WHY YOU CAN'T SIT WITH US

UNDERSTANDING NETWORK ANALYSIS IN PYTHON WITH MEAN GIRLS

Richard Harris - Data Scientist - Braintree Payments

ABOUT ME (AND A WARNING!)

WHY THIS MATTERS

WHAT WE'RE DOING TODAY:

- Basics of Network Analysis
- Introduction to NetworkX
  - NetworkX / Matplotlib
  - Standard Library / Techniques (list comprehension, etc.)
- Application
  - Creating the Network
  - You've Probably Been Personally Assaulted by Regina George
  - Why You Can't Sit with Us
  - (Glen Coco)
  - (Why Fetch Won't Happen)

MEAN GIRLS

WHAT IS THE DATA?

- All lines spoken in the film
- Edges drawn between two characters if one speaks directly to another one
- Voiceovers, speaking to a group, or cases where it's not clear are coded only with speaker
- Data was collected by hand, through repeated viewings of the film.

A QUICK INTRODUCTION TO NETWORK ANALYSIS

BASICS

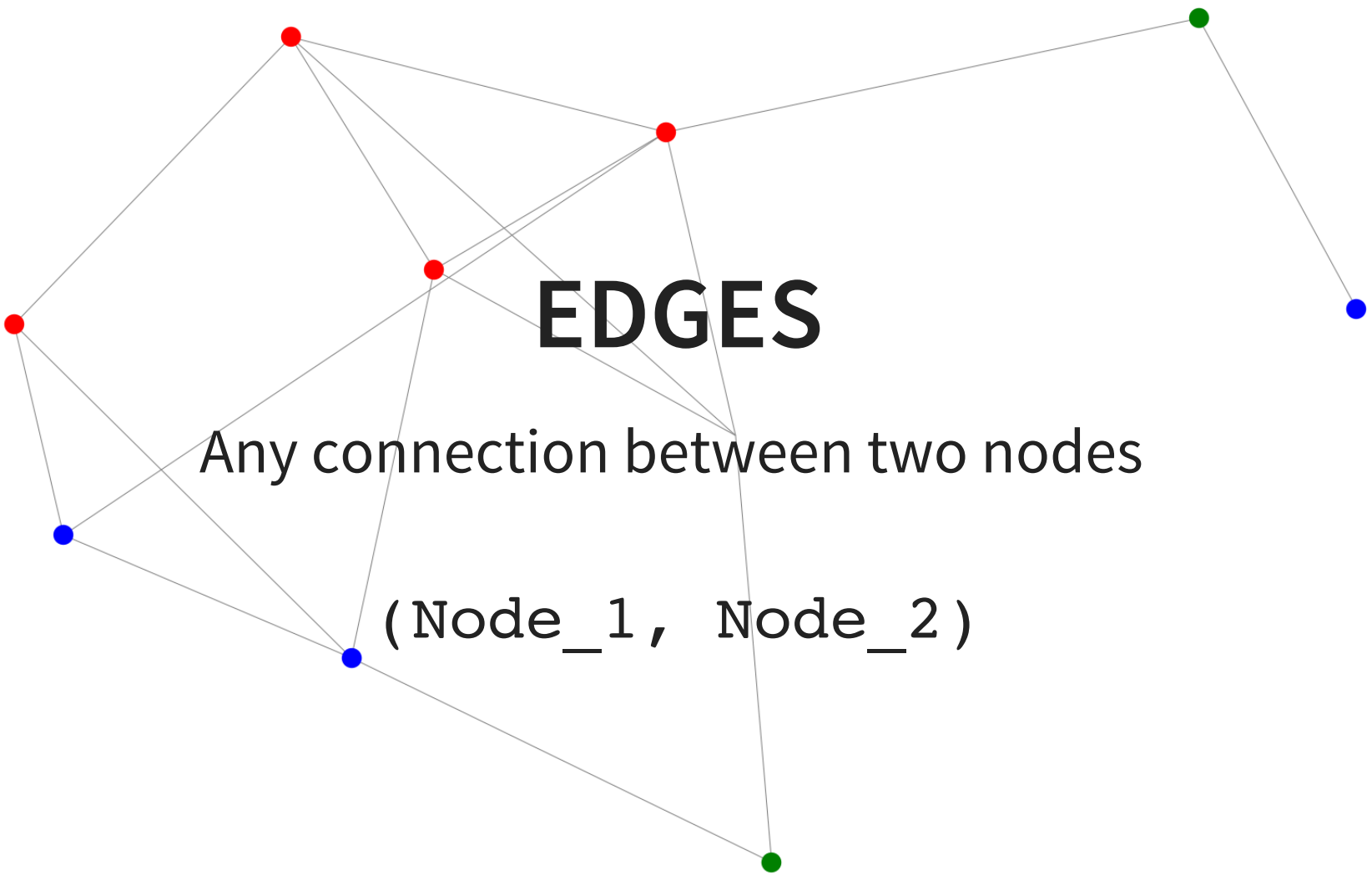
Network Analysis is based off two things:

