{i=1}^{N-1} DOS_i}$$

Closeness Centrality

Name	Closeness Centrality
Will	0.64
Tina	0.56
Sam	0.50
Tim	0.47
Kat	0.45
Mary	0.45
Larry	0.43
Bill	0.41
Jade	0.39
Pat	0.31

Eigenvector Centrality

	Bill	Larry	Tina	Pat	Will	Kat	Sam	Tim	Jade	Mary
Bill	0	1	1	0	0	0	0	0	0	0
Larry	1	0	1	1	0	0	0	0	0	0
Tina	1	1	0	0	1	0	0	0	0	0
Pat	0	1	0	0	0	0	0	0	0	0
Will	0	0	1	0	0	1	1	1	0	1
Kat	0	0	0	0	1	0	1	0	0	0
Sam	0	0	0	0	1	1	0	1	1	0
Tim	0	0	0	0	1	0	1	0	1	0
Jade	0	0	0	0	0	0	1	1	0	1
Mary	0	0	0	0	1	0	0	0	1	0

Eigenvector Centrality

- A node is high in eigenvector centrality if it is connected to many other nodes who are themselves well connected.
- A node's centrality is dependent on the centrality of adjacent nodes.
- These nodes would be considered influential – closely related to **diffusion** and **adoption**.

Eigenvector Centrality

- Eigenvector centrality for each node is simply calculated as the proportional eigenvector values of the eigenvector with the largest eigenvalue.

$$\boldsymbol{Ax = \lambda x}$$

Eigenvector Centrality

- Eigenvector centrality for each node is simply calculated as the proportional eigenvector values of the eigenvector with the largest eigenvalue.

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{\lambda}\boldsymbol{x}$$

Find largest eigenvalue

Eigenvector Centrality

- Eigenvector centrality for each node is simply calculated as the proportional eigenvector values of the eigenvector with the largest eigenvalue.

$$\boldsymbol{Ax} = \boldsymbol{\lambda x}$$

Find corresponding eigenvector

Eigenvector Centrality

Name	Scaled Eigenvector Centrality
Will	1.00
Sam	0.94
Tim	0.80
Jade	0.69
Kat	0.59
Mary	0.52
Tina	0.43
Larry	0.21
Bill	0.19
Pat	0.06

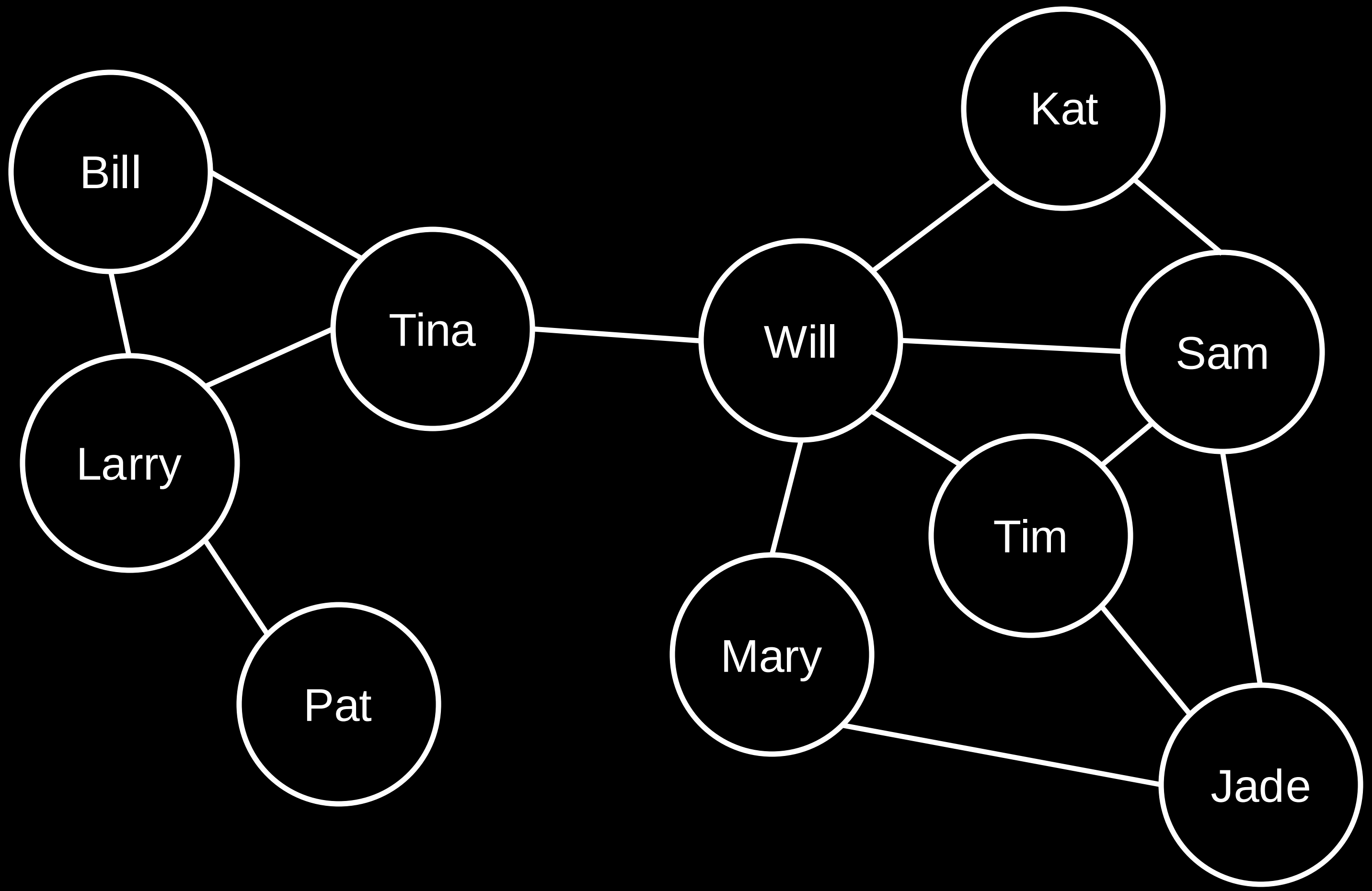
Social Structure

- There are many different summaries and important calculations obtained from sociograms.
- Here are a few we will focus on:
  - Subgroups
  - Centers and Closeness
  - Brokers and Bridges
  - Diffusion and Adoption

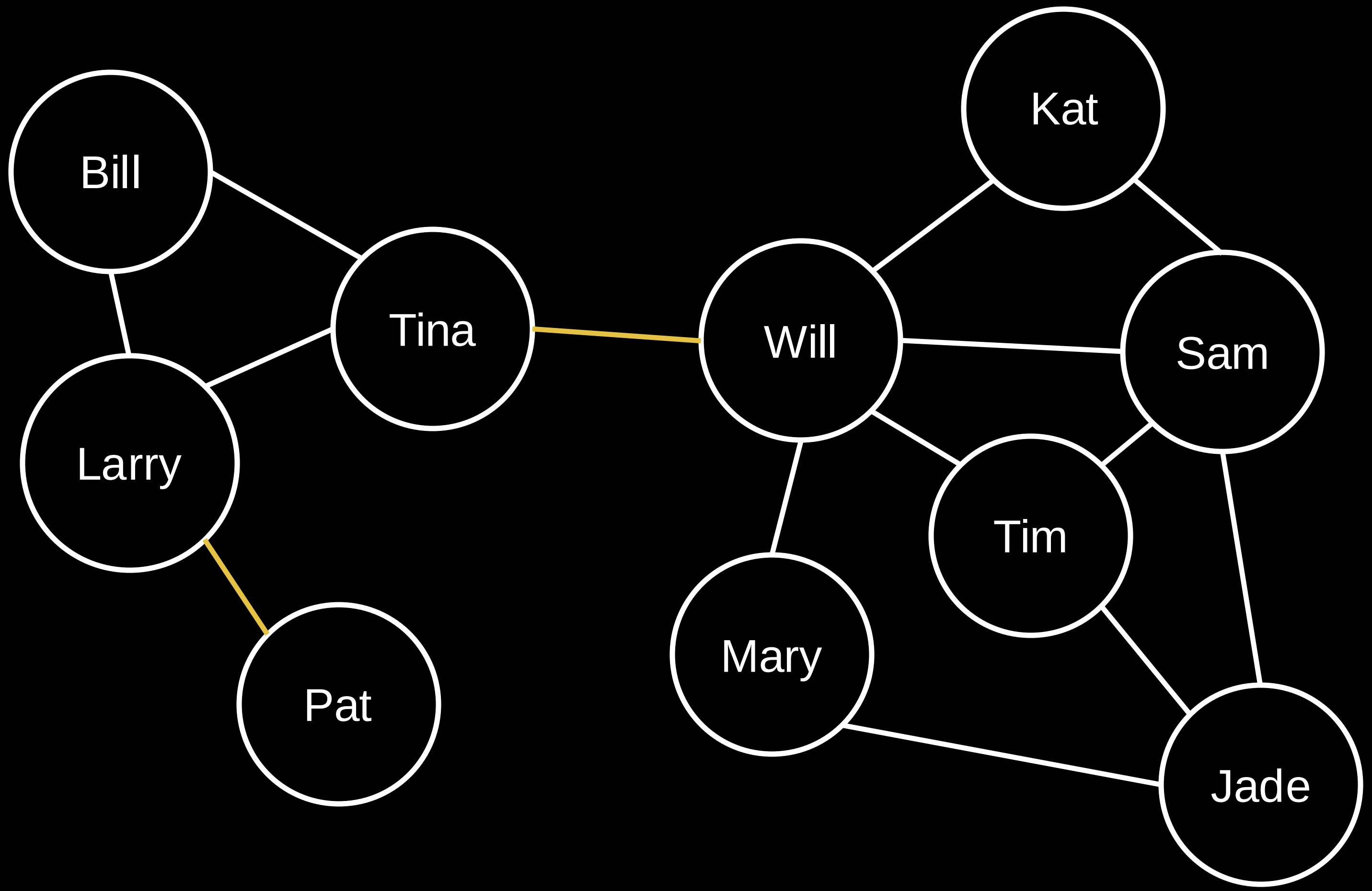
Different Links

- Not only are number of links important, but the kind of link is extremely important as well.
- Links with individuals who are linked themselves is not as strong as links with individuals who are not linked together.
- Links within a subgroup yield little new information compared to links with other subgroups.
- A **bridge** is a link whose removal increases the number of isolated nodes.

