Interpreting the results of community detection algorithms

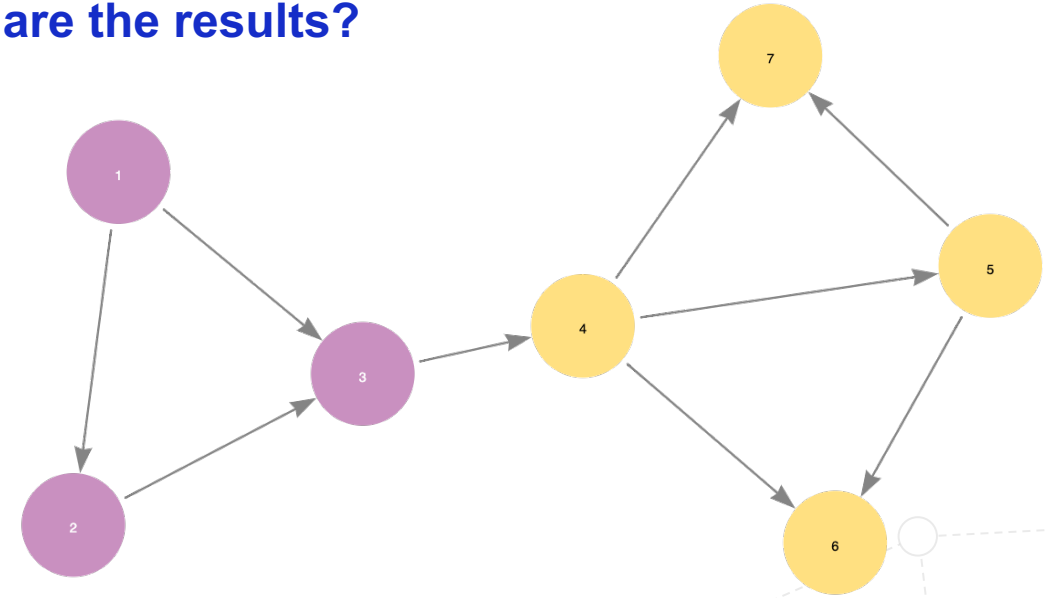
Nathan Smith,
Senior data scientist

Sample code repo

Congratulations!

You have successfully run a community detection algorithm!

But how useful are the results?



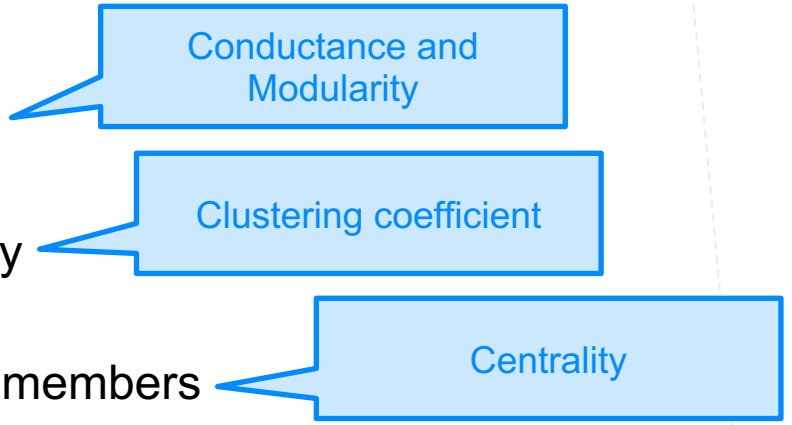
graphconnect

What makes a good community?

- Weak relationships to other communities
- Strong relationships within the community
- Identifiable characteristics of community members

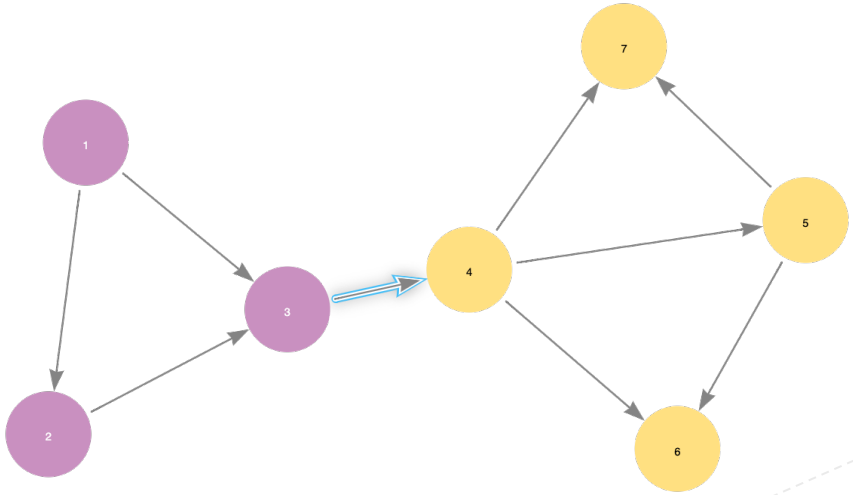
We can measure community quality!