- Nodes
- Edges

NODES

- People
- Locations
- Genes
- Websites

NODES CAN ALSO HAVE ATTRIBUTES



EDGES

Any connection between two nodes

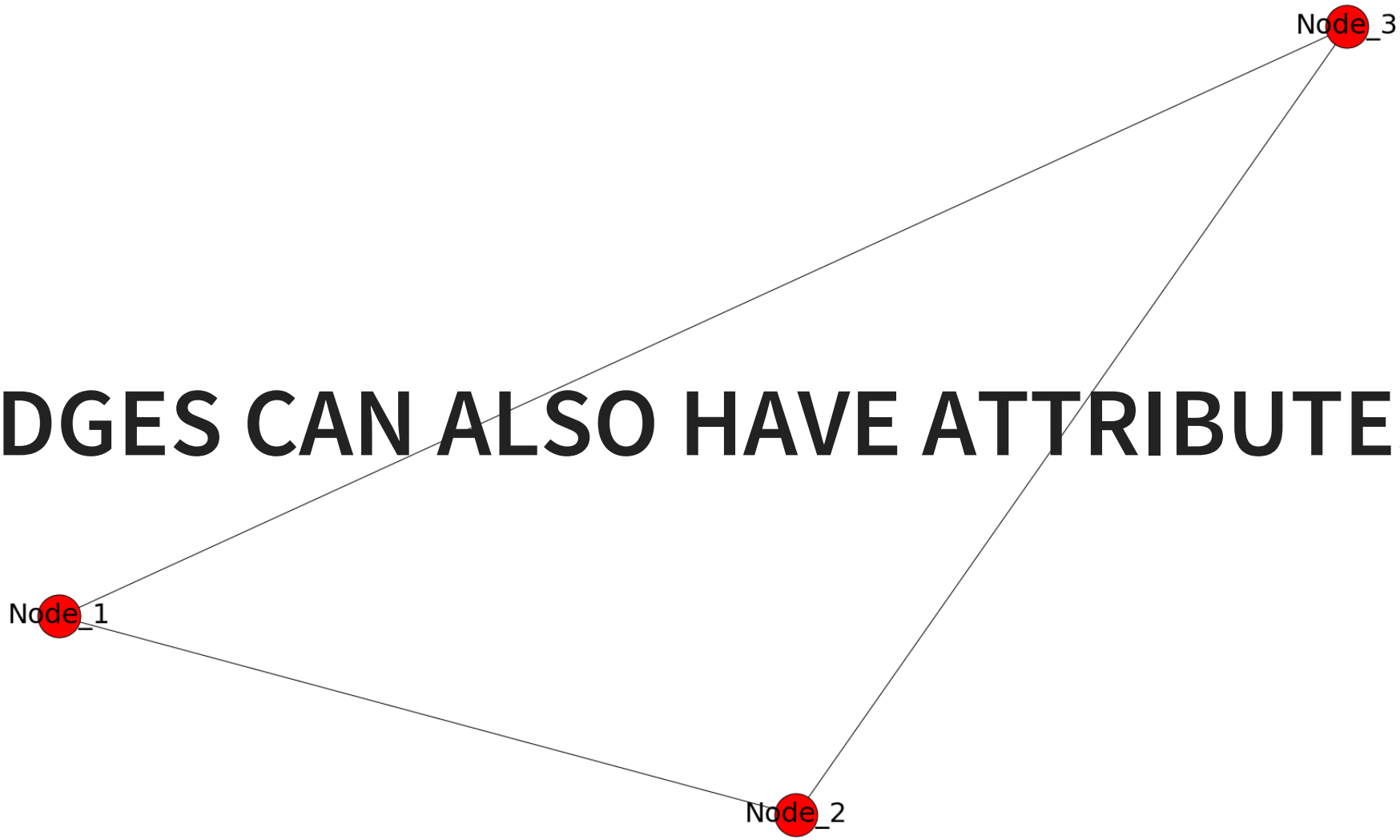
`(Node_1, Node_2)`

EDGES

Nodes can have any number of edges

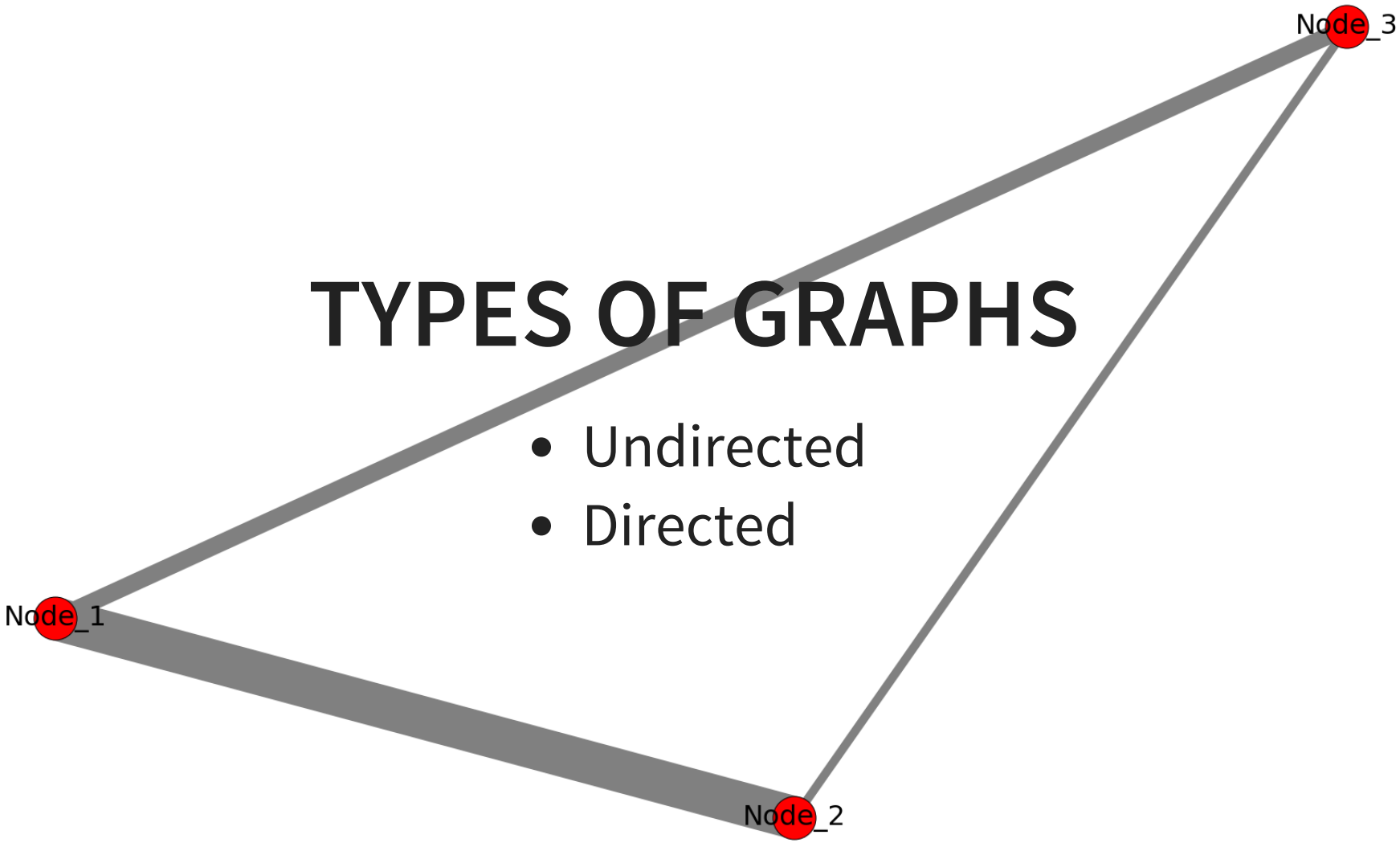
```
(Node_1, Node_2)
(Node_1, Node_3)
(Node_2, Node_3)
```