Bridge



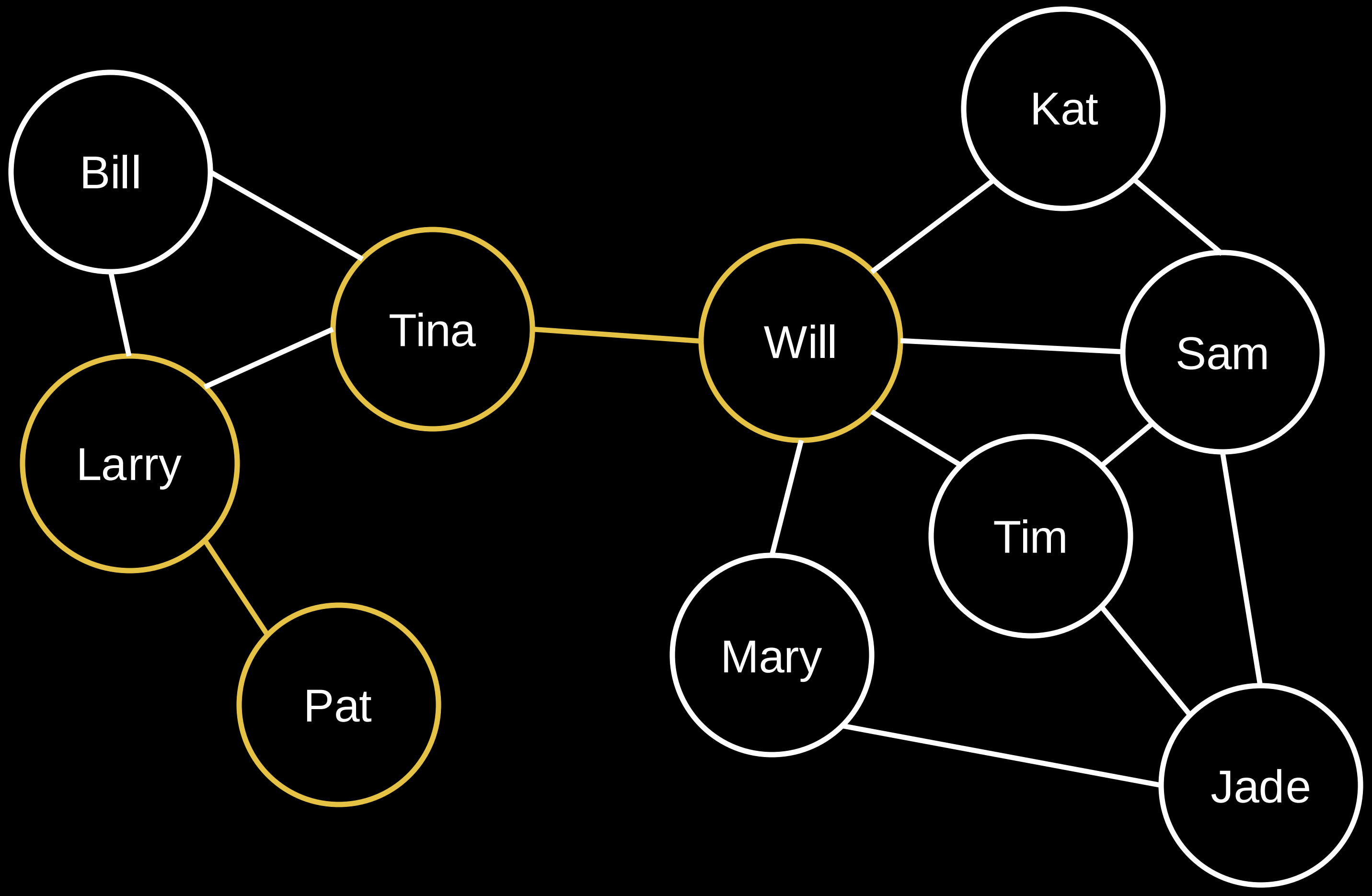
Bridge



Brokers

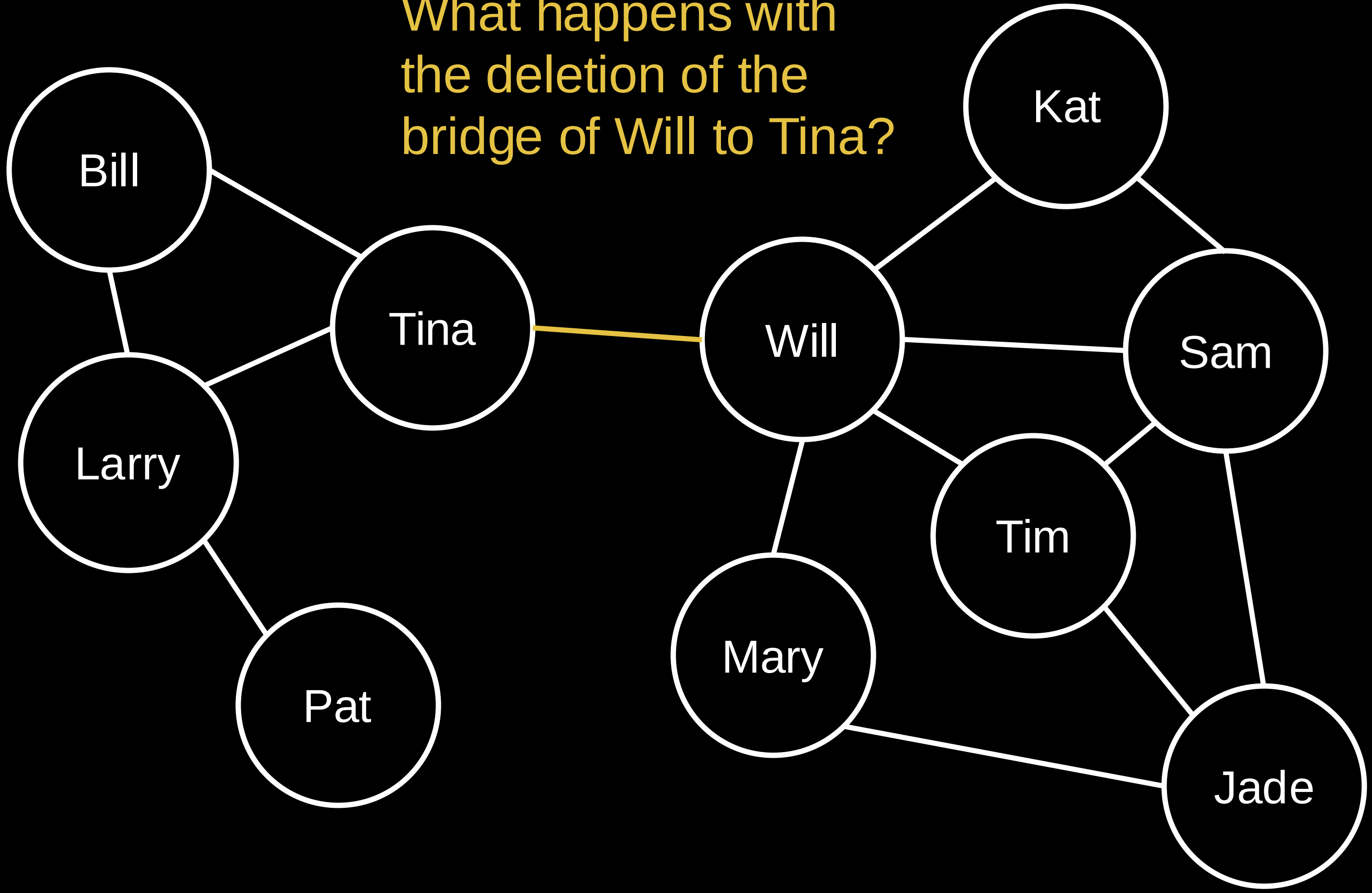
- These bridges are important because they are a potential bottleneck of information.
- The individuals that are connected to these bridges are called **brokers** because they facilitate the information between the two sides of the bridge.
- By eliminating either the bridge or the broker, the spread of information across the network becomes limited.
- Important Applications:
  - Fraud detection
  - Disease contamination
  - Marketing campaigns

Brokers

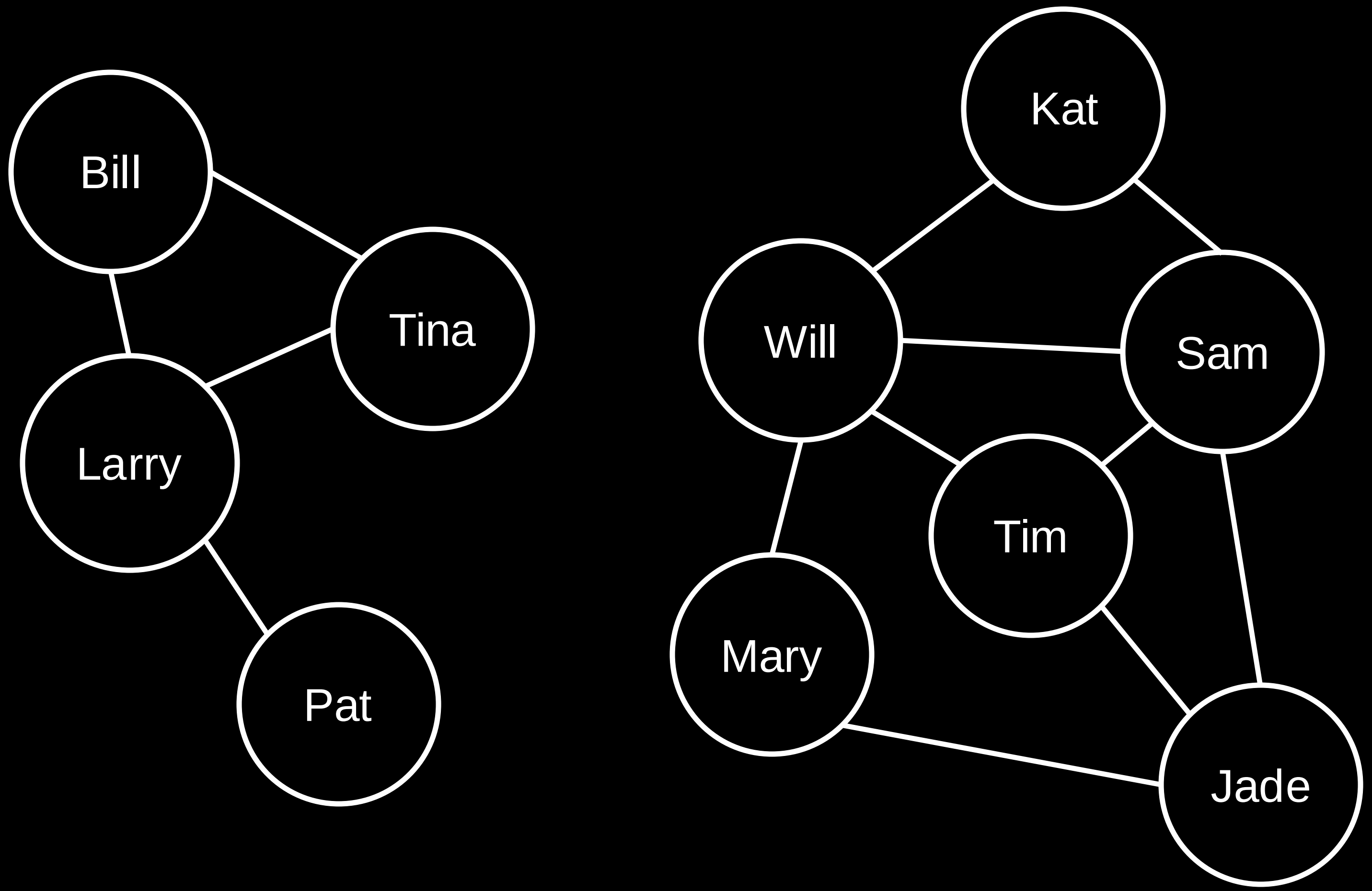


Bridge Elimination

What happens with the deletion of the bridge of Will to Tina?

