Community Detection

Topics

- Introduction
- Taxonomy of Community Criteria
  - Node-Centric
  - Group-Centric
  - Network-Centric
  - Hierarchical Clustering
- Overlapping Communities
- Community Evolution

Introduction

Communities

- Group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities. Two important areas where communities play a role: Social Network and Biological Network.

Social Network

- Explicit Groups: formed by user subscriptions.
- Implicit Groups: implicitly formed by social interactions.

Biological Network

- Proteins that are involved in the same disease tend to interact with each other.

Zachary's Karate Club

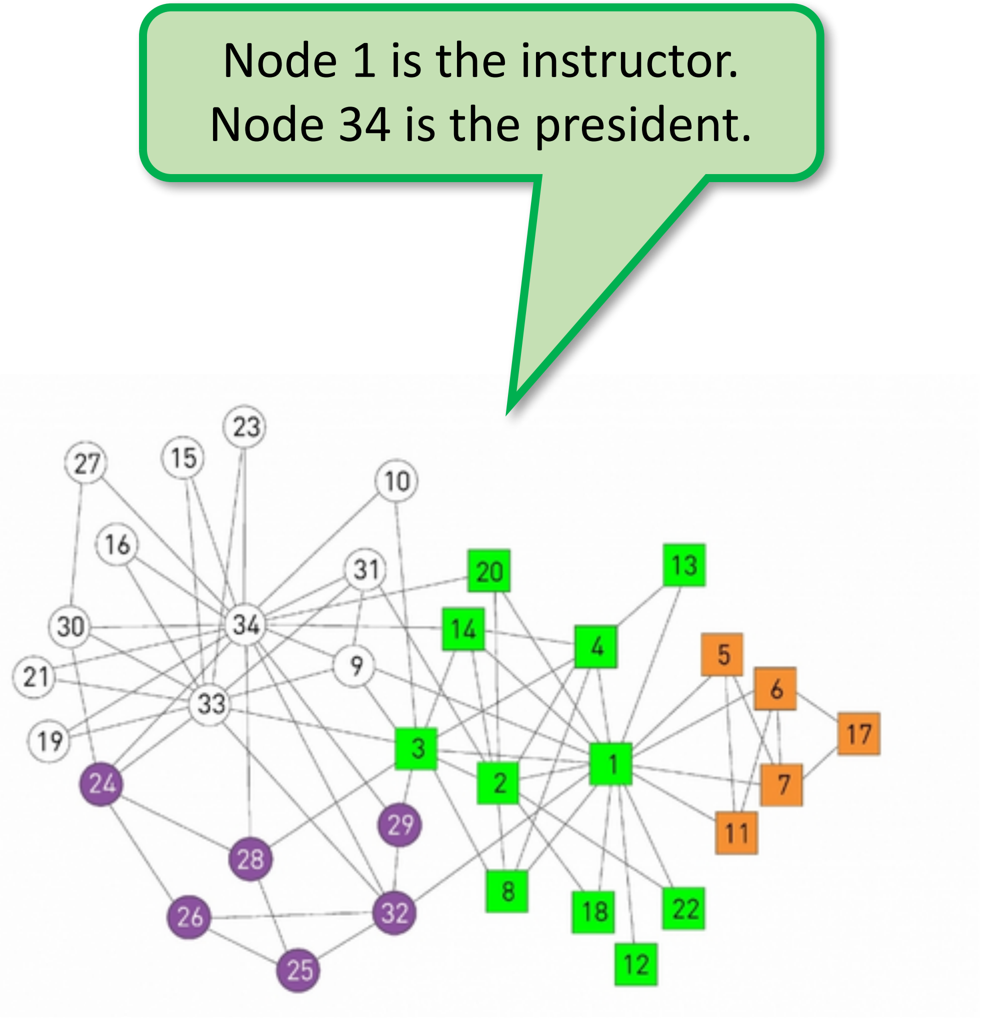
The network captures the links between 34 members of a karate club.

Sociologist Wayne Zachary documented 78 pairwise links between members who regularly interacted outside the club.

The interest in the dataset is driven by a singular event: a conflict between the club's president and the instructor split the club into two.

About half of the members followed the instructor and the other half the president - a breakup that unveiled the underlying community structure.