EDGES CAN ALSO HAVE ATTRIBUTES



TYPES OF GRAPHS

- Undirected
- Directed

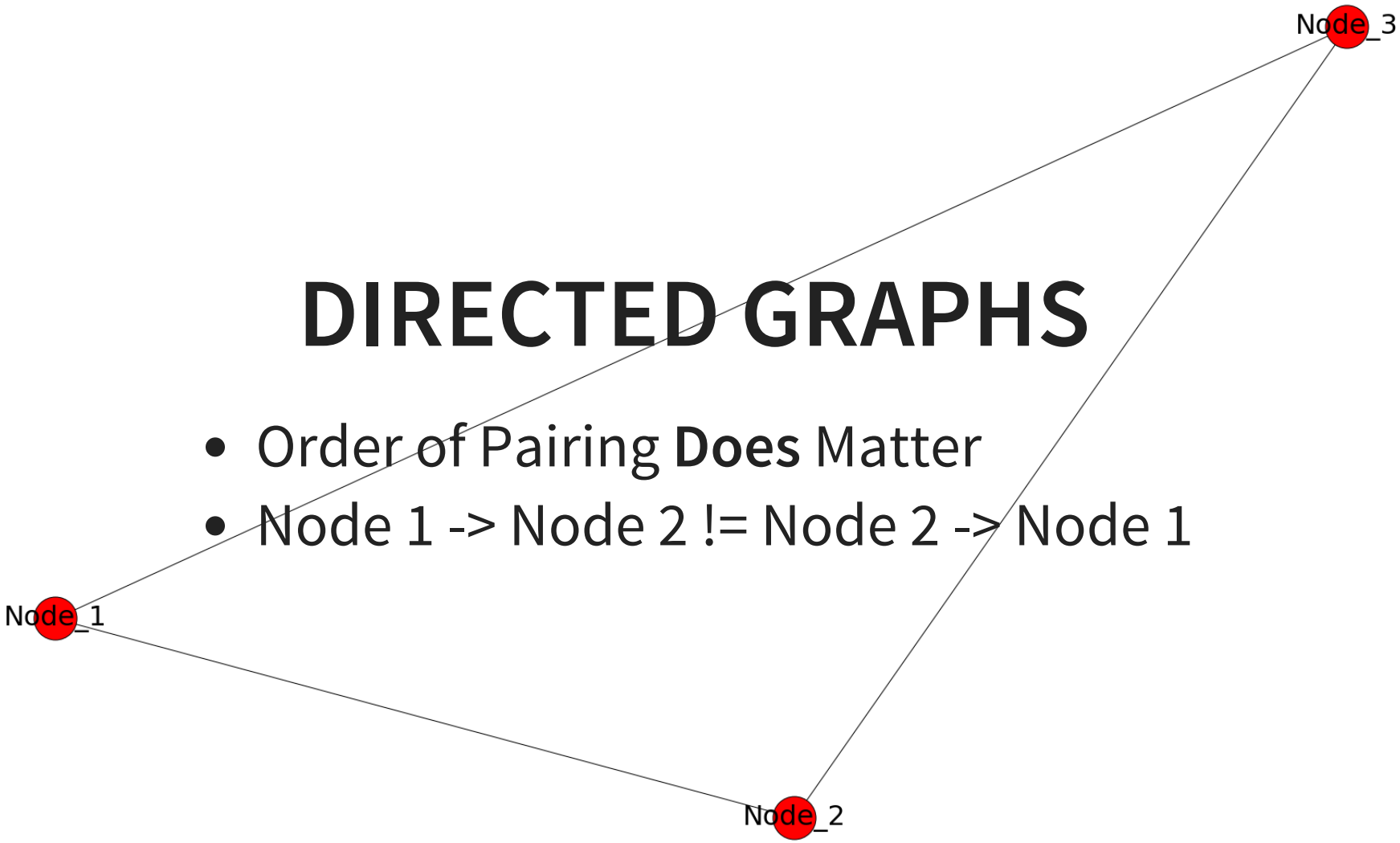


UNDIRECTED GRAPHS

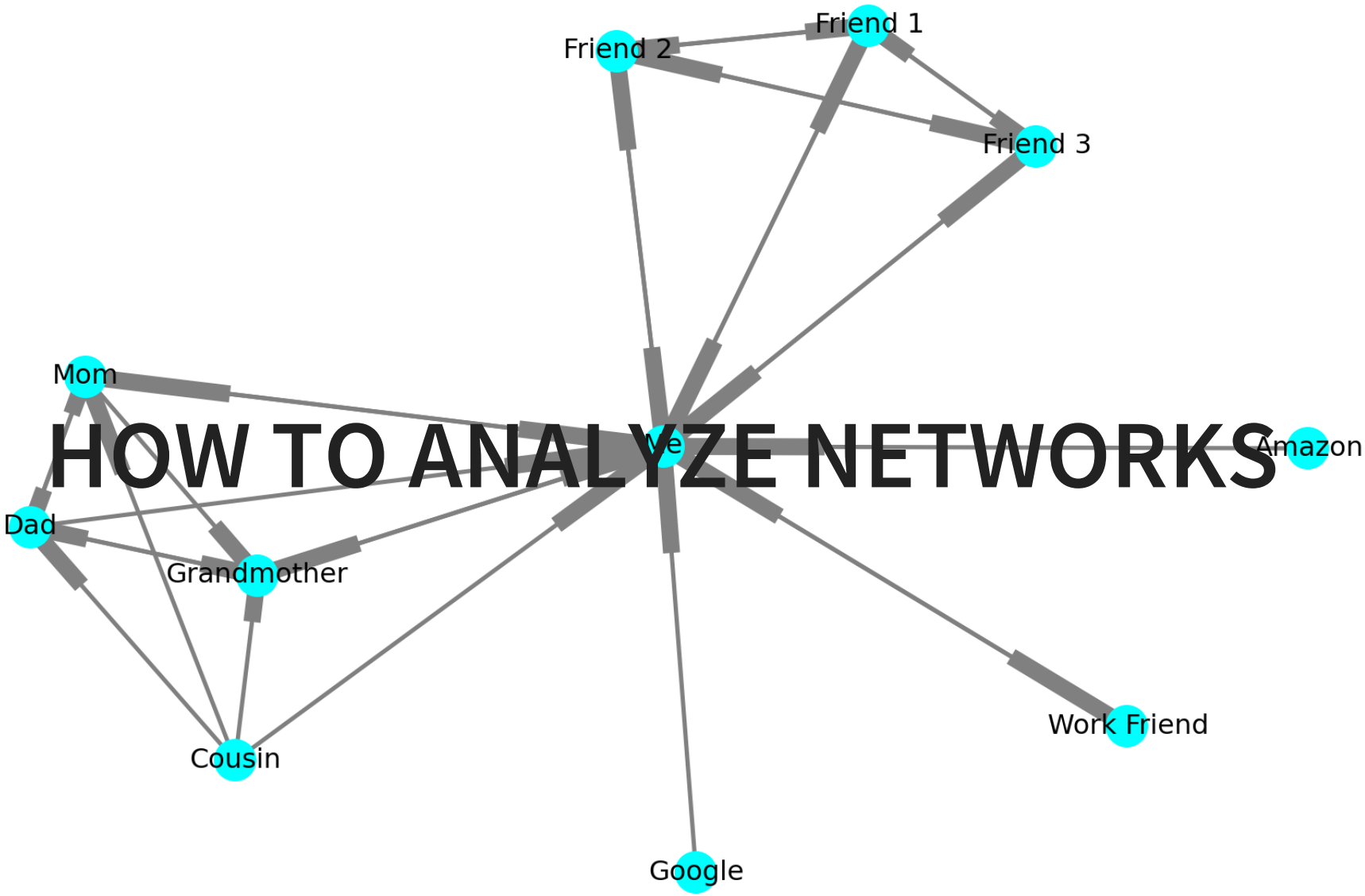
- Order of Pairing Doesn't Matter
  - Node 1 -> Node 2 == Node 2 -> Node 1
- Maximum number of edges (without self-loops): $\frac{n(n-1)}{2}$

DIRECTED GRAPHS

- Order of Pairing **Does** Matter
- Node 1 -> Node 2 != Node 2 -> Node 1



DIRECTED GRAPHS

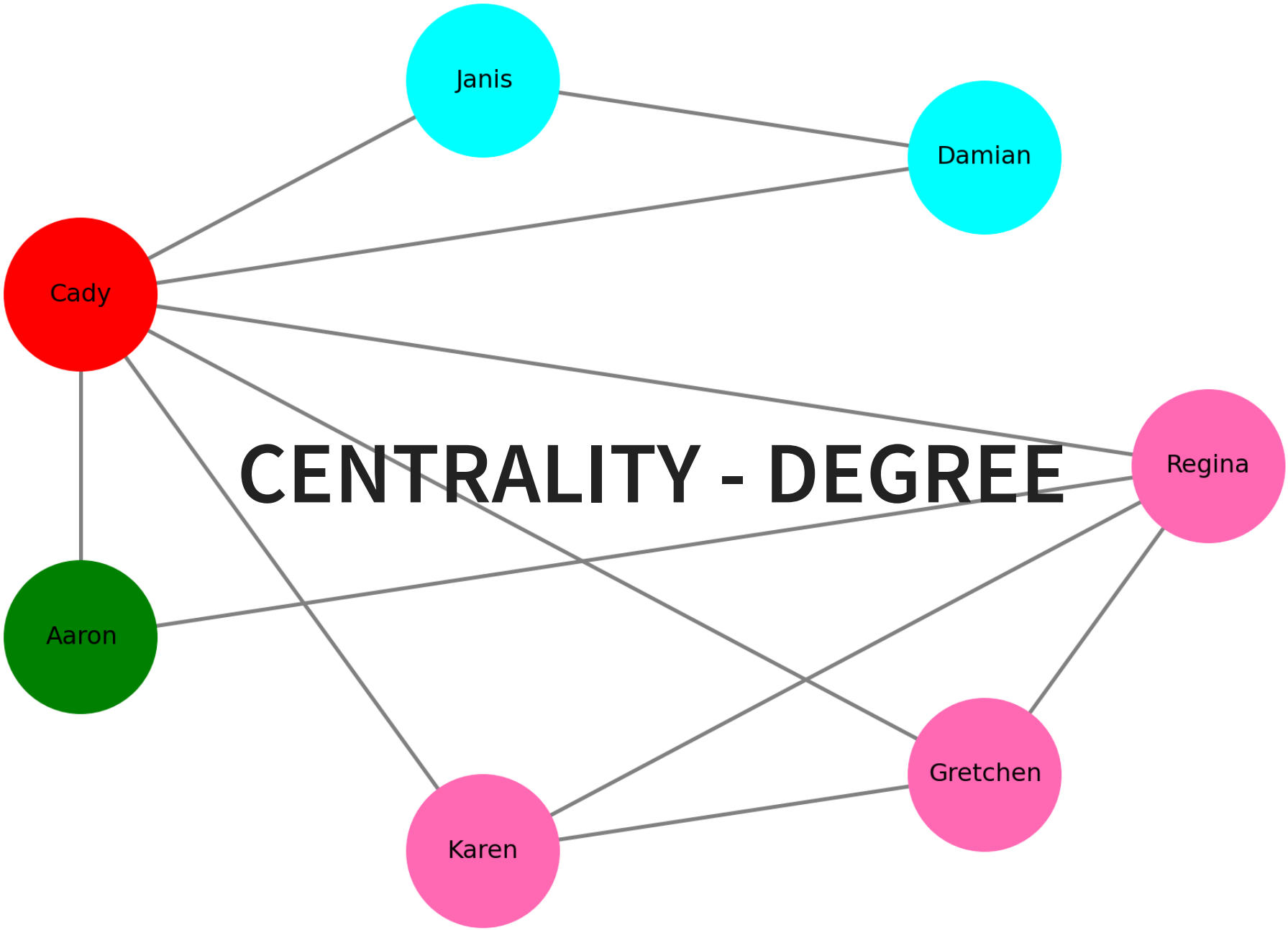

HOW TO ANALYZE NETWORKS
Friend 1
Friend 2
Friend 3
Mom
Me
Amazon
Dad
Grandmother
Work Friend
Cousin
Google