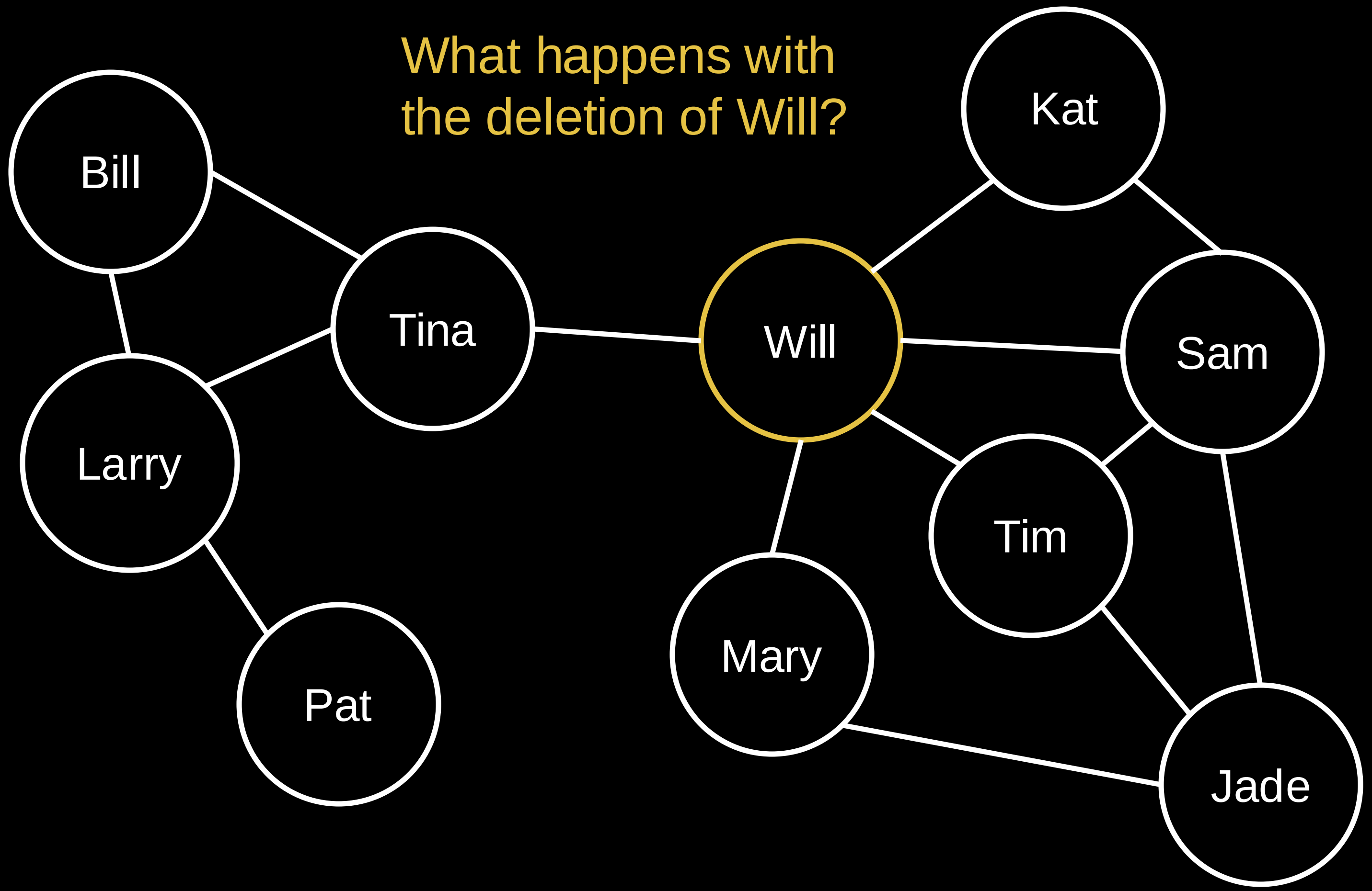


Bridge Elimination

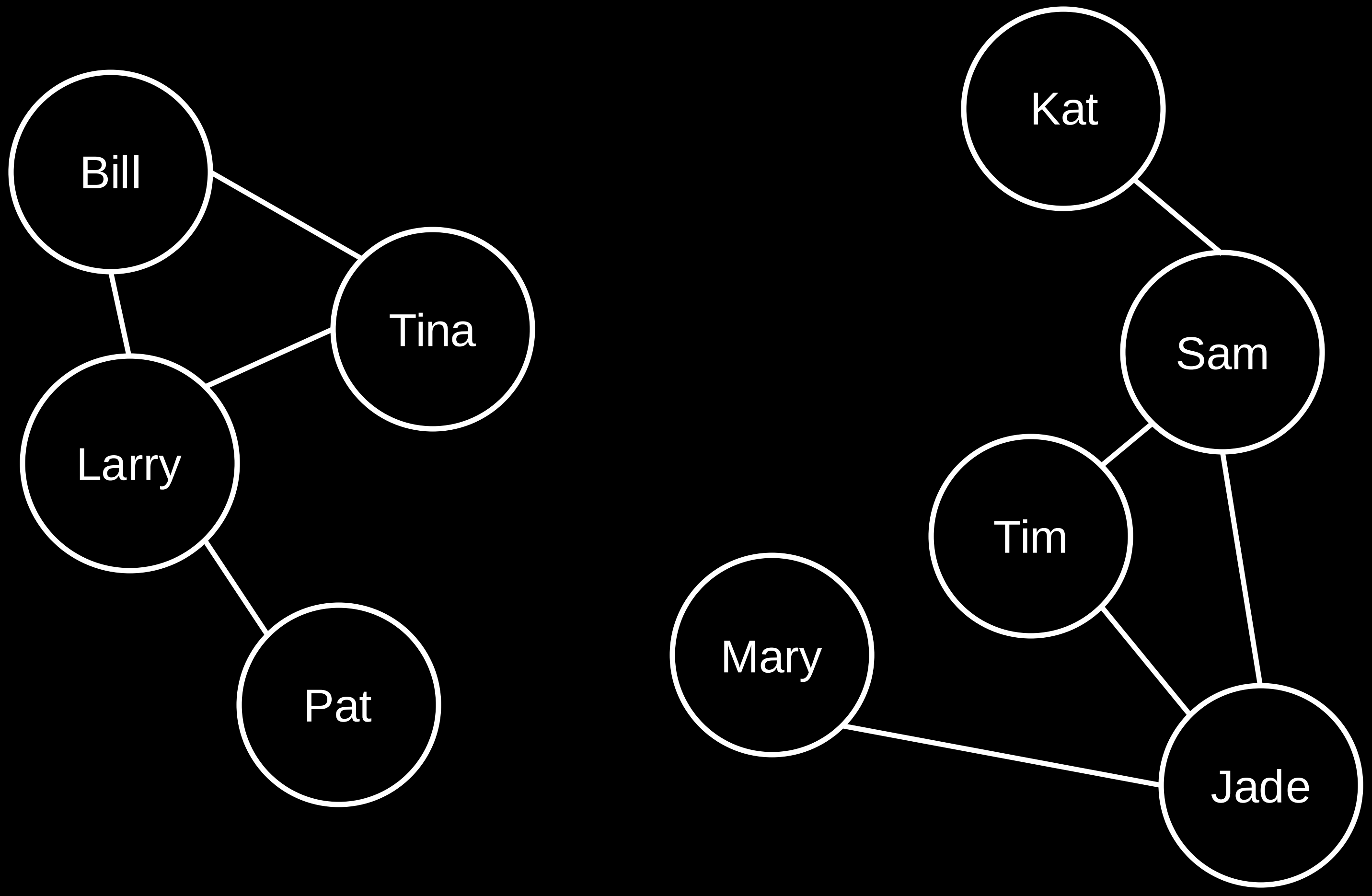


Broker Elimination

What happens with
the deletion of Will?



Broker Elimination



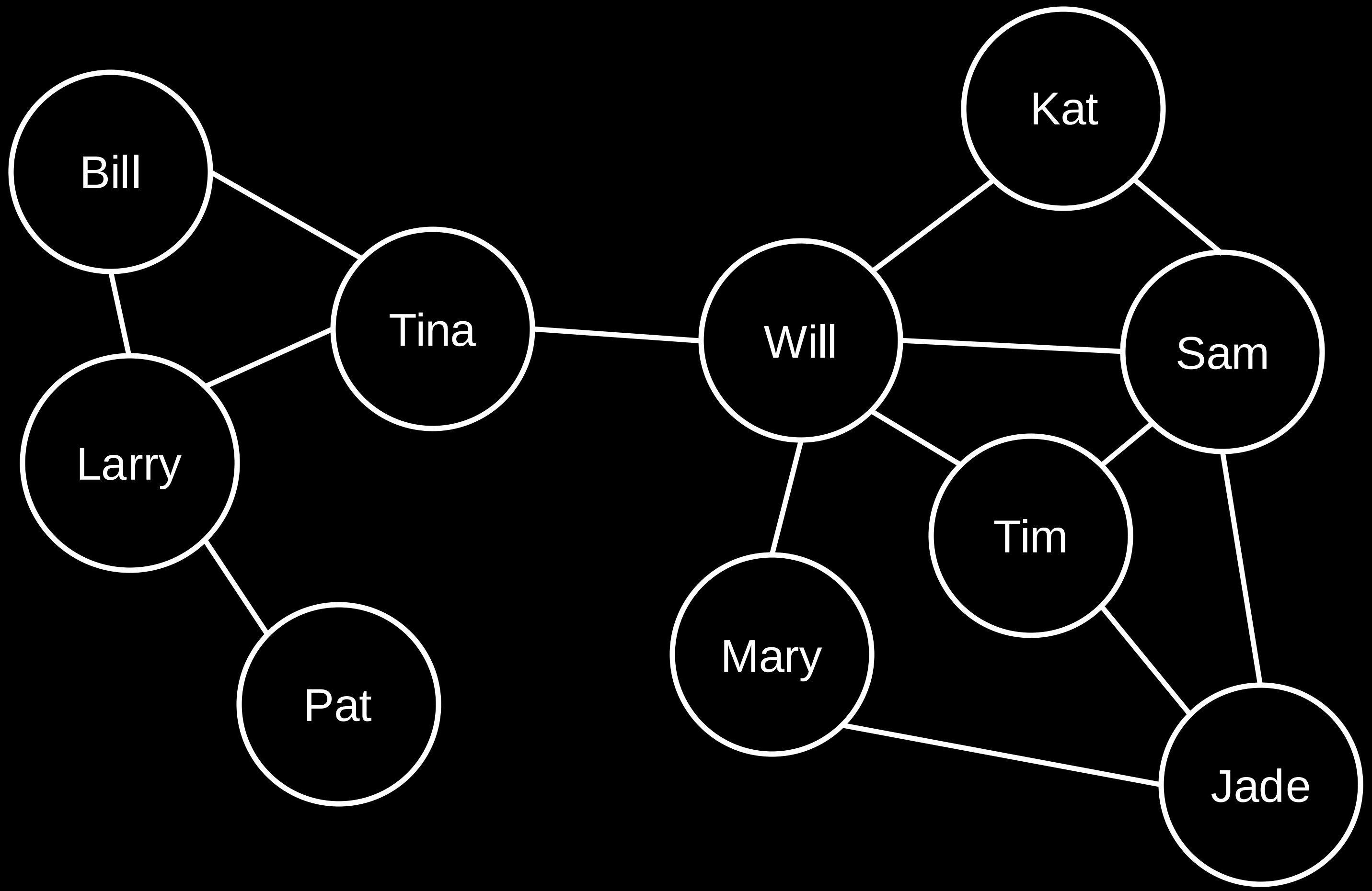
Social Structure

- There are many different summarizes and important calculations obtained from sociograms.
- Here are a few we will focus on:
  - Subgroups
  - Centers and Closeness
  - Brokers and Bridges
  - Diffusion and Adoption

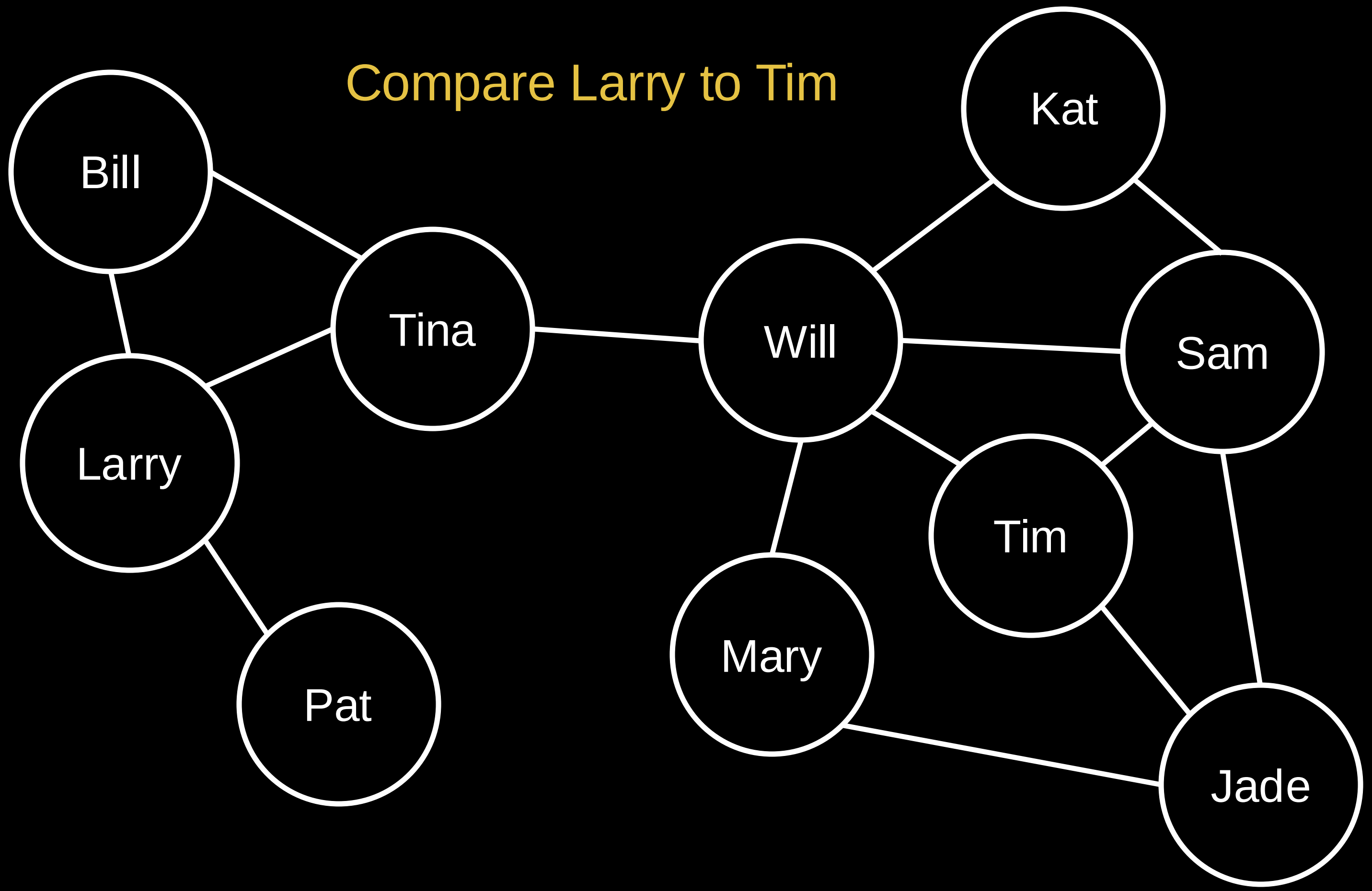
Diffusion and Adoption

- Diffusion and adoption add a sense of time to a sociogram.
- How long does it take for the entire network to **adopt** an idea based on initial location?

