Interpreting the results of community detection algorithms

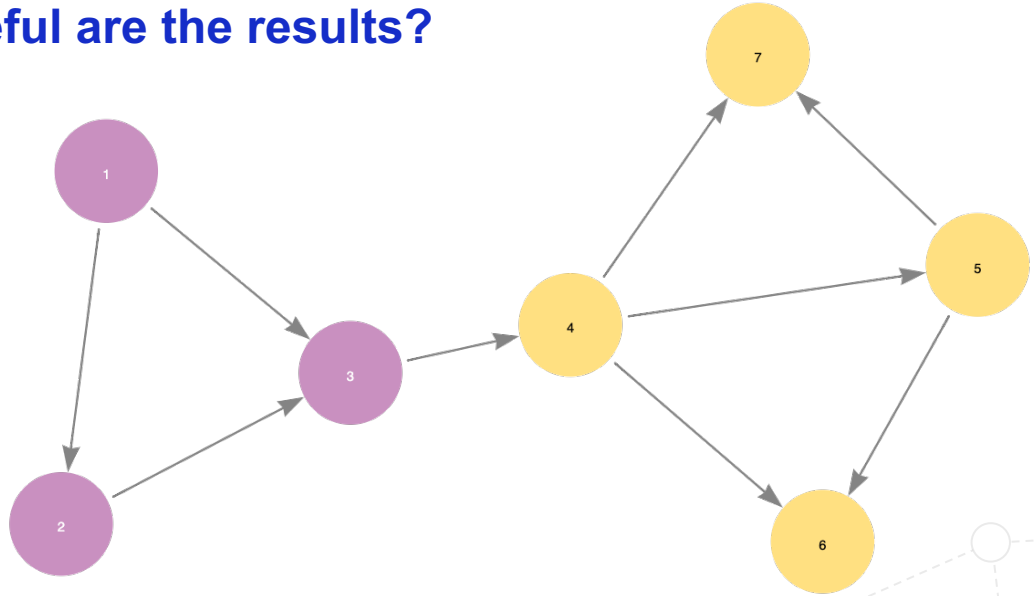
Nathan Smith,
Senior Data Scientist

graphconnect

Congratulations!

You have successfully run a community detection algorithm!

But how useful are the results?



What makes a partitioning useful?

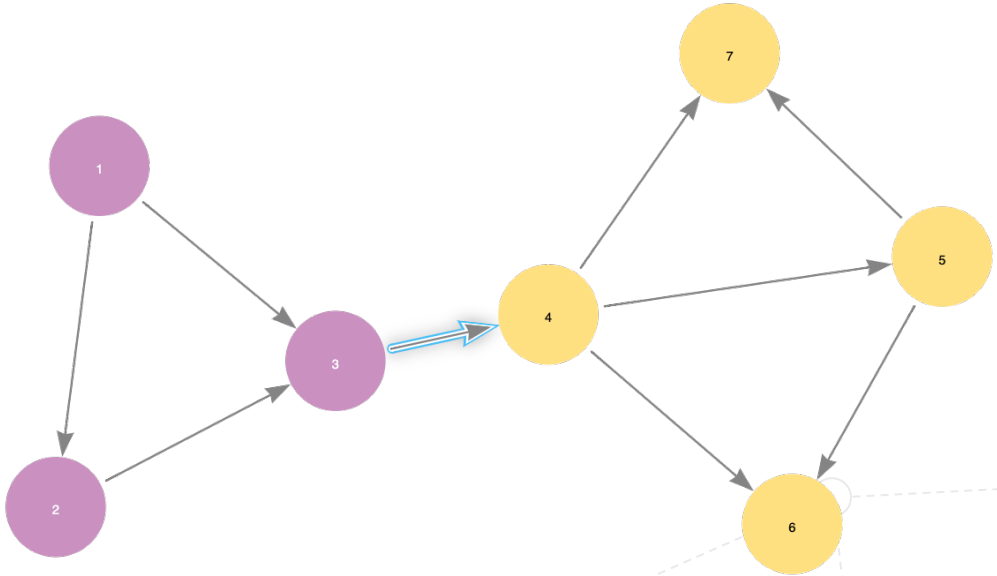
- Communities are distinct from each other
- Members of a community are closely related
- We can identify distinctive characteristics of different communities

What makes a partitioning useful?