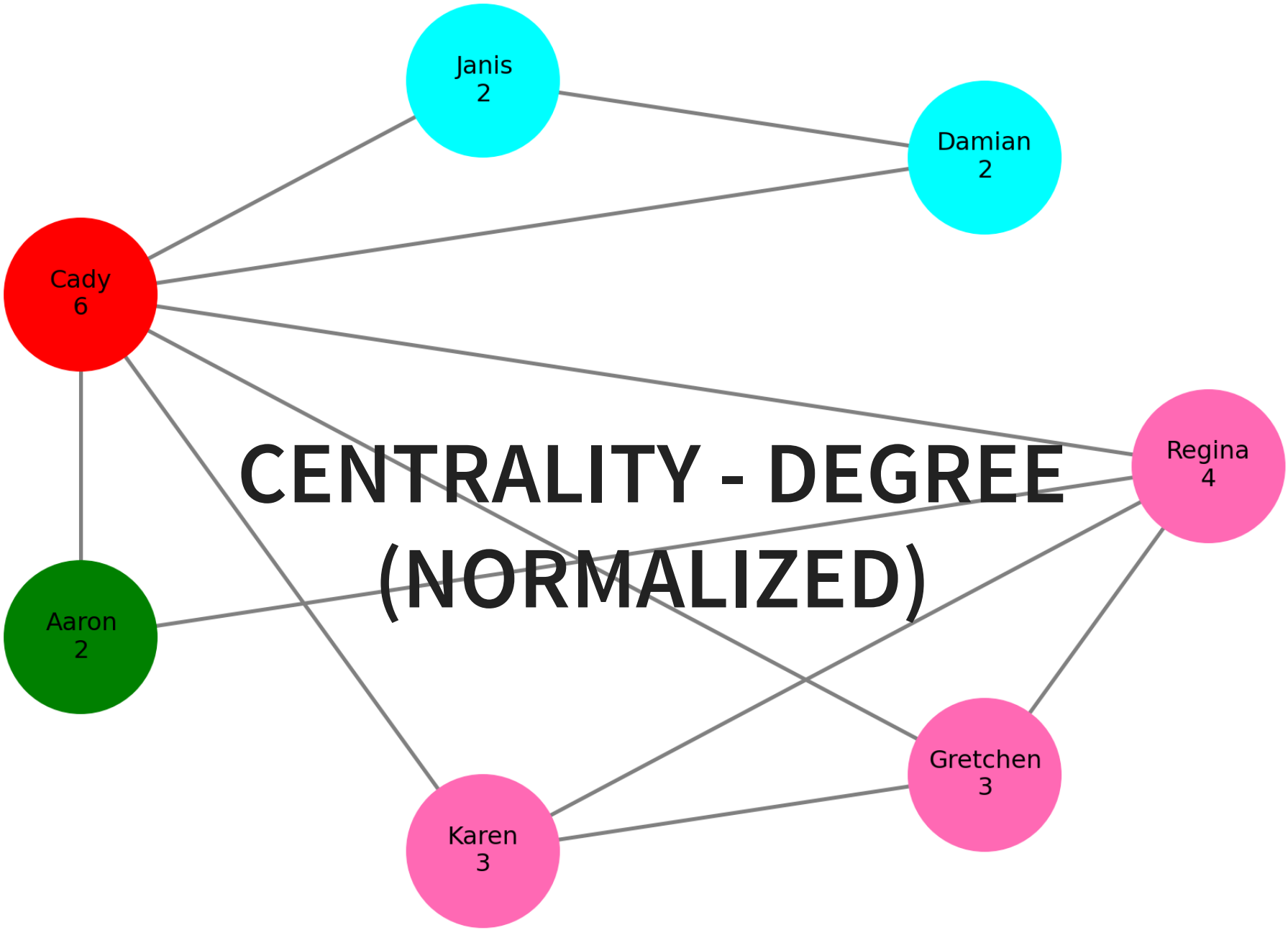
KEY FEATURES

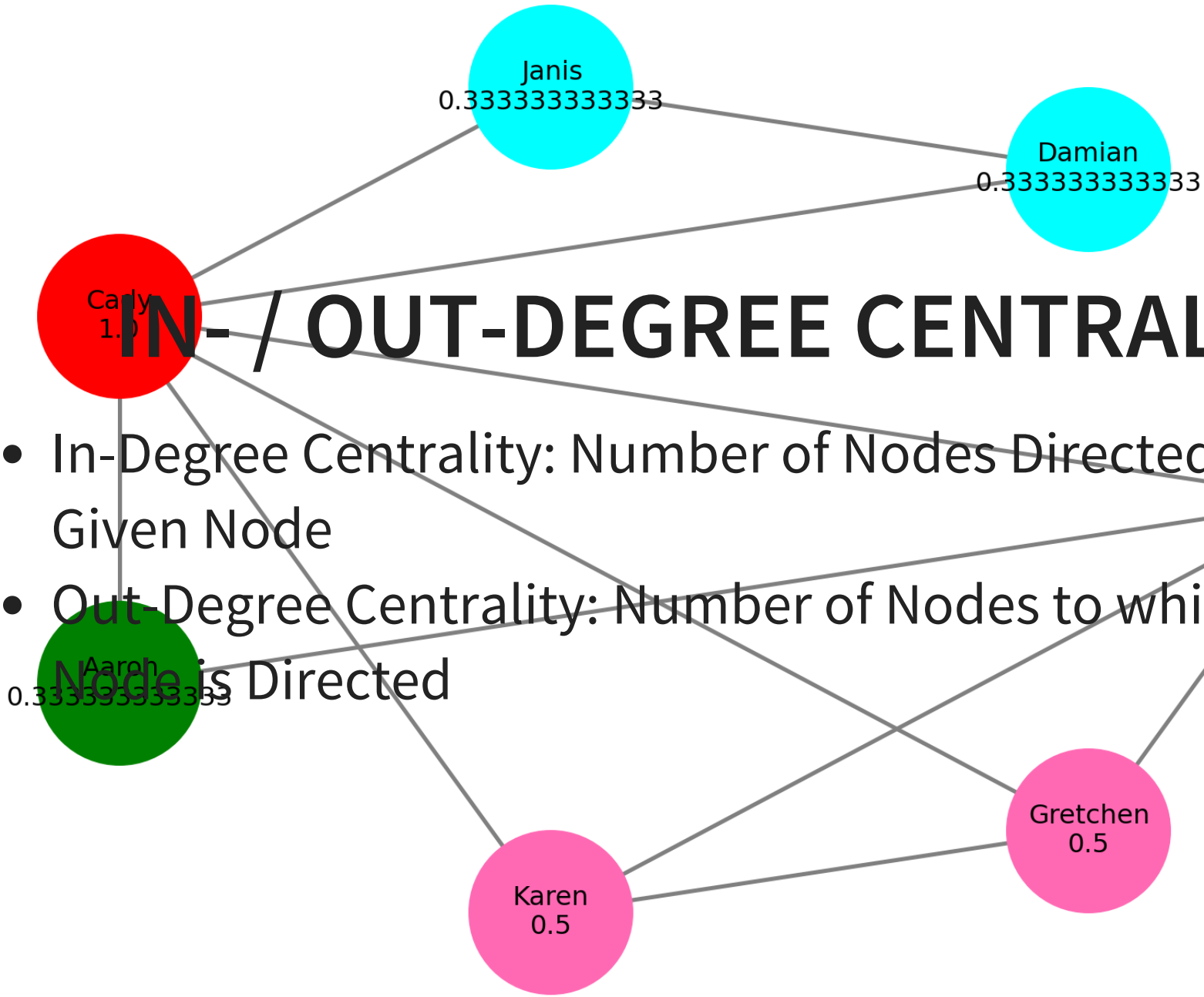
- Centrality
  - Degree
  - Closeness
  - Betweeness
- Connectedness
- Brokers

DEGREE CENTRALITY

...or how many people are you directly connected with?


CENTRALITY - DEGREE
Janis
Damian
Cady
Regina
Aaron
Gretchen
Karen

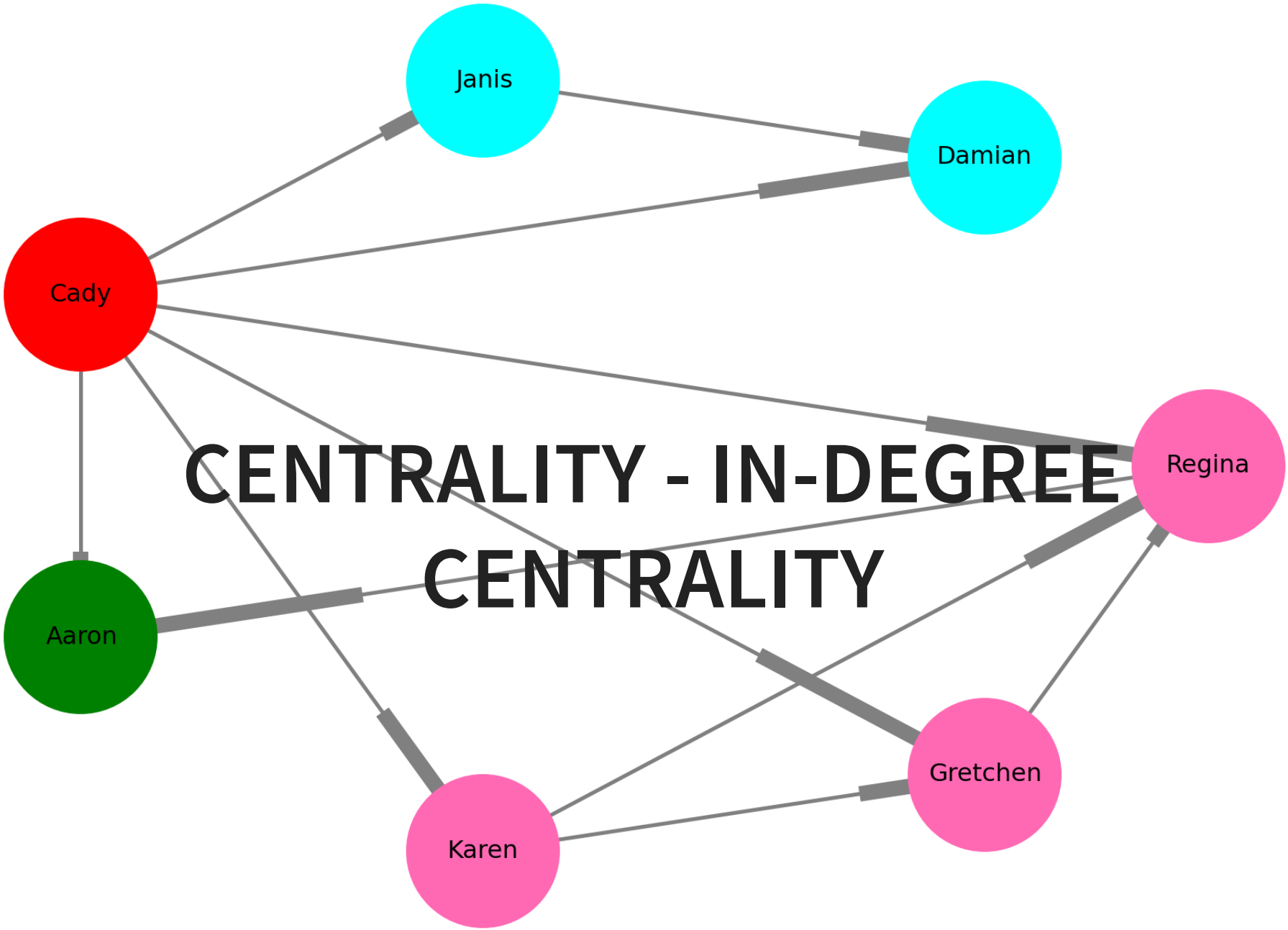

Janis
2
Damian
2
Cady
6
Regina
4
CENTRALITY - DEGREE
(NORMALIZED)
Aaron
2
Gretchen
3
Karen
3