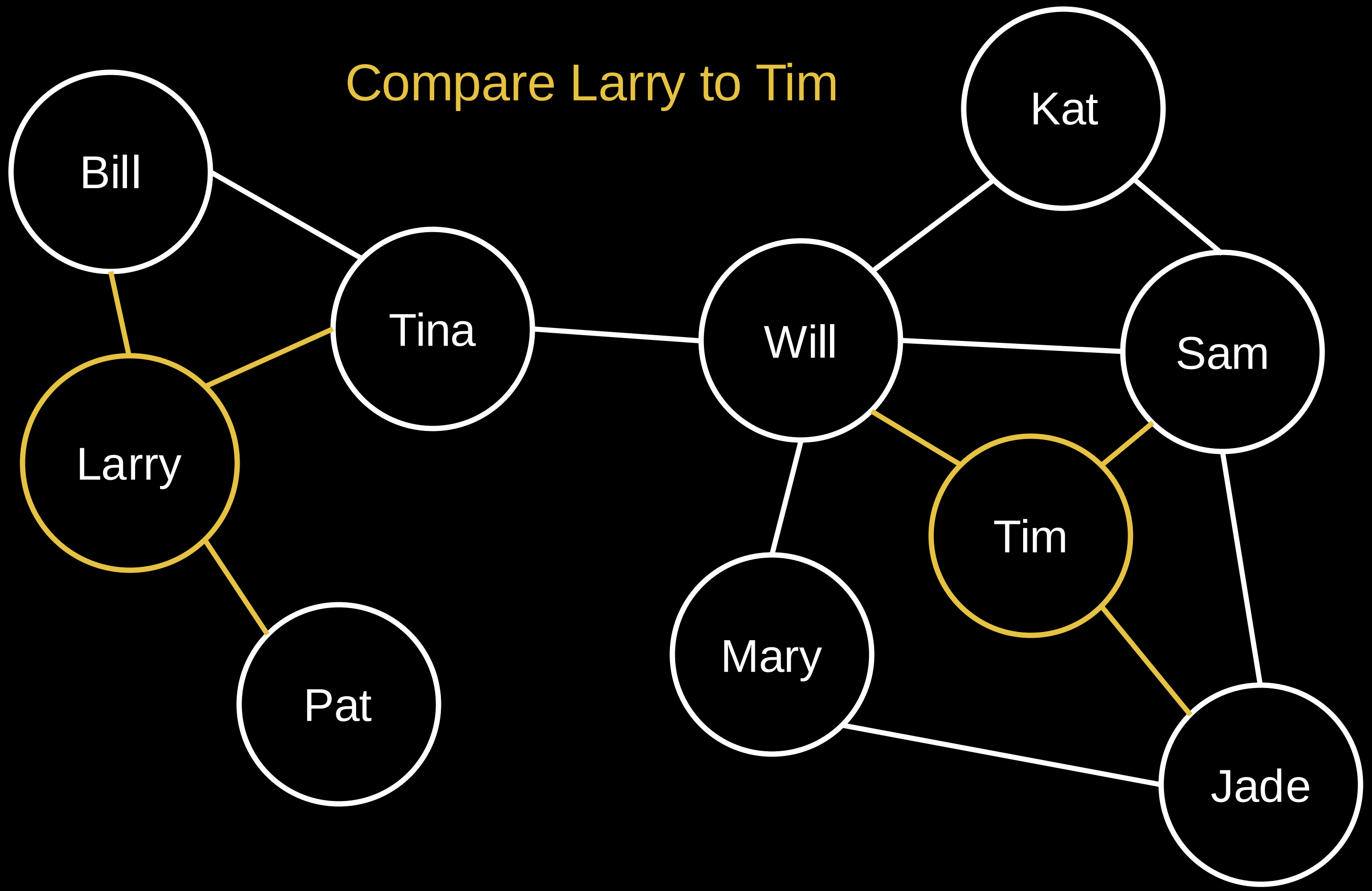
Location Matters



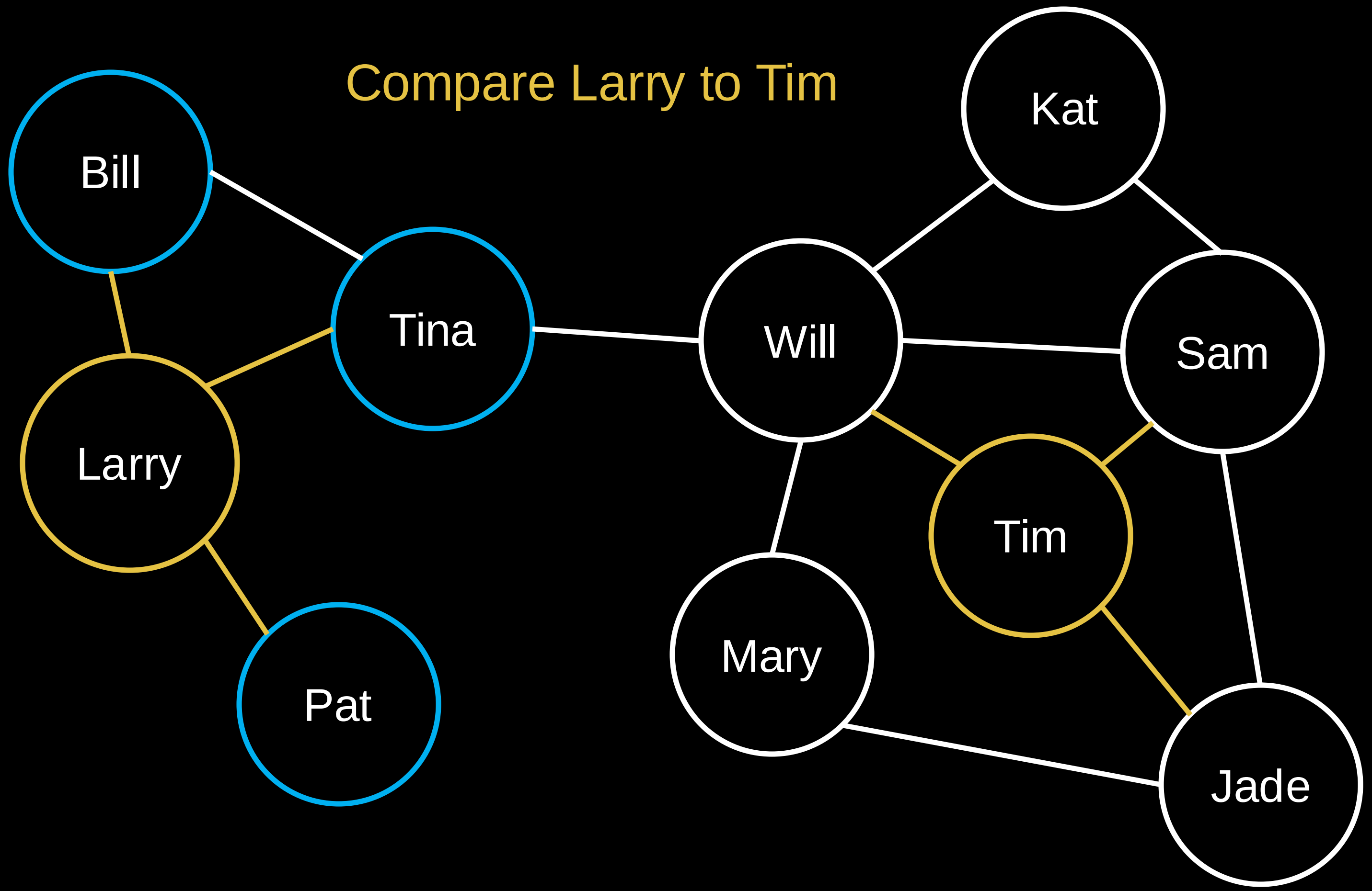
Location Matters



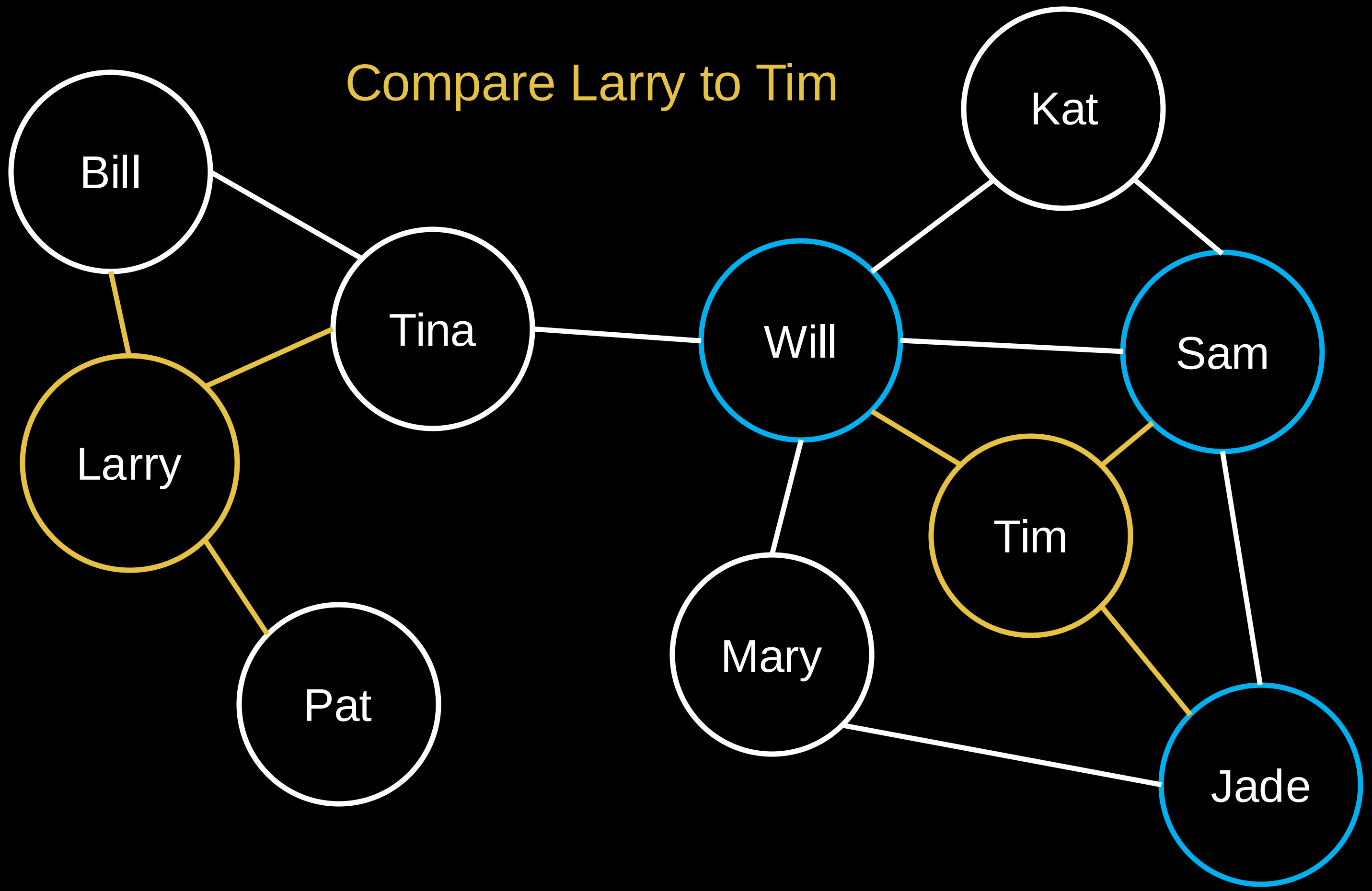
Location Matters



Location Matters



Location Matters



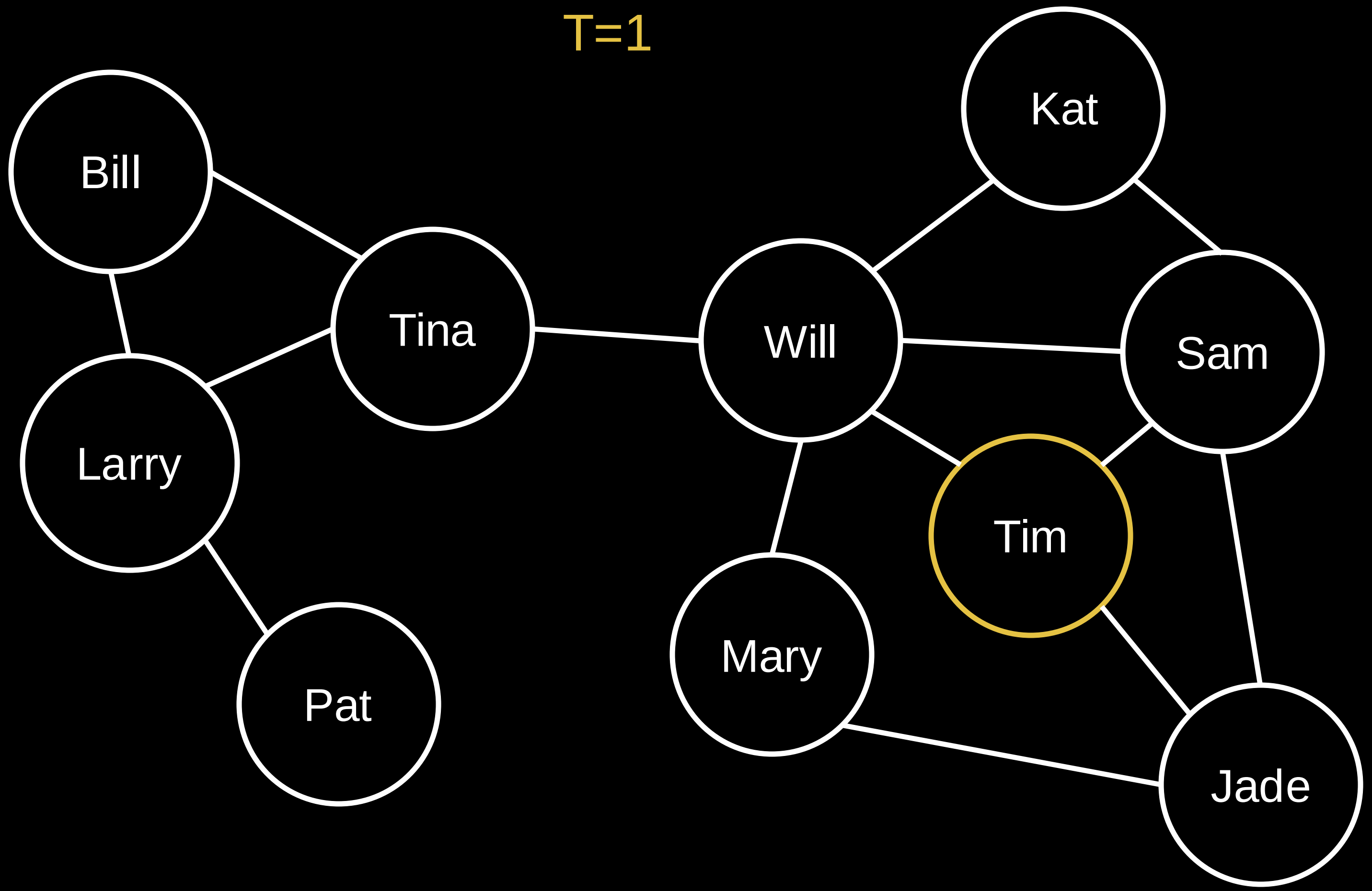
Location Matters

- Only looking at the counts of the links wouldn't be able to explain the information that is summarized in the graph.
- How is this important?
  - Disease prevention – who would you rather get sick, Larry or Tim?
  - Marketing choices – who would you rather sell your product to, Larry or Tim?