Today community finding algorithms are often tested based on their ability to infer these two communities from the structure of the network before the split.



Community Hypothesis

Connectedness Hypothesis

- *A community is a locally dense connected subgraph in a network.*
- Each community corresponds to a connected subgraph.
- A community cannot consist of two subgraphs that do not have a link to each other.

Density Hypothesis

- Nodes in a community are more likely to connect to other members of the same community than to nodes in other communities.

Strong and Weak Communities

Strong Community

A connected subgraph where **each** of it's nodes have more links to other nodes in the same community that to nodes that belong to other communities (fig. b).

Weak Community

A subgraph whose nodes' **total** internal degree exceeds their **total** external degree (fig. c).

Figure b

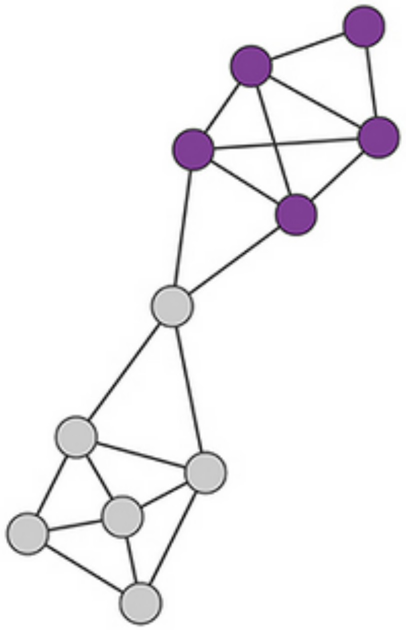
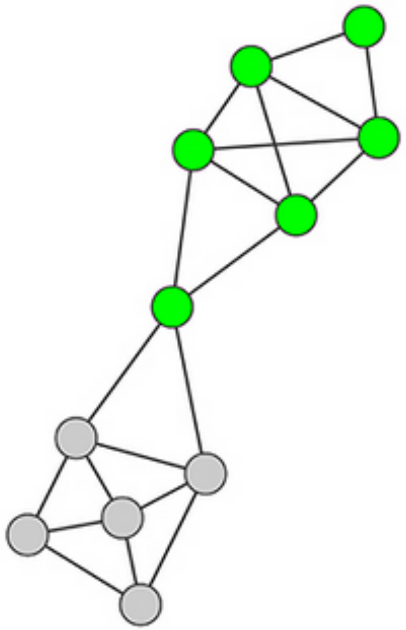


Figure c



Taxonomy of Community Criteria

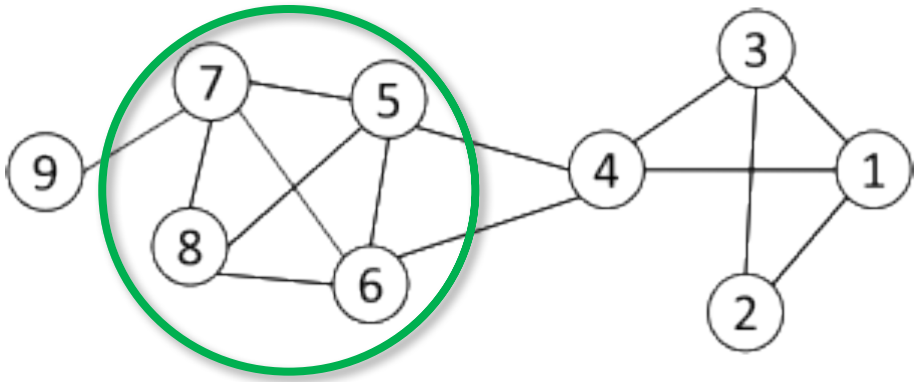
Roughly, community detection methods can be divided into 4 categories (not exclusive):

1. **Node**-Centric Community
   - **Each node** in a group satisfies certain properties.
   - Example: Cliques.
2. **Group**-Centric Community
   - Consider the connections **within a group** as a whole. The group has to satisfy certain properties without zooming into node-level.
3. **Network**-Centric Community
   - Partition **the whole network** into several disjoint sets.
   - Examples: Vertex Similarity, Cut, Modularity.
4. **Hierarchy**-Centric Community
   - Construct a **hierarchical structure** of communities
   - Examples: Divisive (top-down); Agglomerative (bottom-up)

Source: Chapter 3, *Community Detection and Mining in Social Media*. Lei Tang and Huan Liu, Morgan & Claypool, September, 2010.
http://www.cc.gatech.edu/classes/AY2011/cs7270_spring/7270-community-detection.pptx

Node Centric Community: Cliques

Clique: a **maximum complete** subgraph in which all nodes are adjacent to each other.