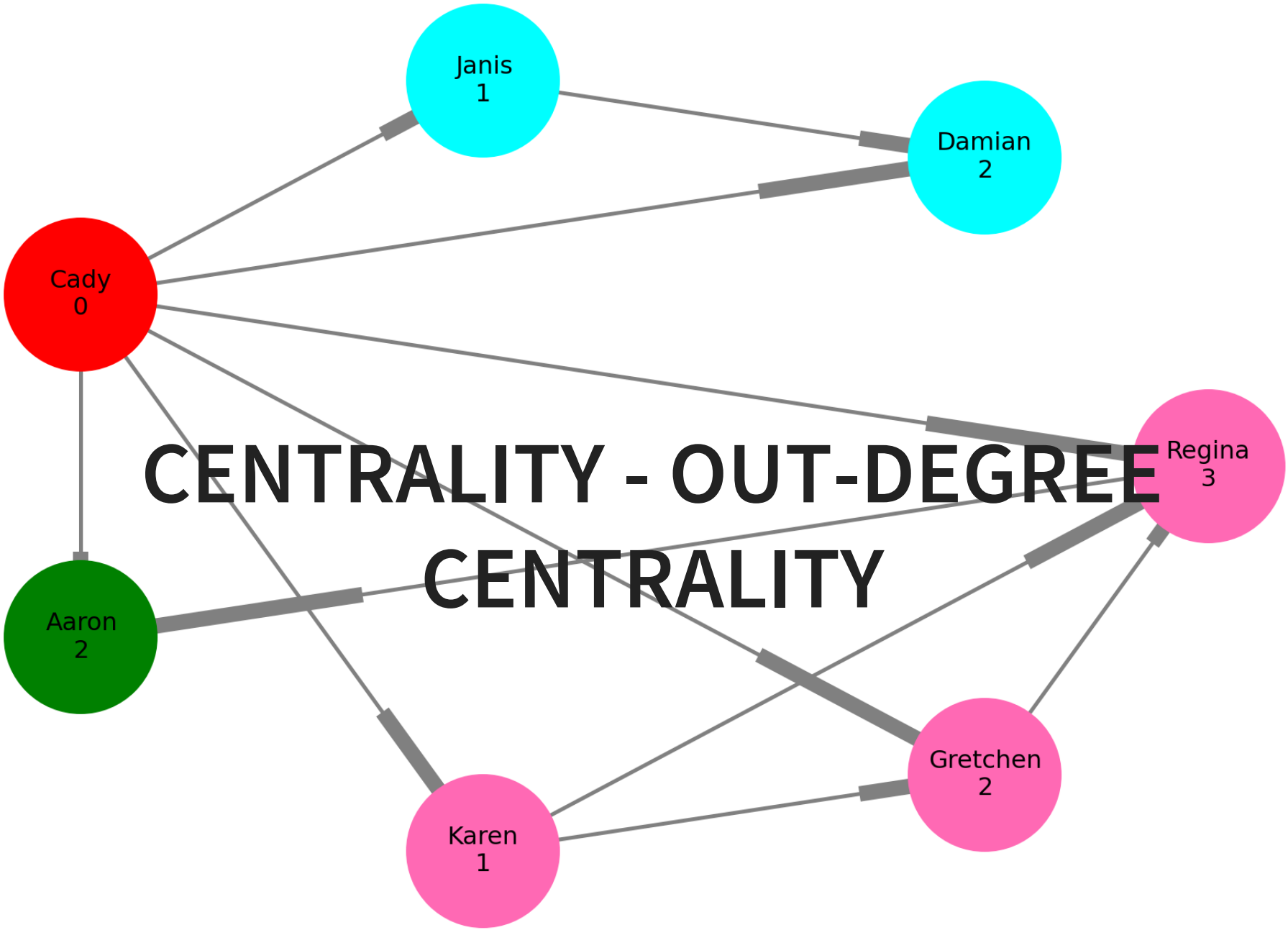


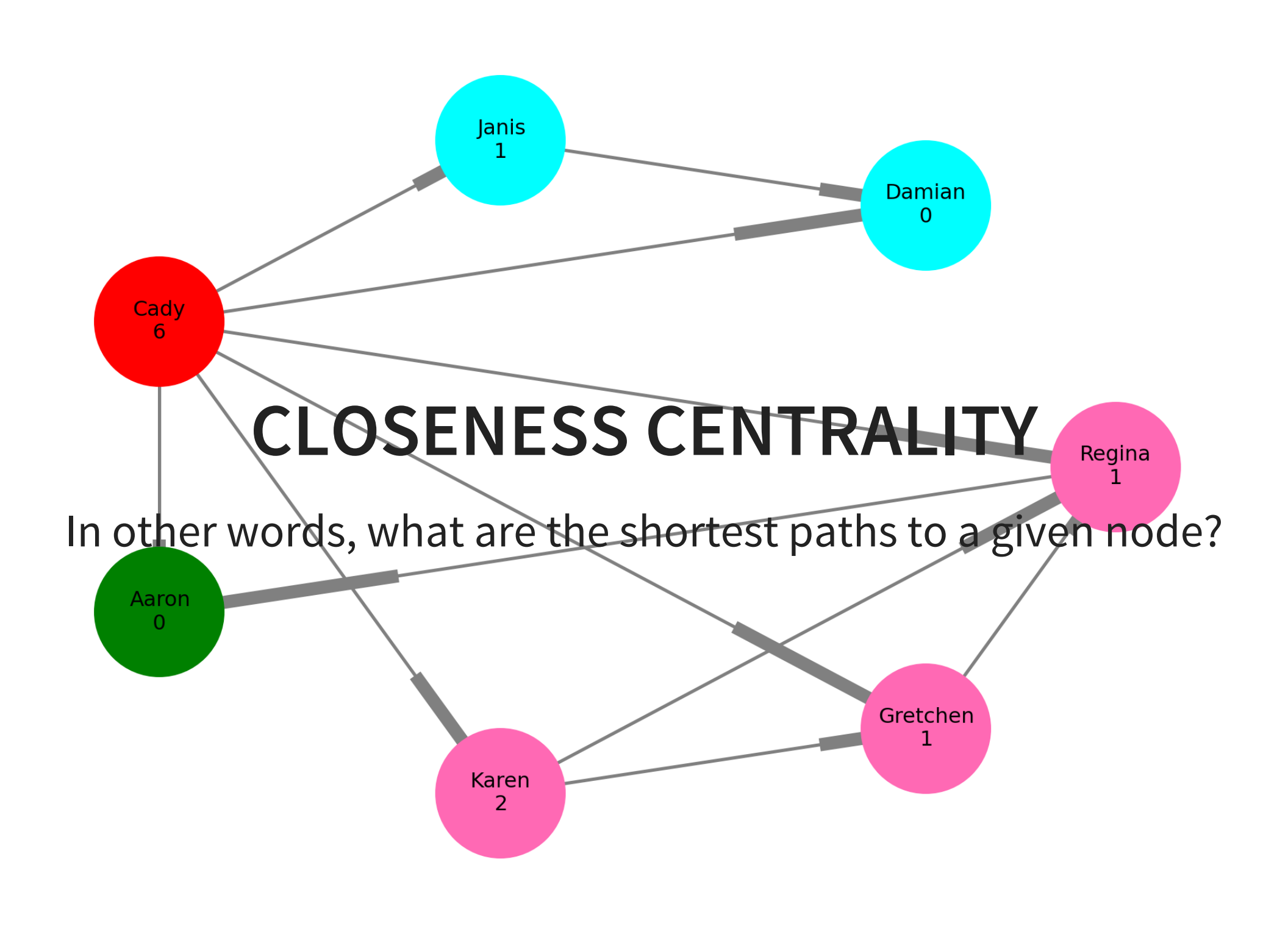
IN- / OUT-DEGREE CENTRALITY

- In-Degree Centrality: Number of Nodes Directed towards a Given Node
- Out-Degree Centrality: Number of Nodes to which a Given Node is Directed