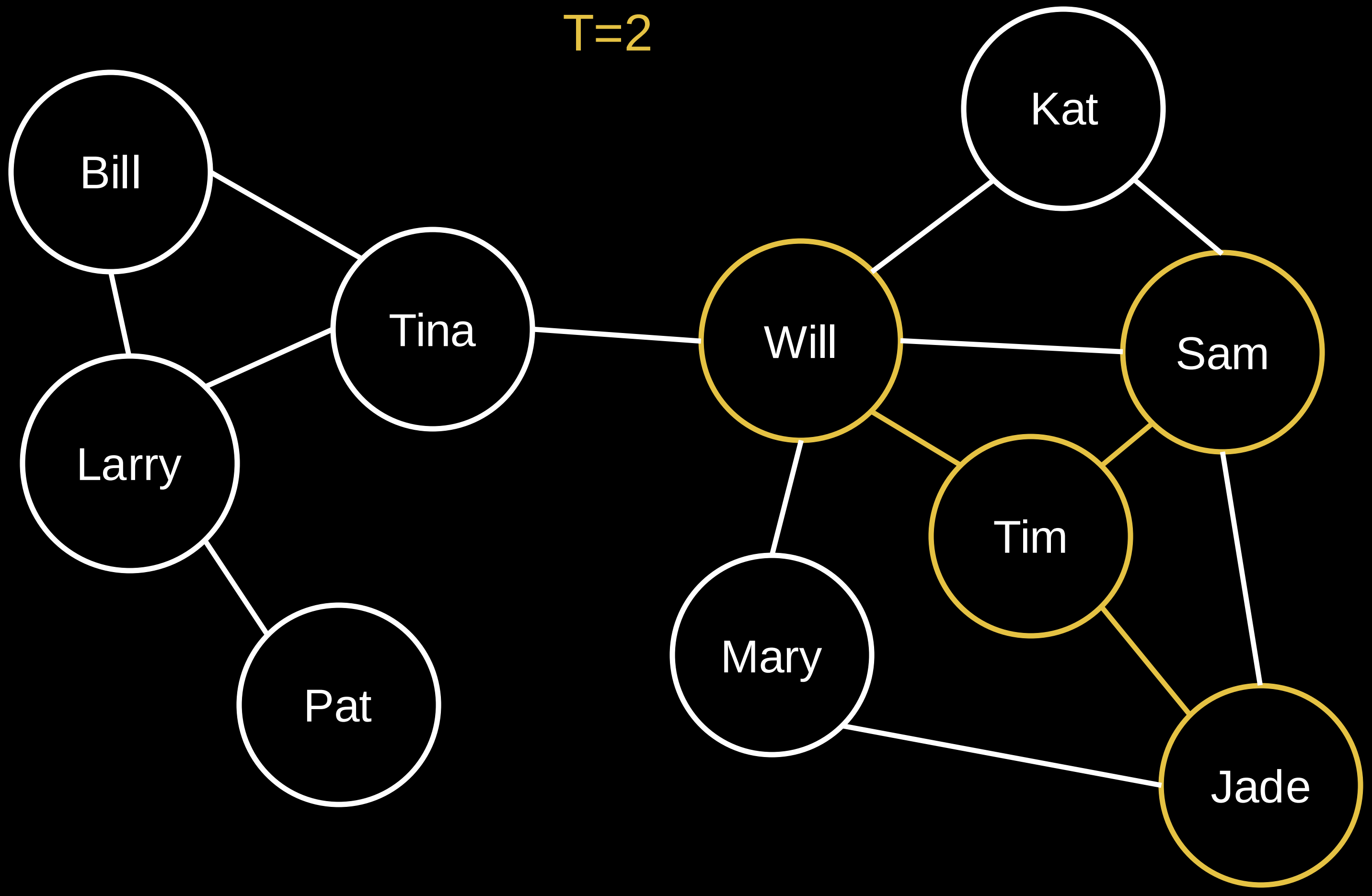
Location Matters

- Let's focus on the disease contamination example.
- Assume the disease moves from one individual to every one of the individual's contacts in one time period.
- This pattern persists in the next time period until all nodes in the network are contaminated.