- Communities are distinct from each other
  - **Conductance** and **Modularity**
- Members of a community are closely related
  - **Clustering coefficient**
- We can identify distinctive characteristics of different communities
  - **Centrality**

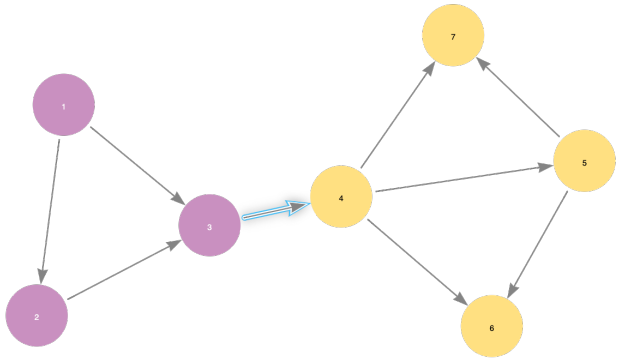
Conductance

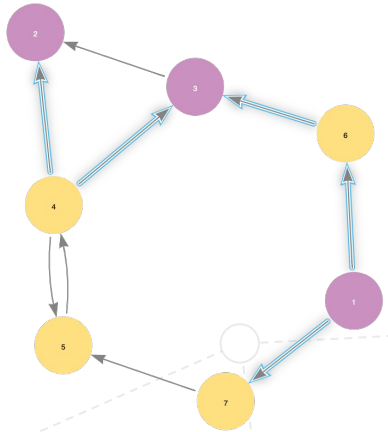
- What percentage of relationships that start in a community end in the same community?
- Lower scores mean more distinct communities.



Modularity

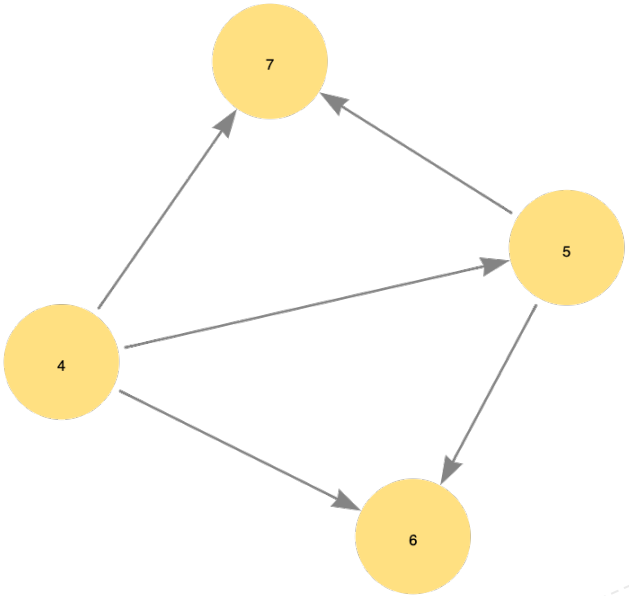
- What is the difference in the ratio of relationships with both endpoints in a community compared to what the ratio would be if the relationships were distributed randomly?
- Higher scores mean more distinct clusters