CENTRALITY - DIRECTED GRAPH

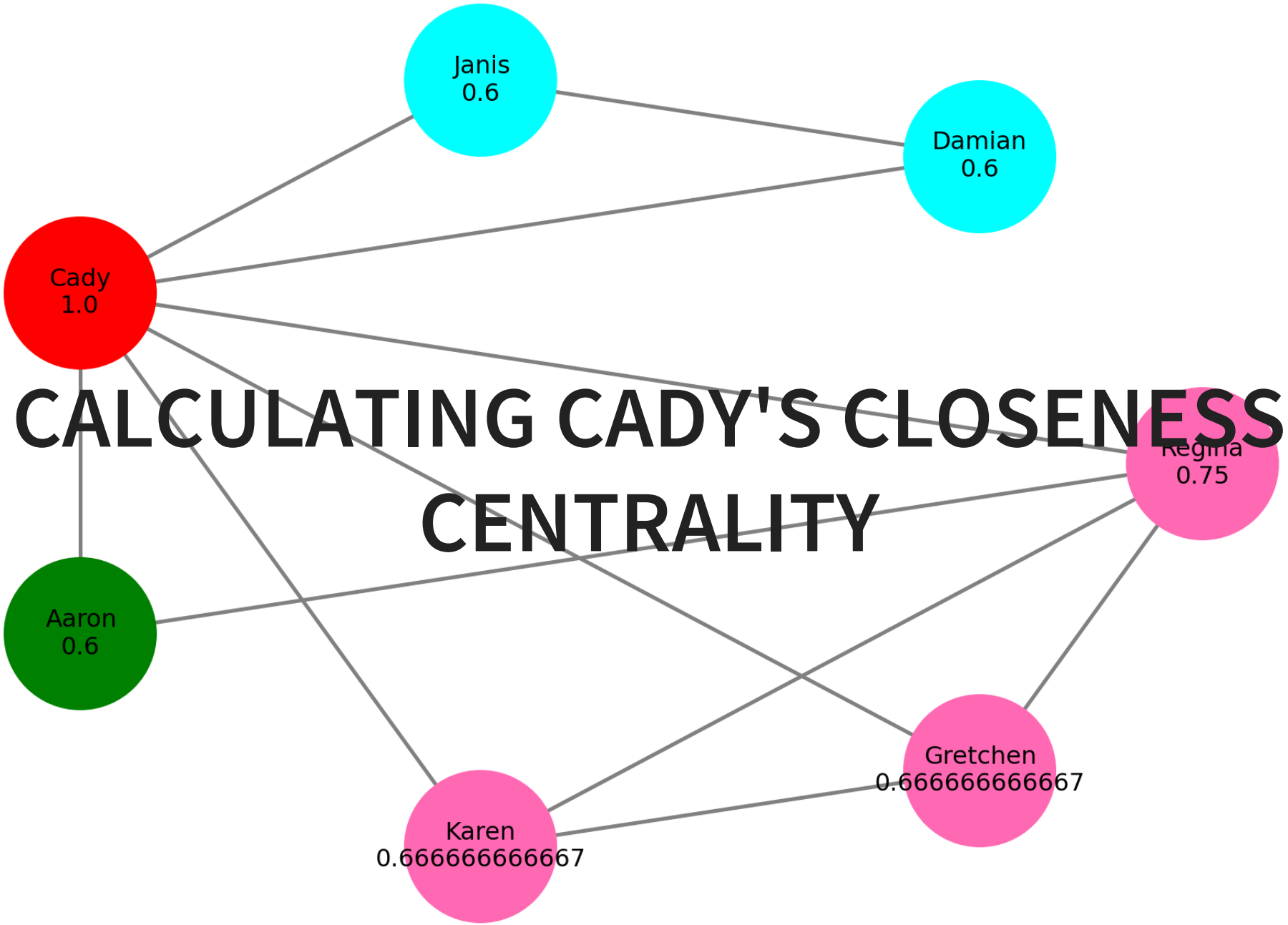

CENTRALITY - IN-DEGREE CENTRALITY
Janis
Damian
Cady
Regina
Aaron
Gretchen
Karen

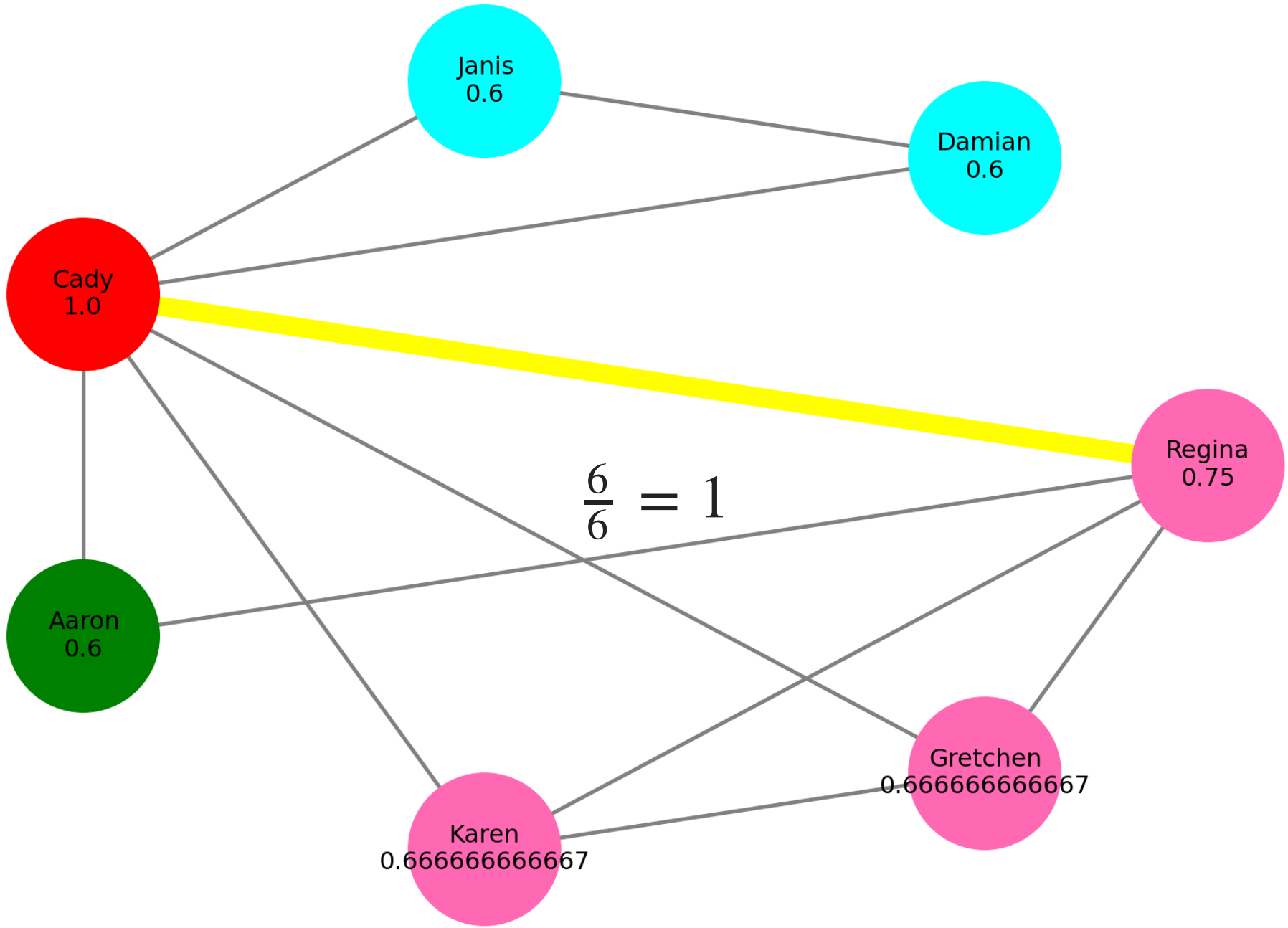

CENTRALITY - OUT-DEGREE CENTRALITY
Janis
1
Damian
2
Cady
0
Regina
3
Aaron
2
Gretchen
2
Karen
1