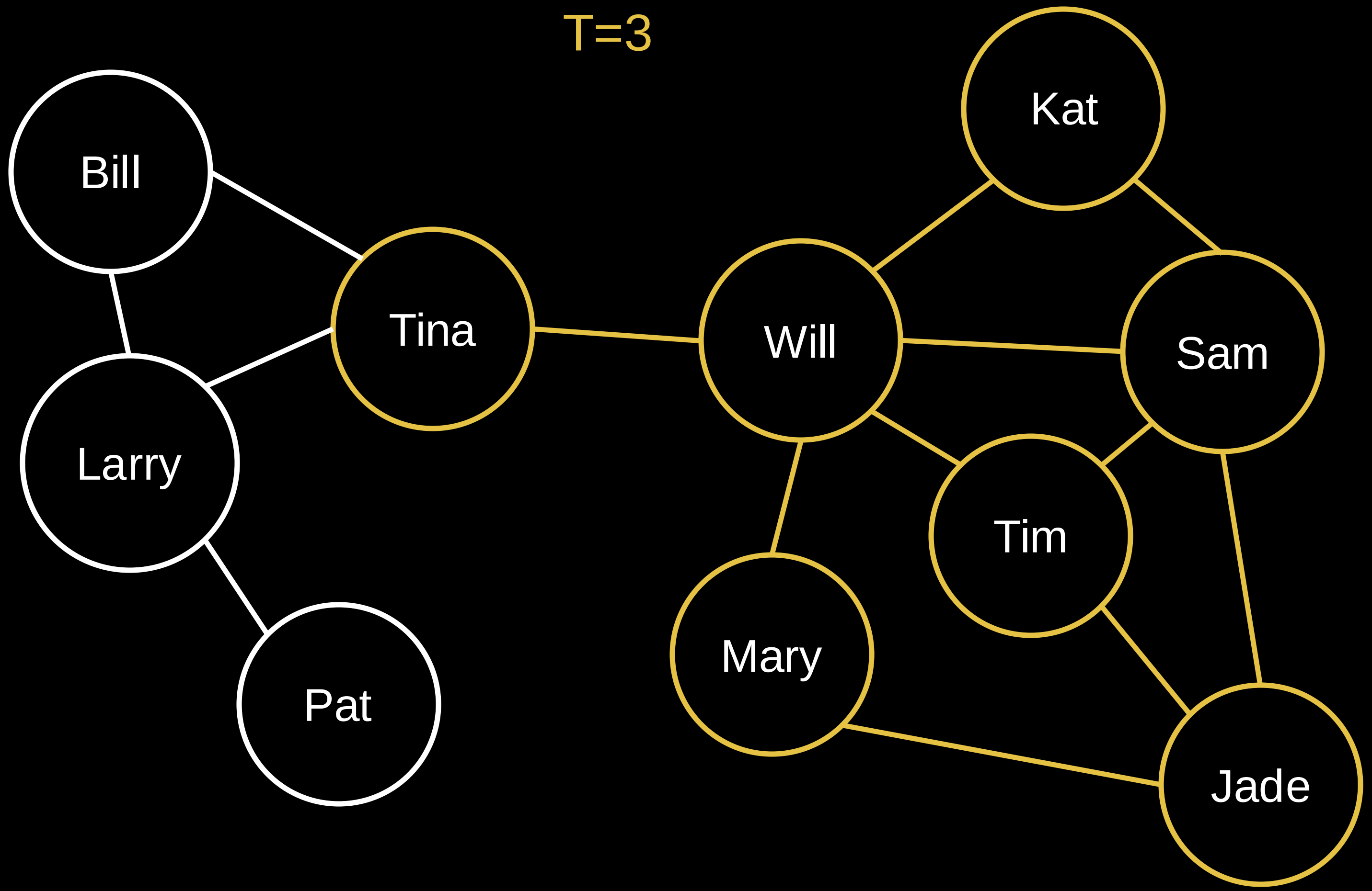
Location Matters



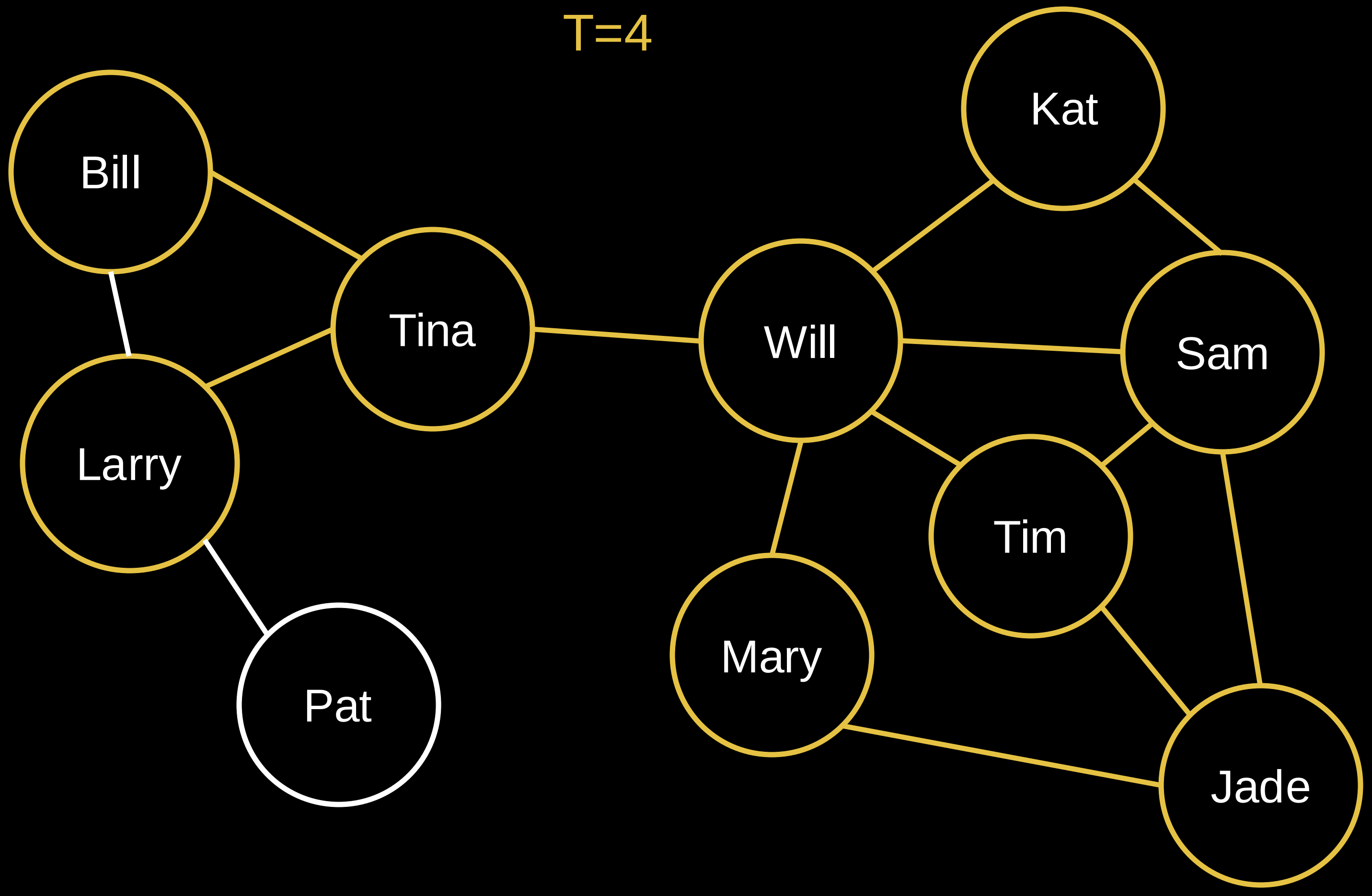
Location Matters



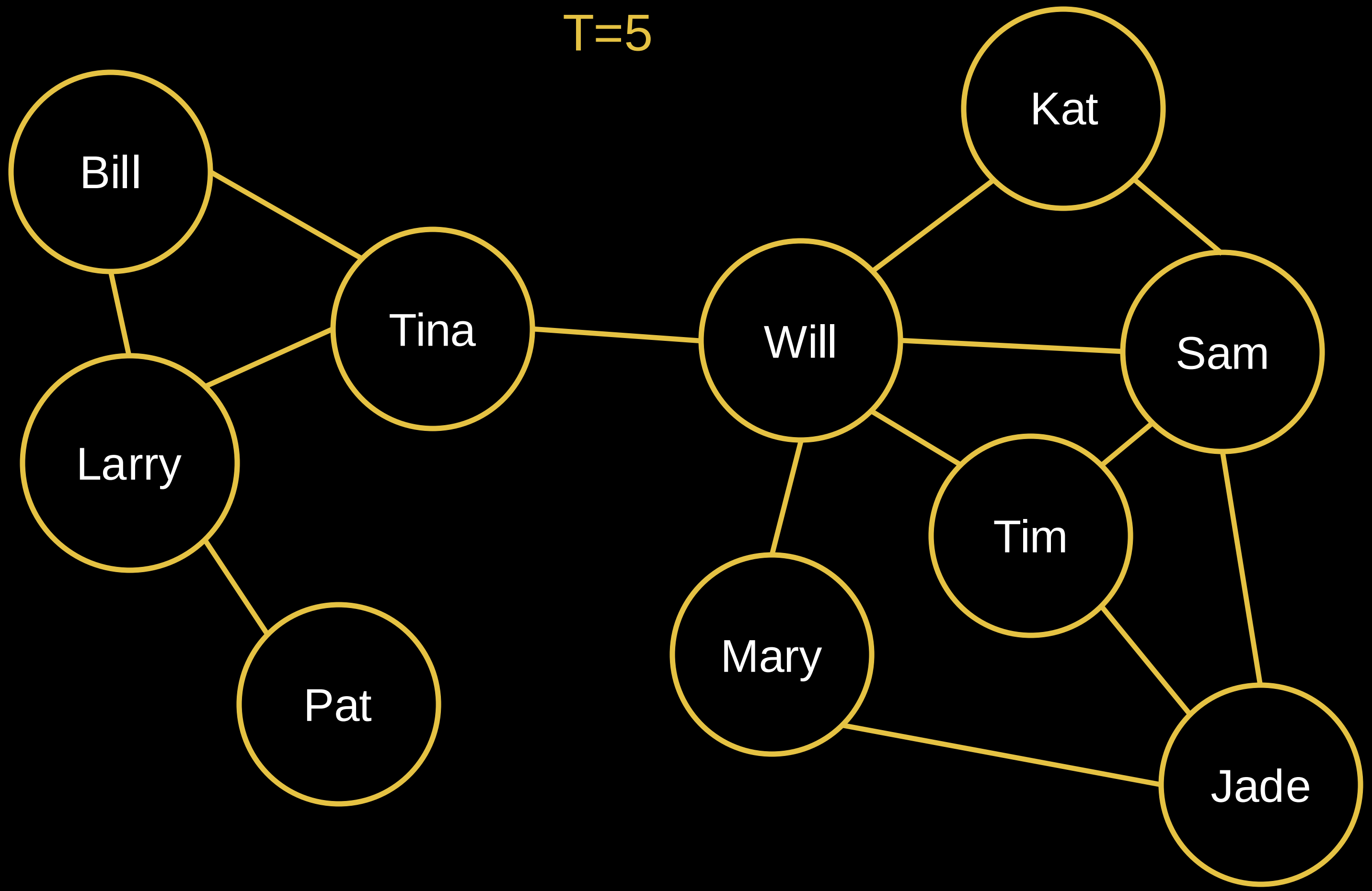
Location Matters



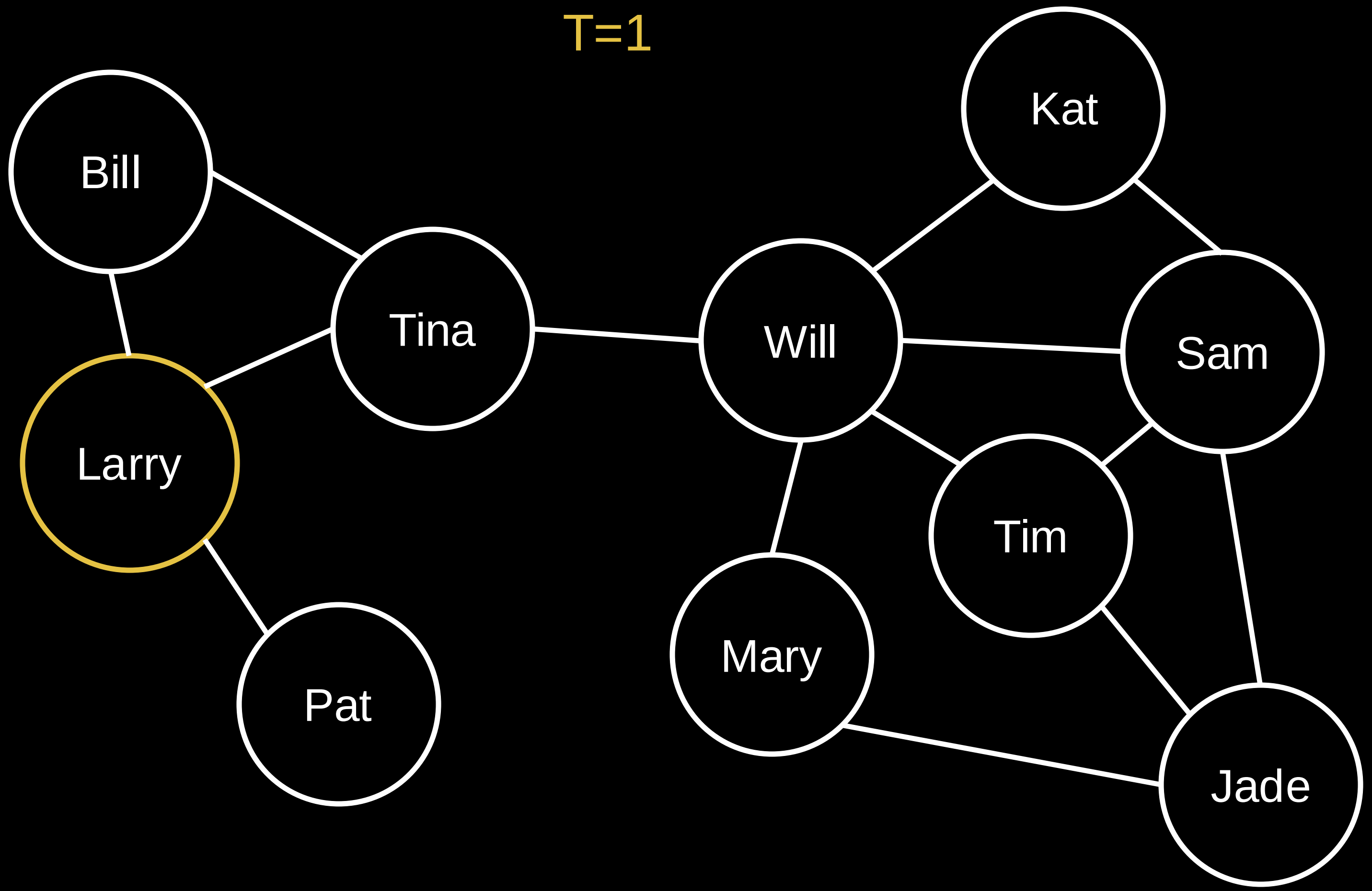
Location Matters



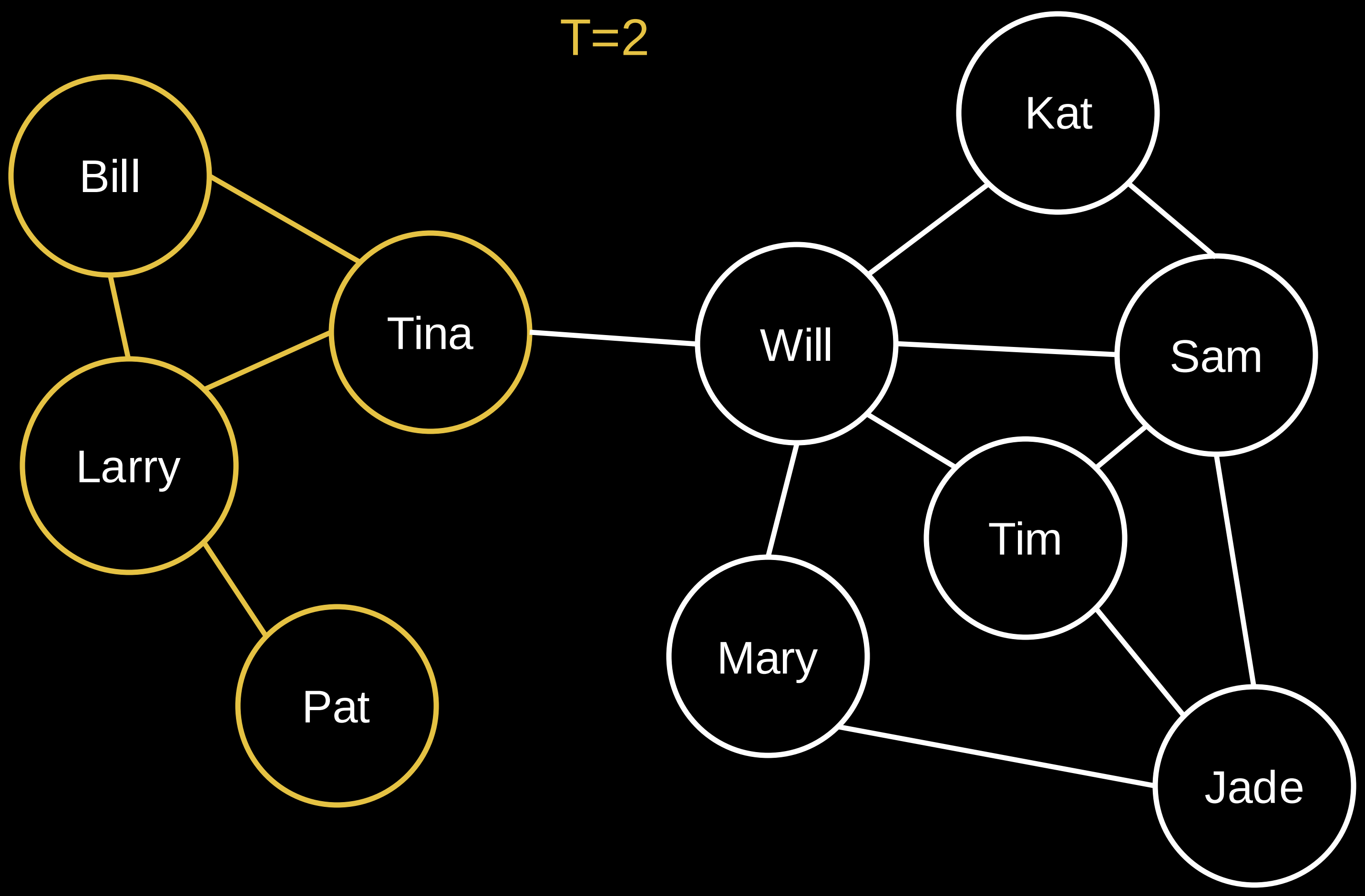
Location Matters



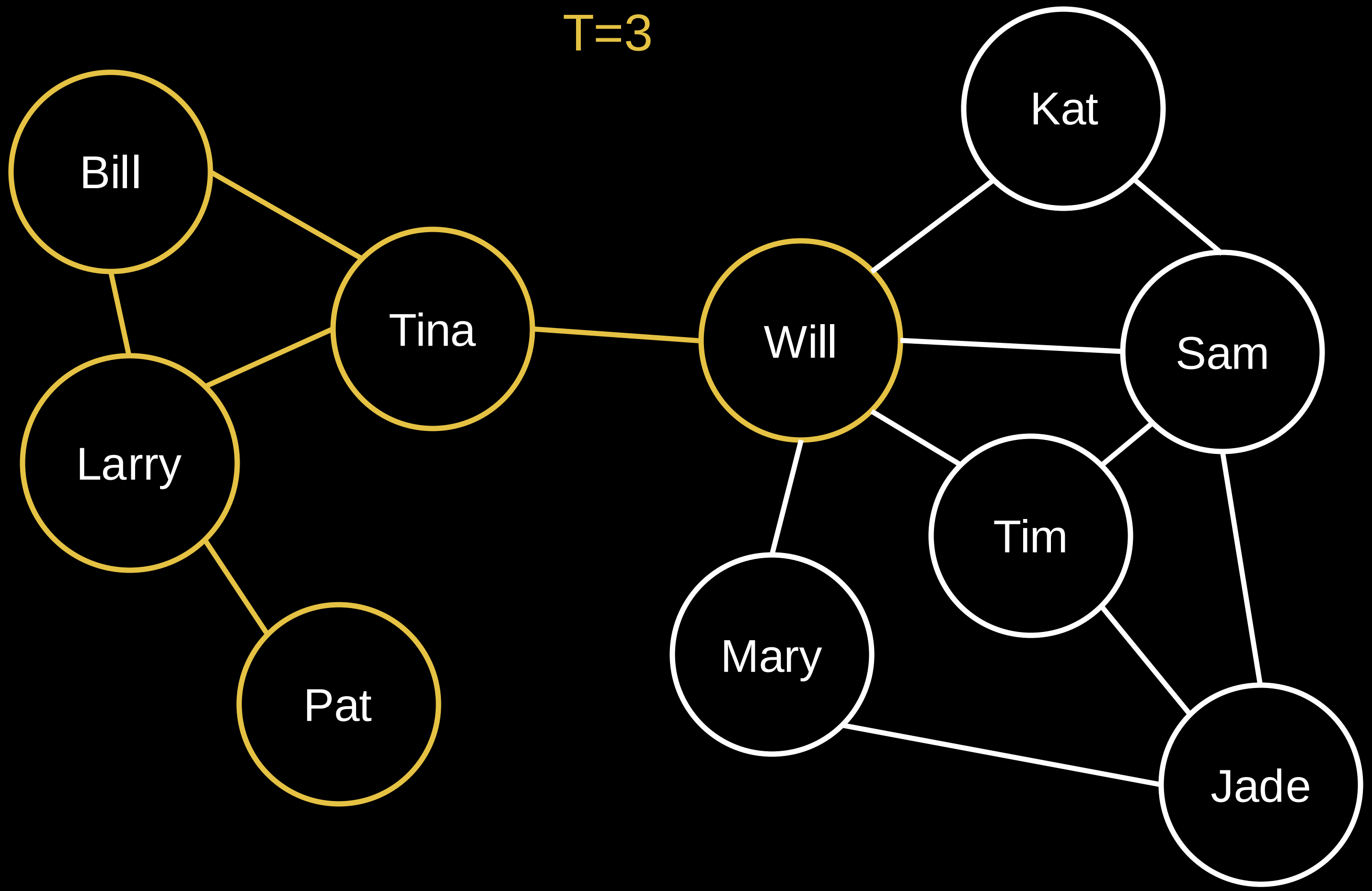
Location Matters



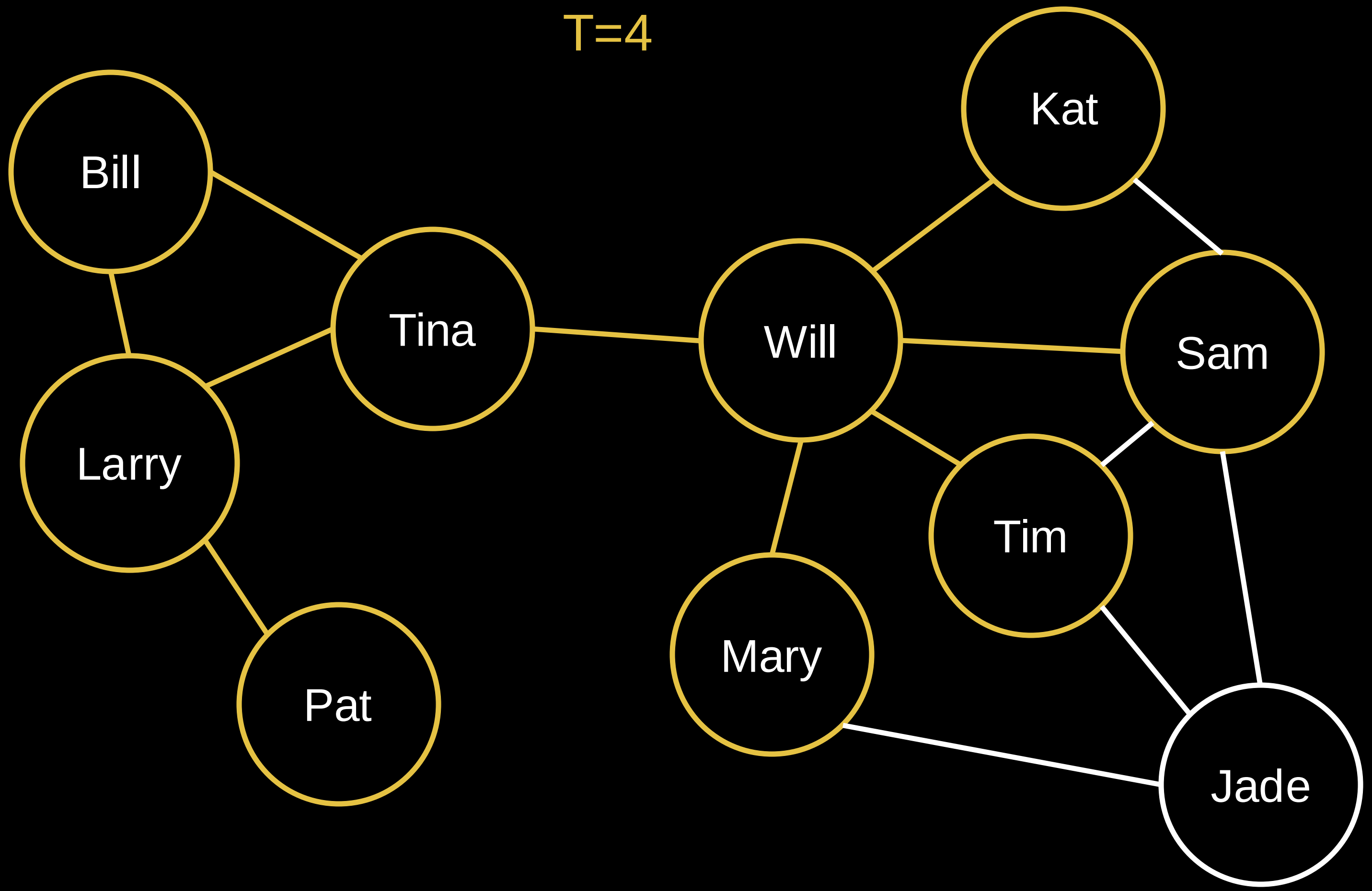
Location Matters



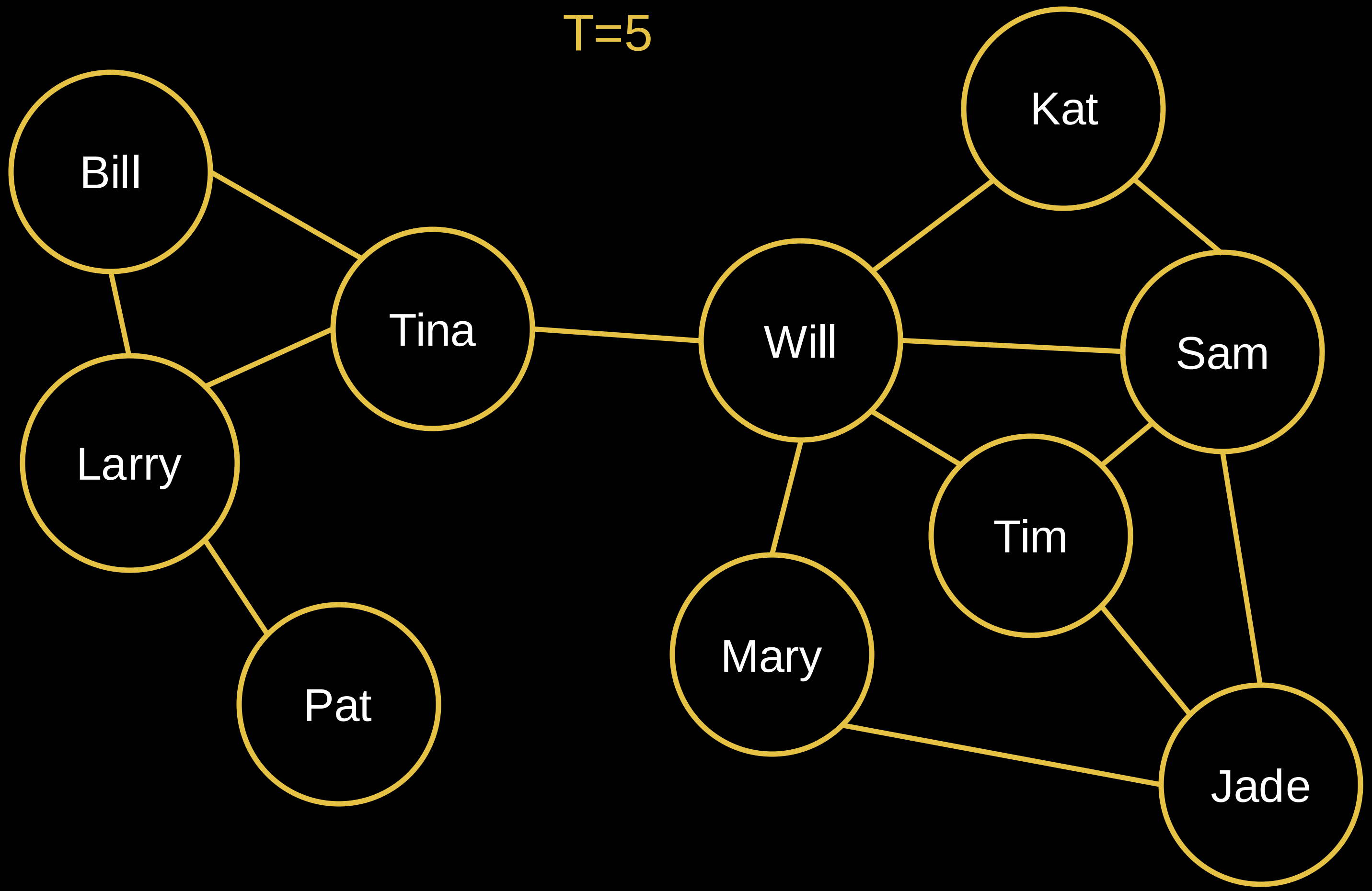
Location Matters



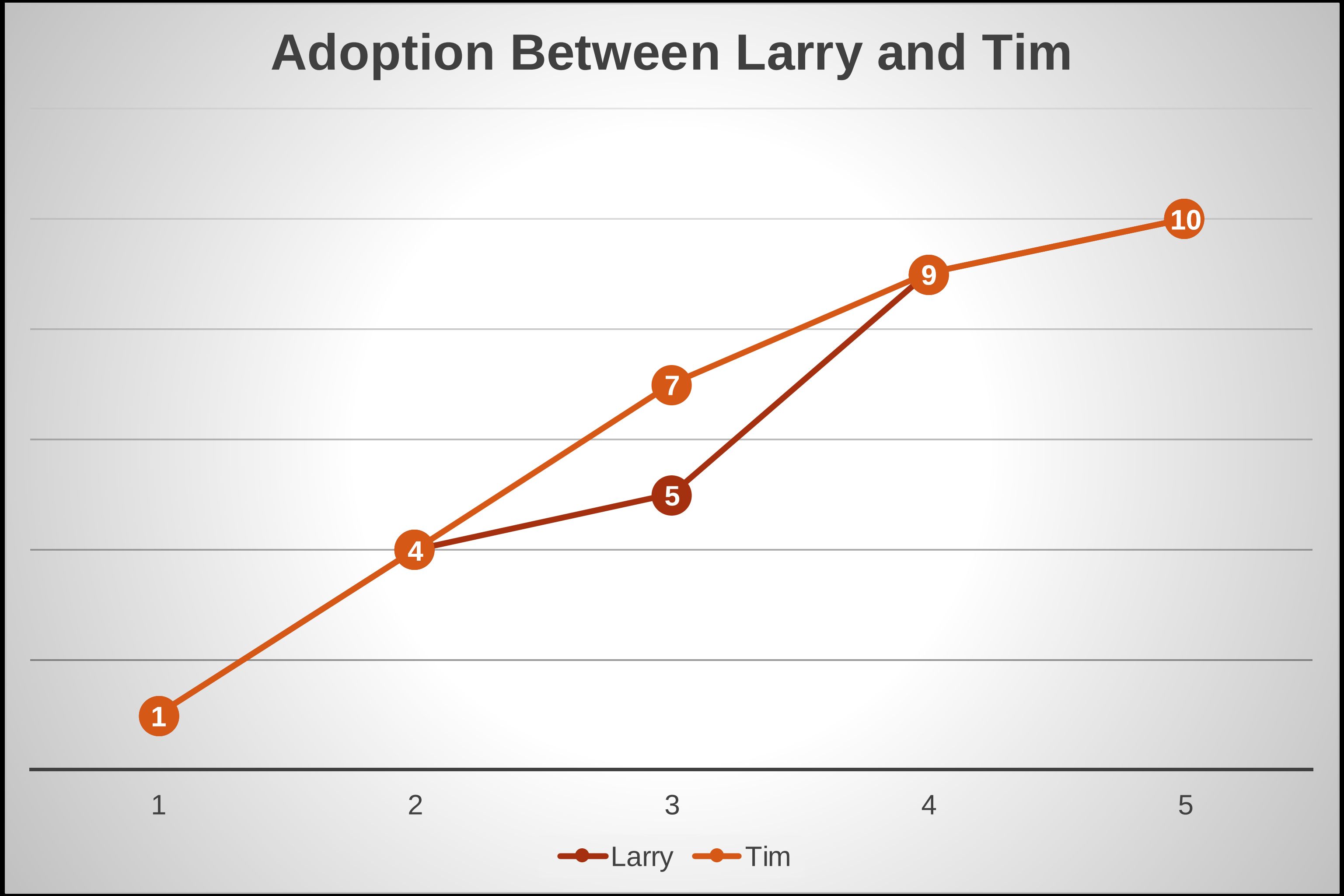
Location Matters



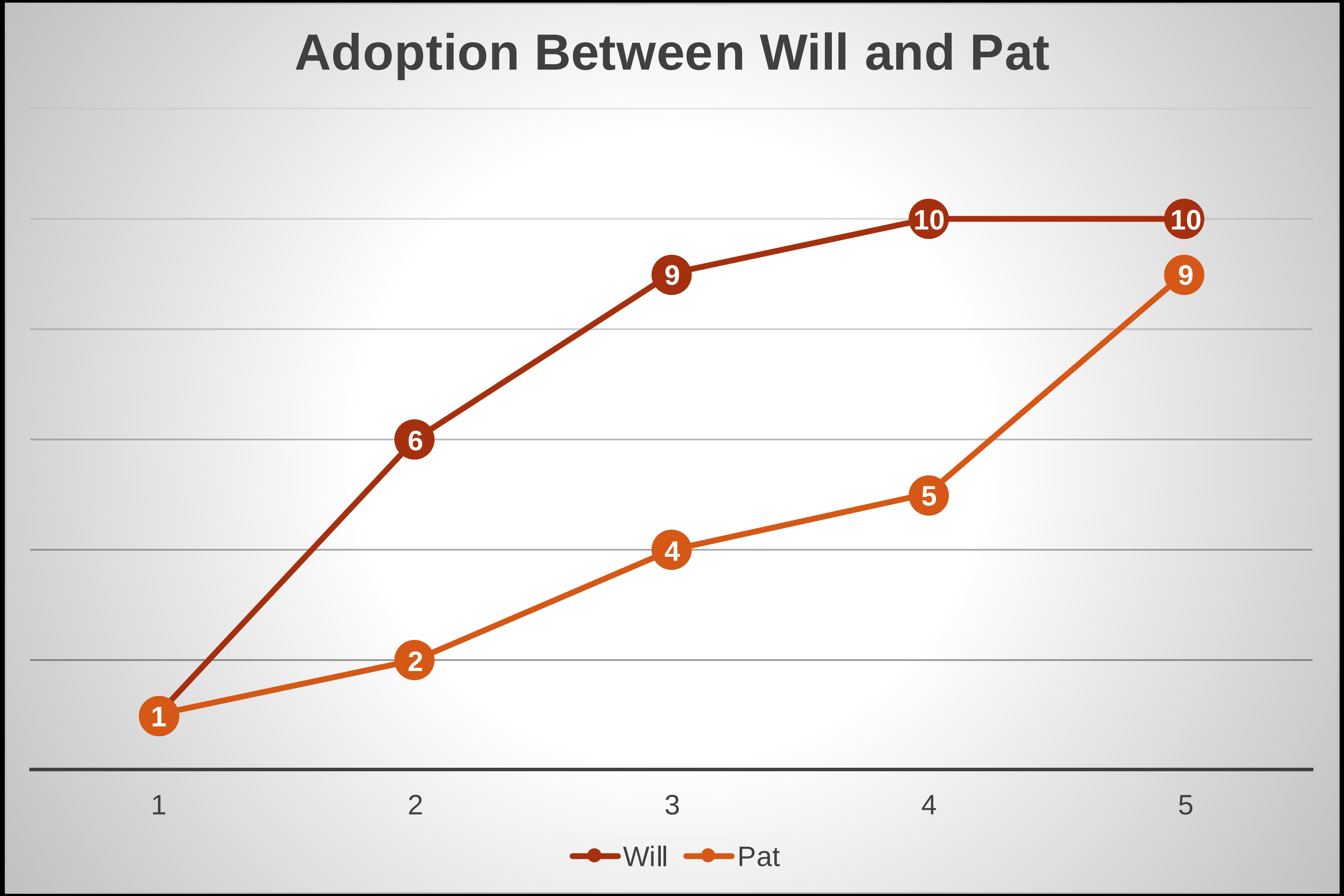
Location Matters



Location Matters



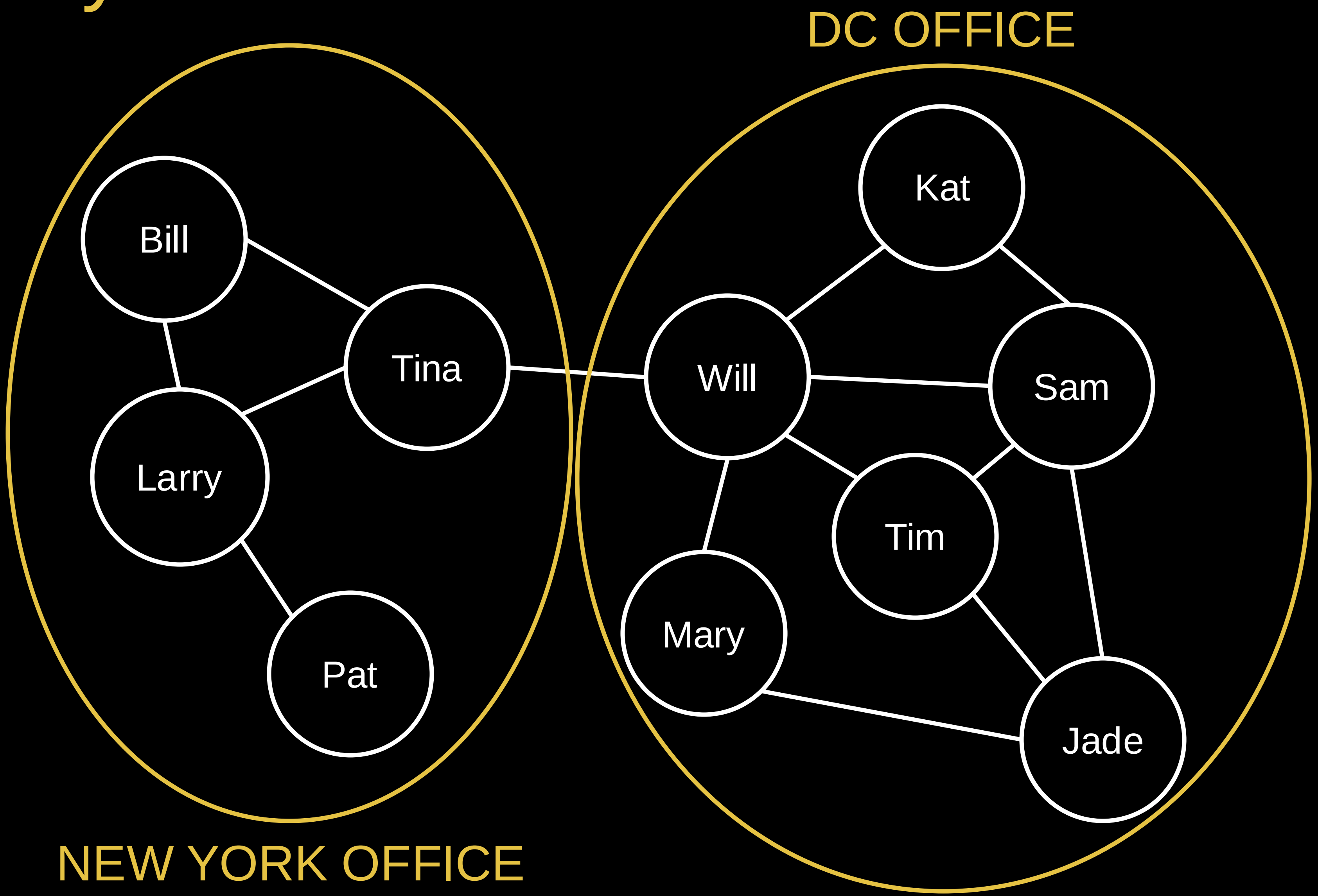
Location Matters



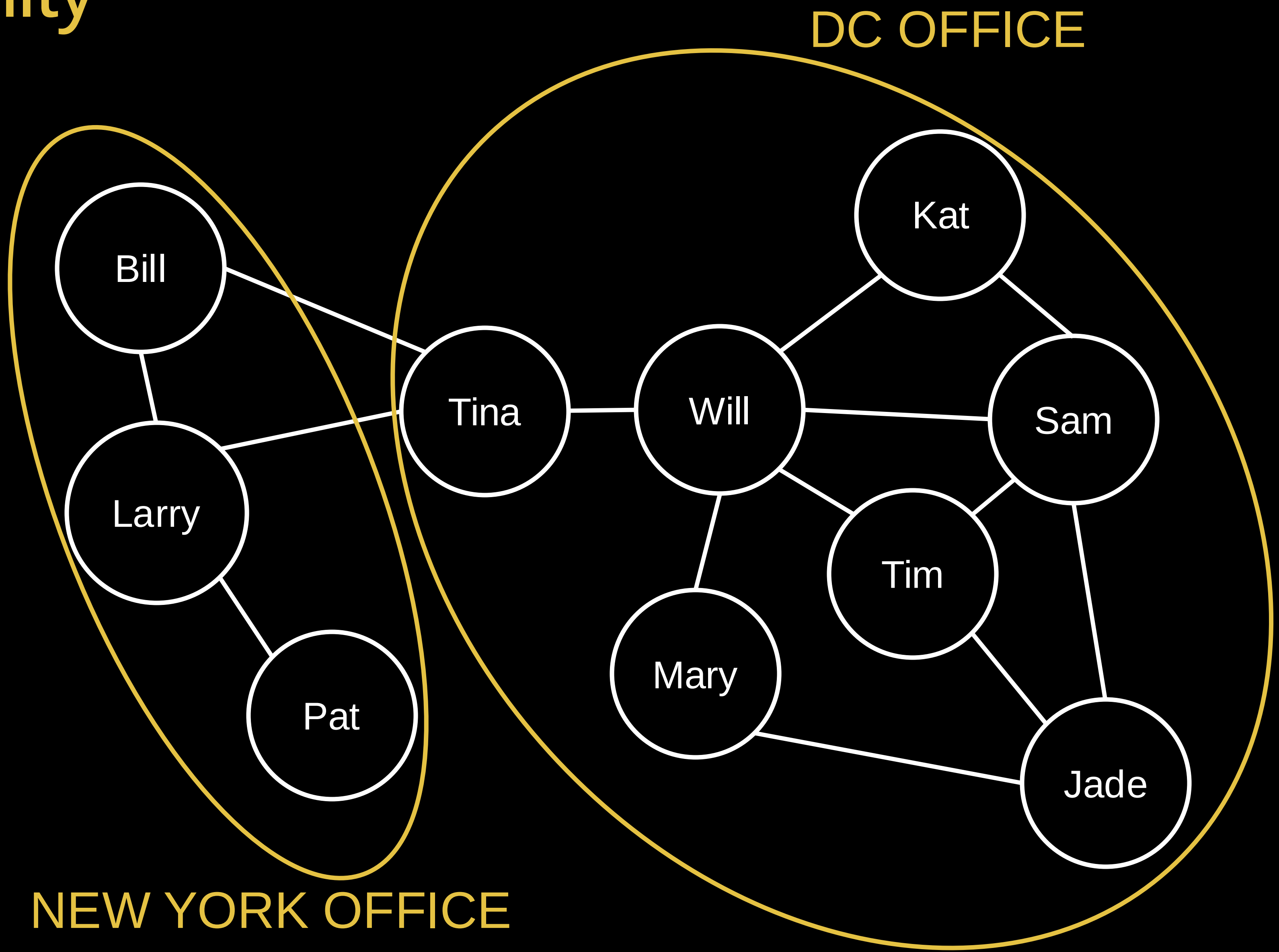
Diffusion and Adoption

- Diffusion and adoption add a sense of time to a sociogram.
- How long does it take for the entire network to **adopt** an idea based on initial location?
- Three other concepts are heavily related to diffusion and adoption:
  1. Proximity
  2. Prestige
  3. Social Conformity

Proximity


DC OFFICE
Bill
Tina
Larry
Pat
Kat
Will
Sam
Tim
Mary
Jade
NEW YORK OFFICE

Proximity


DC OFFICE
Kat
Bill
Tina
Will
Sam
Larry
Tim
Mary
Pat
Jade
NEW YORK OFFICE

Prestige & Social Conformity

- Prestige and Social Conformity are closely related.