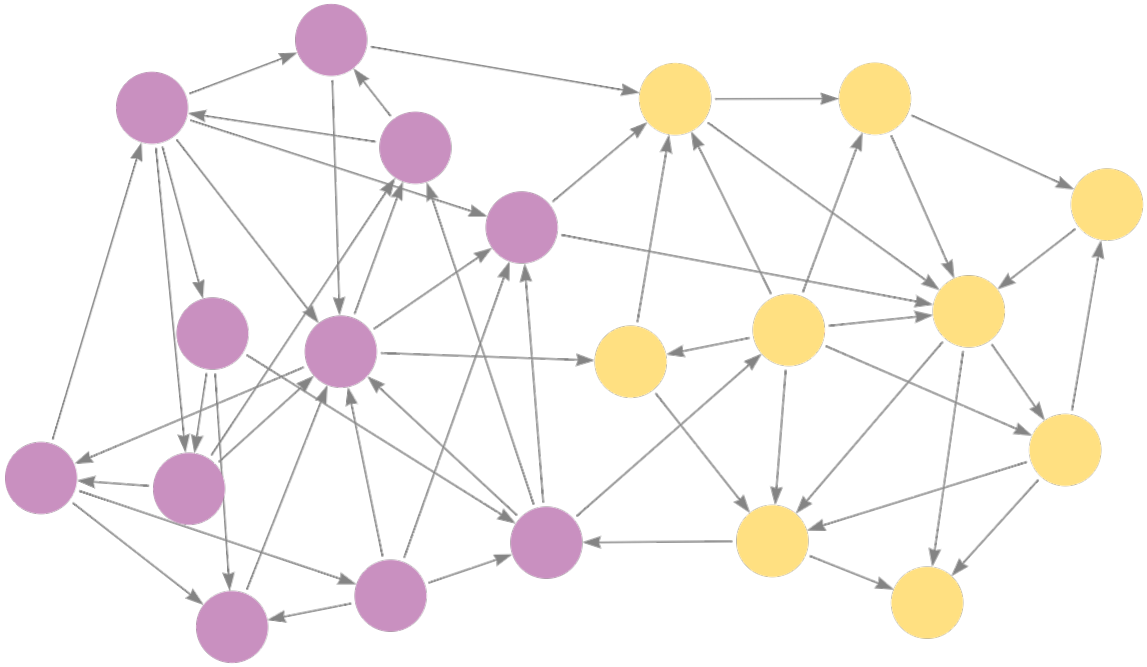


Clustering Coefficient

- What percentage of the neighbors of a node are related to each other?
- Higher scores mean more connected, cohesive clusters

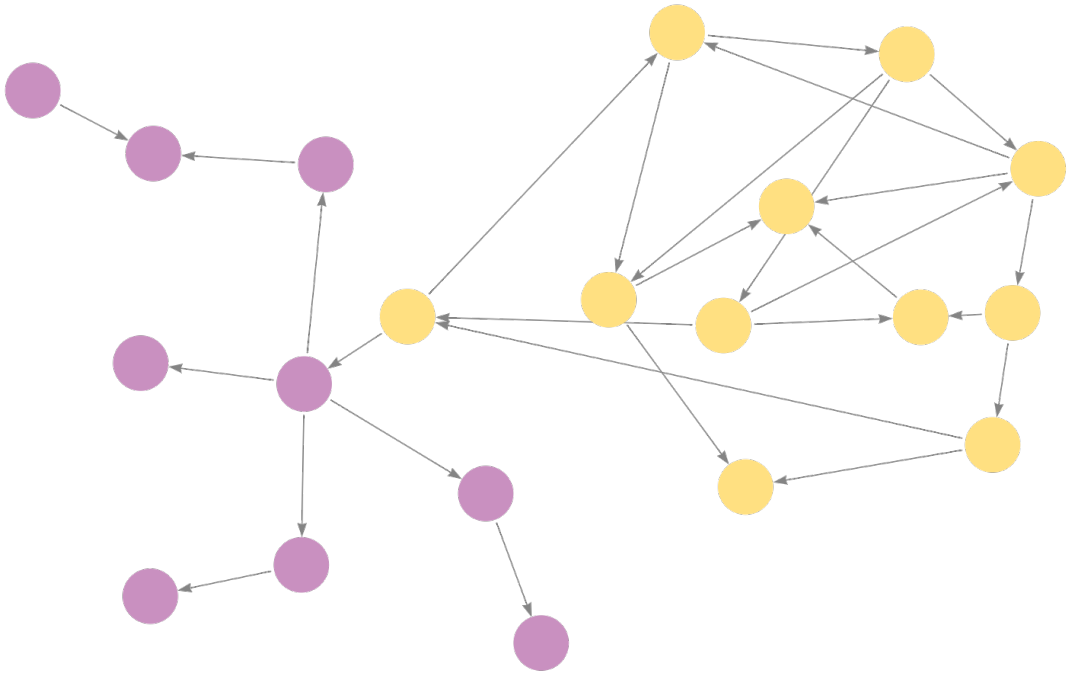


Example graph



- Conductance:
  - Purple: .10
  - Yellow: .13
- Modularity: 0.38
- Clustering Coefficient:
  - Purple: 0.54
  - Yellow: 0.65

Example graph



- Conductance:
  - Purple: .05
  - Yellow: .02
- Modularity: 0.38
- Clustering Coefficient:
  - Purple: 0.00
  - Yellow: 0.12

Demo

https://github.com/smithna/blogs/tree/main/cluster_quality

What to tell your boss

- Are the communities distinct and cohesive enough to be useful?
- What are the high centrality examples within each community?
- Are the community quality statistics changing over time?

Thank you!

Stay in touch at
nathan.smith@neo4j.com
@nsmith_piano on Twitter