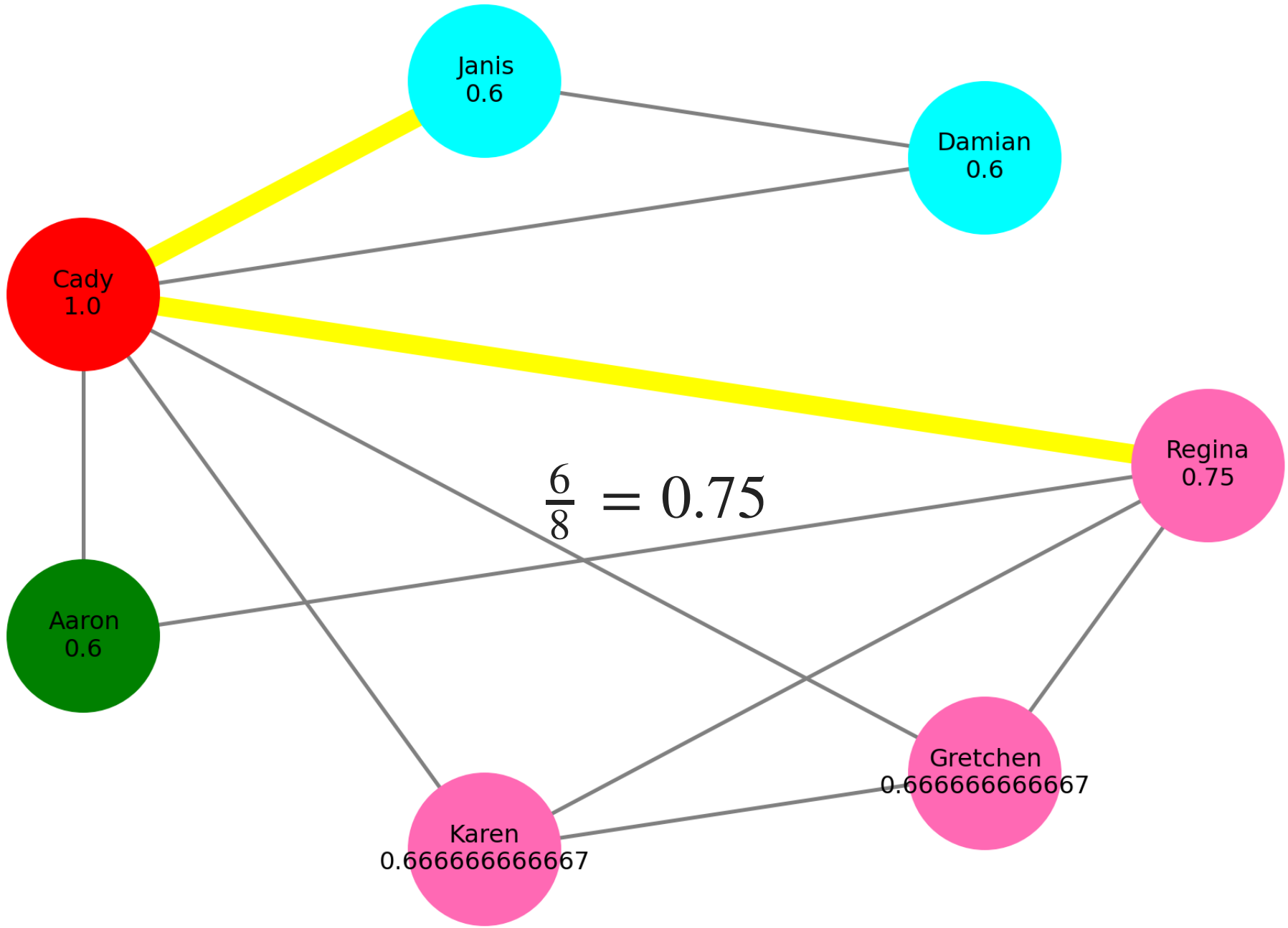
CLOSENESS CENTRALITY

In other words, what are the shortest paths to a given node?


Janis
0.6
Damian
0.6
Cady
1.0
CALCULATING CADY'S CLOSENESS CENTRALITY
Regina
0.75
Aaron
0.6
Gretchen
0.666666666667
Karen
0.666666666667

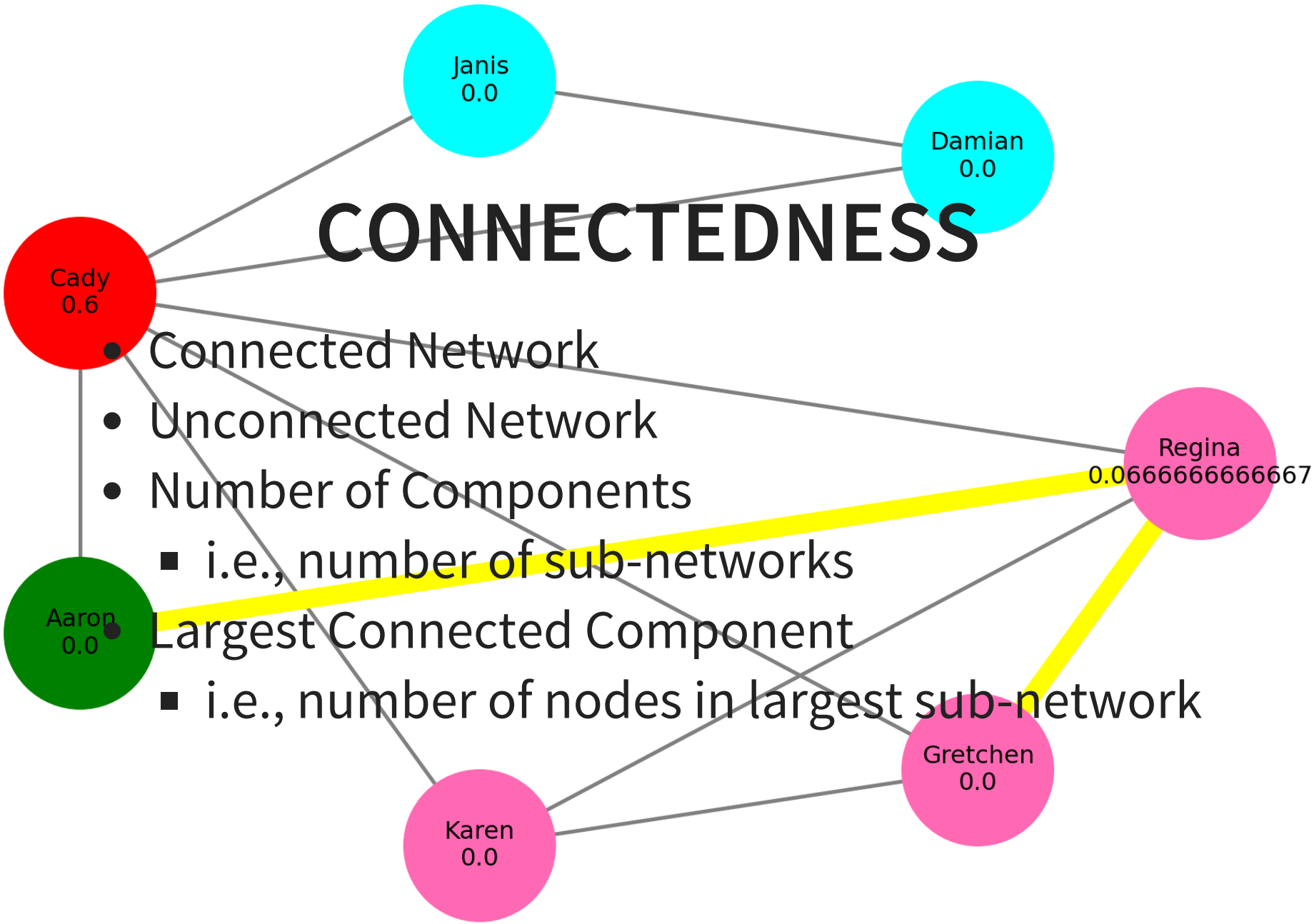

Janis
0.6
Damian
0.6
Cady
1.0
Regina
0.75
$\frac{6}{6} = 1$
Aaron
0.6
Gretchen
0.666666666667
Karen
0.666666666667

VERSUS REGINA'S CLOSENESS CENTRALITY


Janis
0.6
Damian
0.6
Cady
1.0
Regina
0.75
$\frac{6}{8} = 0.75$
Aaron
0.6
Gretchen
0.666666666667
Karen
0.666666666667

BETWEENESS CENTRALITY

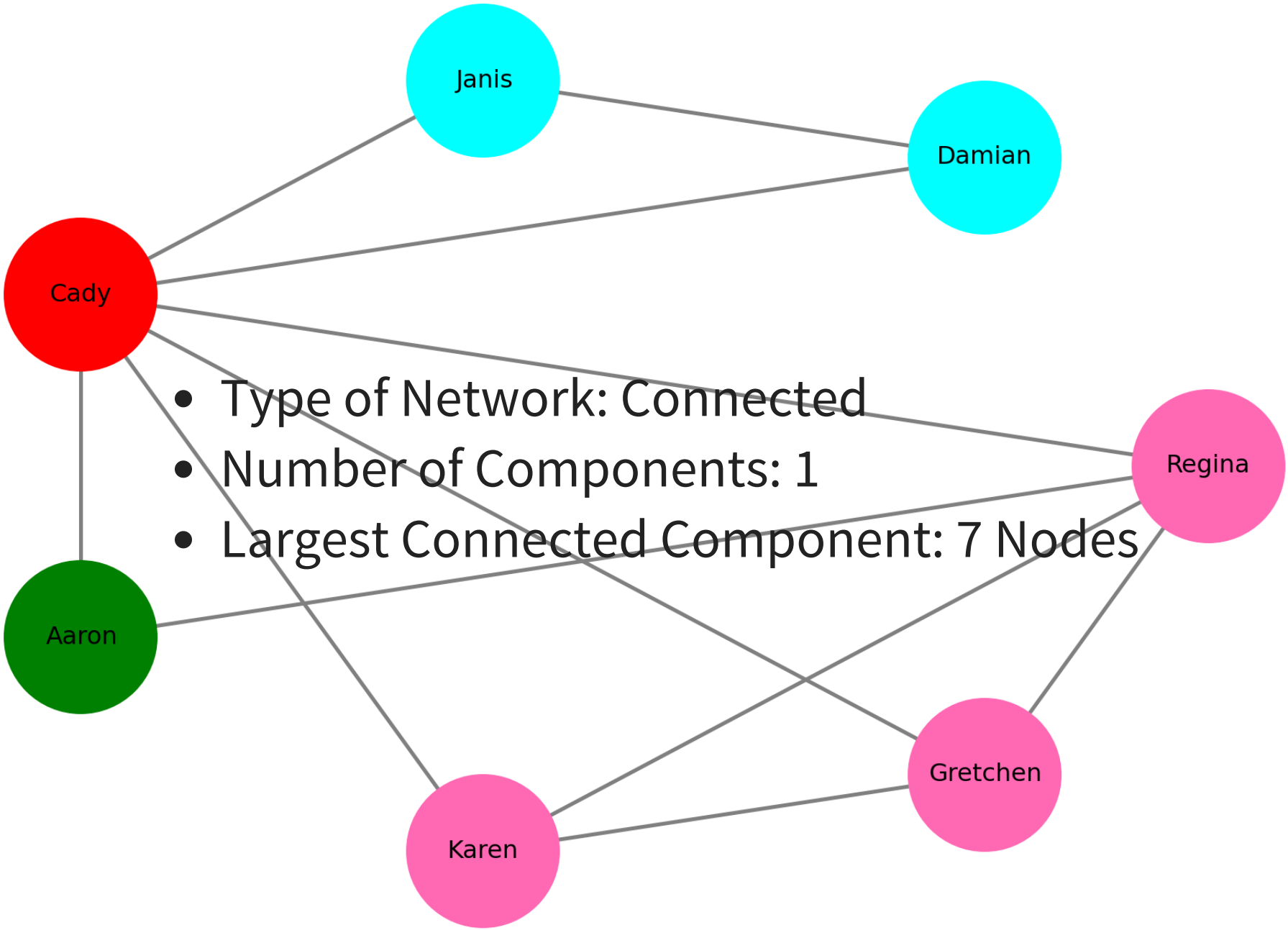
Of all shortest paths between two nodes, how many of them go through the node in question?