- Individuals who epitomize social norms and values of a group that are perceived by others to be valuable have **prestige**.
- **Social conformity** allows people to validate their own sense of self-worth in a group.
  - Example: Will is the prototypical DC office type employee, so Jade wants to be like Will.

?

ACCOUNTING FOR TIME

OPTIONAL SELF STUDY

Aggregating Time

- Certain transactions are expected to occur at certain times.
- Anomalies might be detected outside of “normal” hours.
- Dealing with time averages and confidence intervals can be tricky.

Aggregating Time

- What is the arithmetic average between 1 and 23?

Aggregating Time

- What is the arithmetic average between 1 and 23? **12!**

Aggregating Time

- What is the arithmetic average between 1 and 23? **12!**

- What is the arithmetic average between 1:00AM and 11:00PM? **NOON?**

Aggregating Time

- What is the arithmetic average between 1 and 23? **12!**
- What is the arithmetic average between 1:00AM and 11:00PM? **NOON?**
- What is the **periodic** average between 1:00AM and 11:00PM? **MIDNIGHT!**

Arithmetic Mean

```
set.seed(12345)
timestamp <- as.POSIXlt("2020-02-03 00:30:00")
             + rnorm(1000, 0, 60*60*4)
head(timestamp)
```

```
## [1] "2020-02-03 02:50:31 EST" "2020-02-03 03:20:16 EST"
## [3] "2020-02-03 00:03:46 EST" "2020-02-02 22:41:09 EST"
## [5] "2020-02-03 02:55:24 EST" "2020-02-02 17:13:41 EST"
```

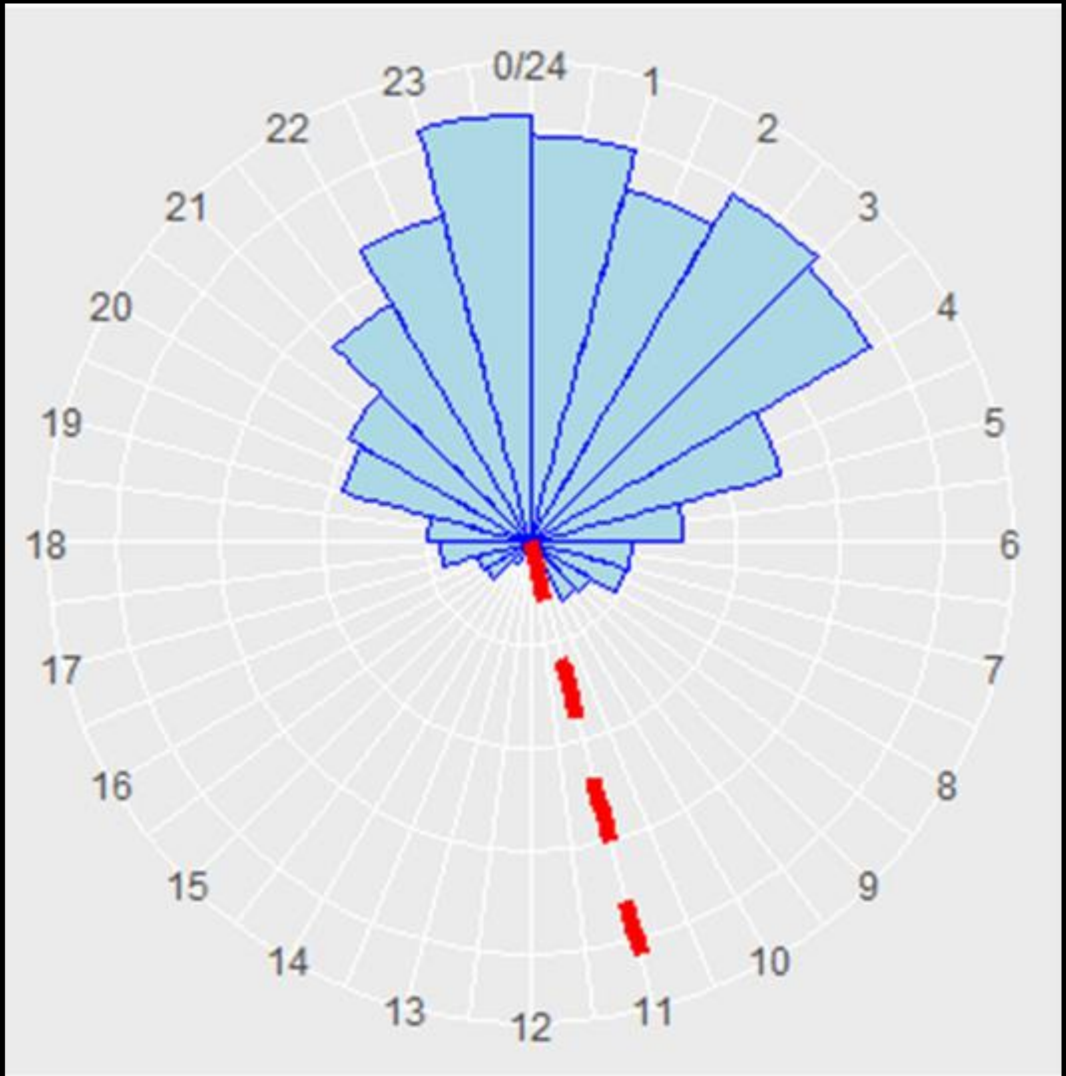
Arithmetic Mean

```
timestamp_hms <- strftime(timestamp, format = "%H:%M:%S")
ts <- as.numeric(hms(timestamp_hms))/3600
mean_a <- mean(ts)

clock <- ggplot(data.frame(ts), aes(x = ts)) +
         geom_histogram(breaks = seq(0, 24), colour = "blue",
                        fill = "lightblue") +
         coord_polar() +
         scale_x_continuous("", limits = c(0, 24),
                            breaks = seq(0, 24))

clock + geom_vline(xintercept = mean_a,
                   linetype = 2, color = "red", size = 2)
```

Arithmetic Mean



Periodic Mean

```
ts <- circular(ts, units = "hours", template = "clock24")
head(ts)
```

```
## Circular Data:
## Type = angles
## Units = hours
## Template = clock24
## Modulo = asis
## Zero = 1.570796
## Rotation = clock
## [1]  2.84194444  3.33777778  0.06277778 22.68583333  2.923333
33 17.22805556
```

Periodic Mean

```
estimates <- mle.vonmises(ts)
mean_p <- estimates$mu %% 24
concentration <- estimates$kappa

clock + geom_vline(xintercept = mean_p,
                   linetype = 2, color = "red", size = 2)
```

Periodic Mean