- Weak relationships to other communities
- Strong relationships within the community
- Identifiable characteristics of community members



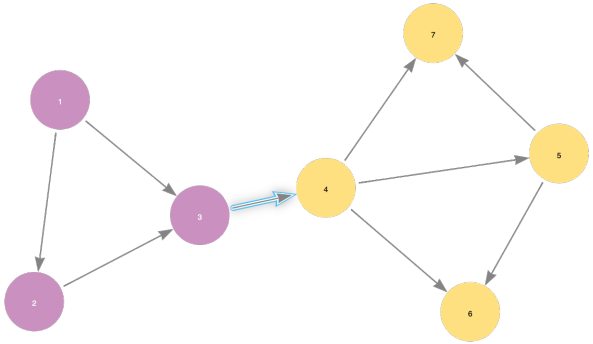
Conductance

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

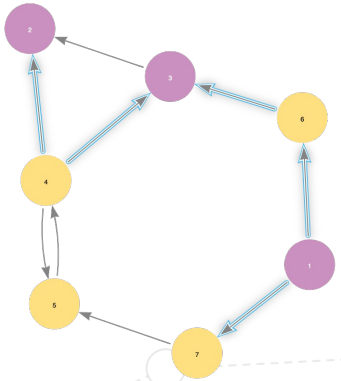


Modularity

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



Original relationships

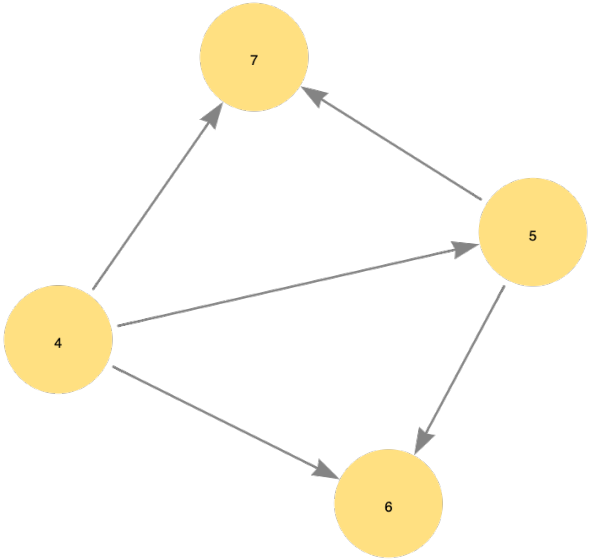


Randomly reassigned relationships

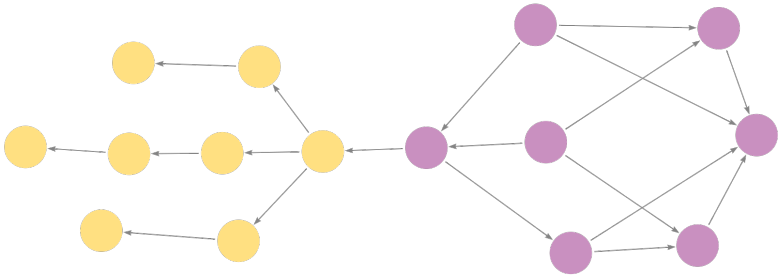
graphconnect

Clustering coefficient

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



Compare sample graph statistics



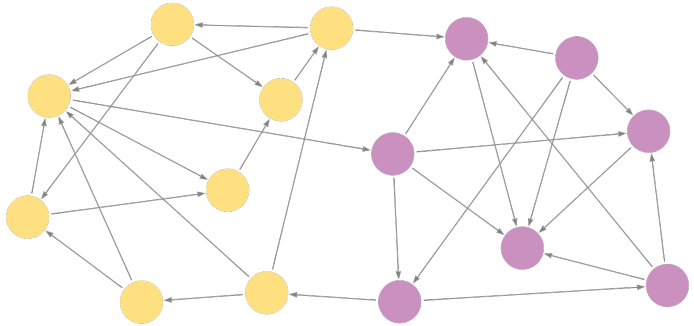
Conductance:

- Purple: 0.043
- Yellow: 0.067

Modularity: 0.425

Clustering Coef:

- Purple: 0.238
- Yellow: 0.0



Conductance:

- Purple: 0.097
- Yellow: 0.090

Modularity: 0.406

Clustering Coef:

- Purple: 0.371
- Yellow: 0.487

Demo

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

What to tell your boss

- Are the communities distinct and cohesive enough to be useful?
- Are the community quality statistics changing over time?
- What are high centrality examples within each community?
- Thank you for sending me to GraphConnect!

Thank you!

Contact us at
nathan.smith@neo4j.com
@nsmith_piano on Twitter