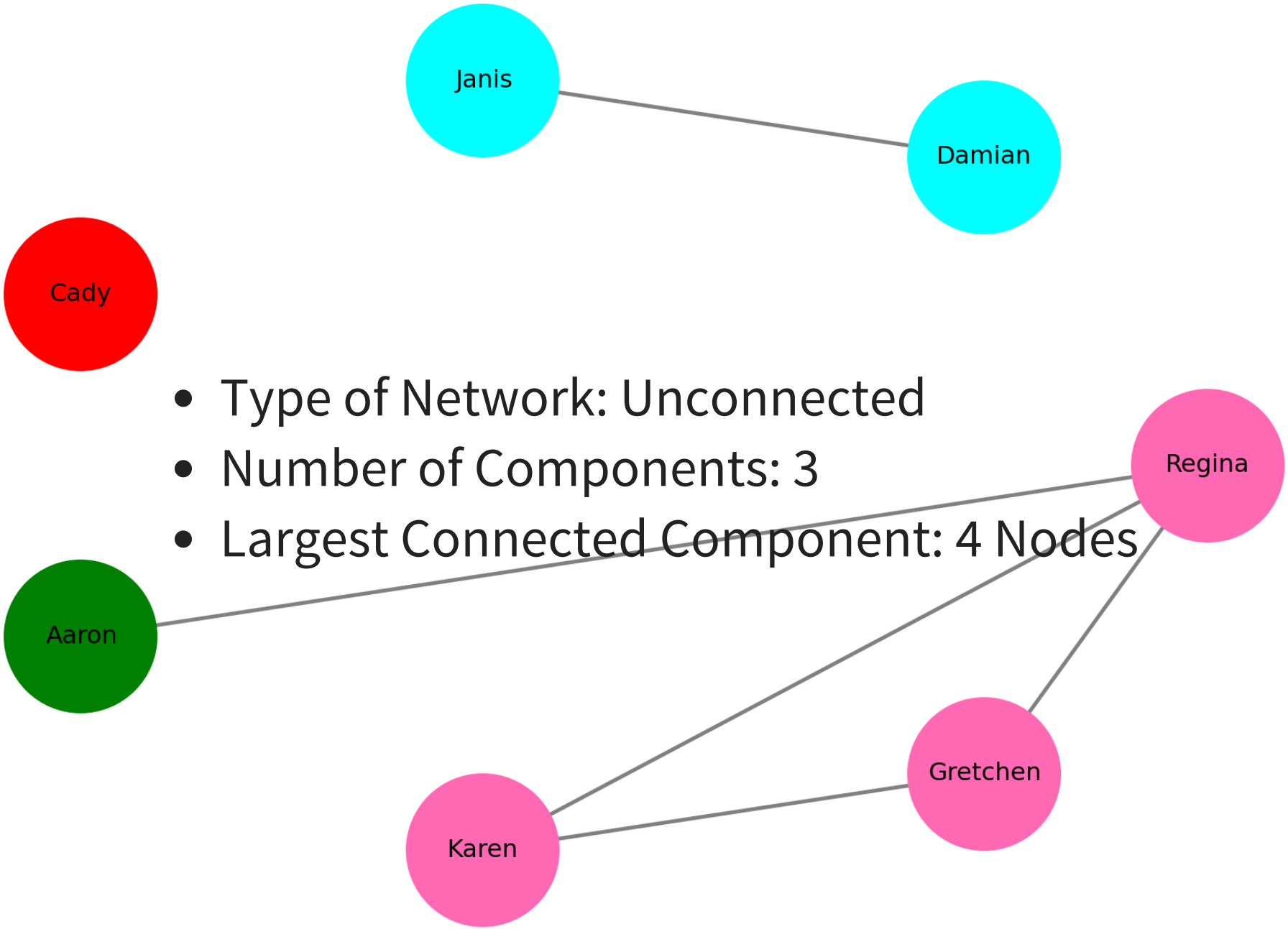
CONNECTEDNESS

- Connected Network
- Unconnected Network
- Number of Components
  - i.e., number of sub-networks
- Largest Connected Component
  - i.e., number of nodes in largest sub-network

CONNECTED NETWORK


Janis
Damian
Cady
Regina
Aaron
Gretchen
Karen
• Type of Network: Connected
• Number of Components: 1
• Largest Connected Component: 7 Nodes

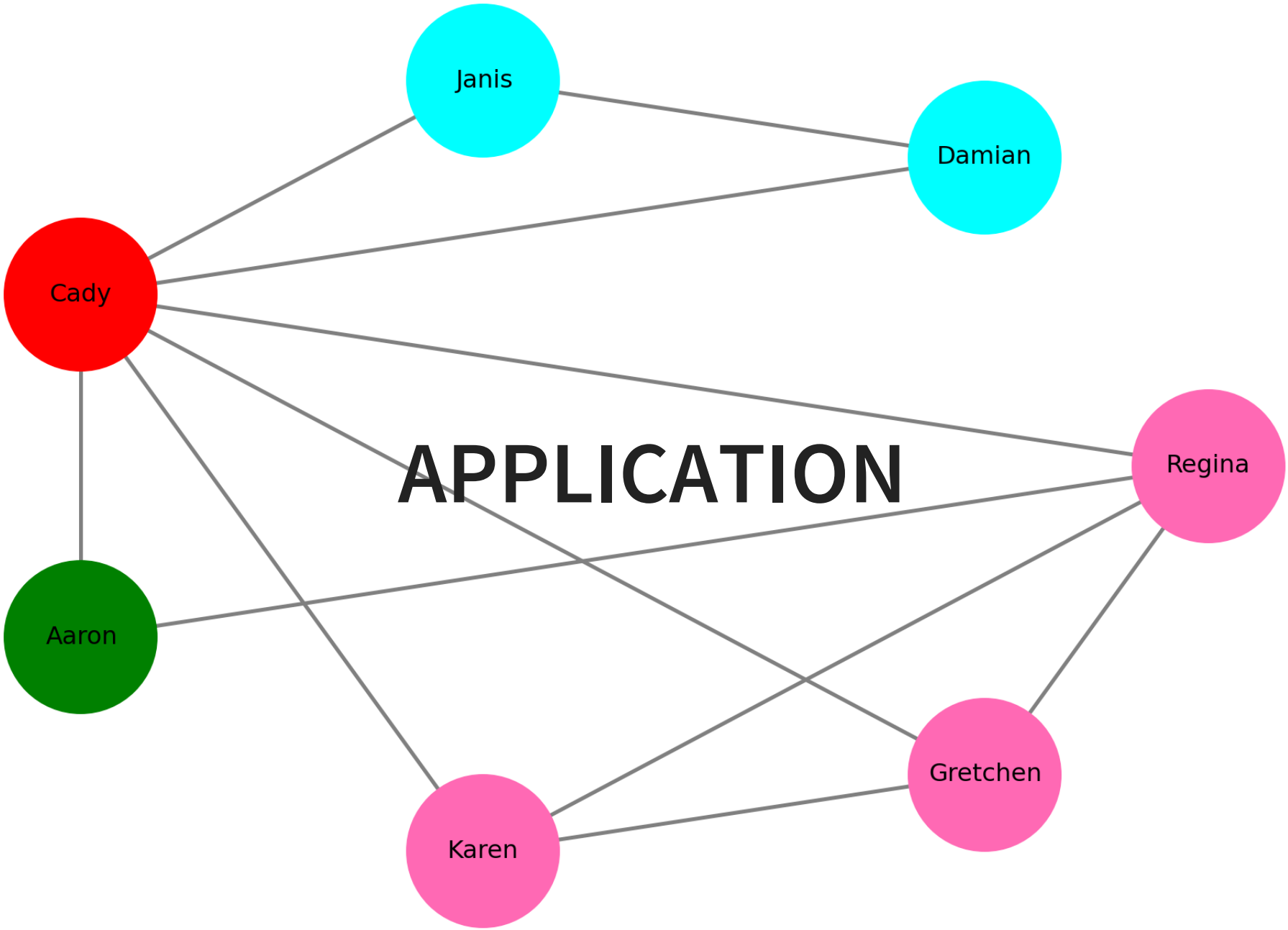
UNCONNECTED NETWORK



- Type of Network: Unconnected
- Number of Components: 3
- Largest Connected Component: 4 Nodes

BROKERS

- Connect Disperate Groups in Network
- How to Identify:
  - High Betweenness Centrality
  - Increase in Network Components if Removed


APPLICATION
Janis
Damian
Cady
Regina
Aaron
Gretchen
Karen