Nodes 5, 6, 7 and 8 form a clique

Straightforward implementation to find cliques is very expensive in time complexity.

Group-Centric Community Detection: Density-Based Groups

The group-centric criterion requires the whole group to satisfy a certain condition.

- In other words, the group density >= a given threshold.

Network-Centric Community Detection

Network-centric criterion needs to consider the connections within a network **globally.**

Goal: partition nodes of a network into **disjoint** sets.

Approaches:

- Clustering based on vertex similarity
- Cut
- Modularity maximization

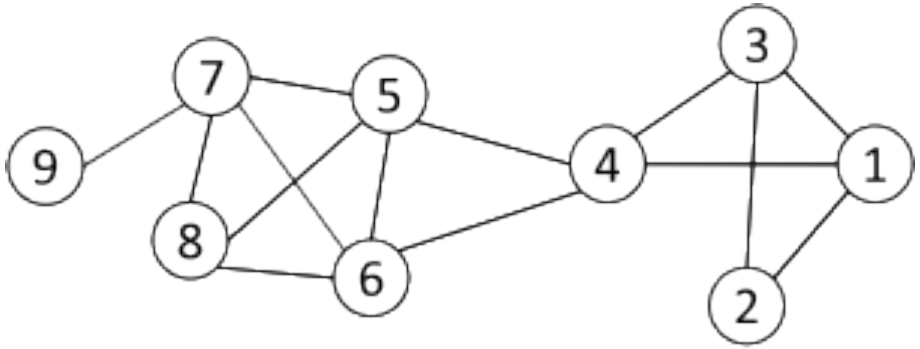
Clustering Based on Vertex Similarity

Apply k-means or similarity-based clustering to nodes.

Vertex similarity is defined in terms of **the similarity of their neighborhood.**

Structural equivalence: two nodes are structurally equivalent if they connect to the same set of actors.

Nodes 1 and 3 are structurally equivalent; So are nodes 5 and 6.



Structural equivalence is too restrictive for practical use.

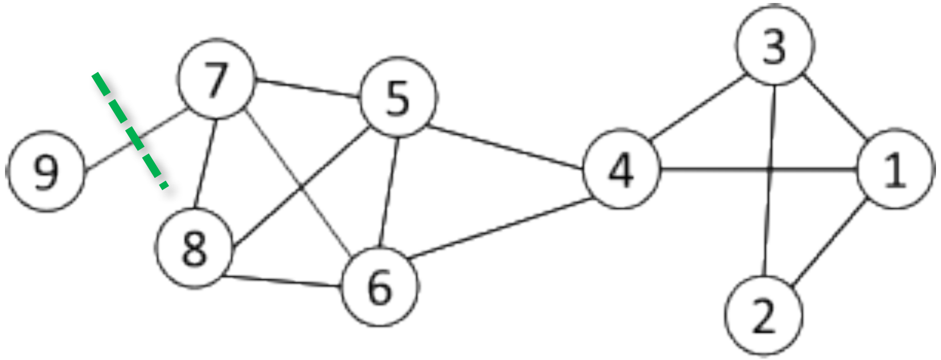
Cut

Most interactions are within group whereas interactions between groups are few.

community detection → **minimum cut problem**

Cut: A partition of vertices of a graph into two disjoint sets

Minimum cut problem: find a graph partition such that the number of edges between the two sets is minimized



Modularity

Community exists if the number of links present in a network is equal to or more than the expected number of links in a random network of similar size.

Represents actual number of links in the network

$$M_c = \frac{1}{2L} \sum_{(i,j) \in C_c} \left(A_{ij} - p_{ij}\right)$$

Represents expected number of links in a random network

If M_c is positive, then the subgraph C_c has more links than expected by chance, hence it represents a potential community. If M_c is zero then the connectivity between the N_c nodes is random, fully explained by the degree distribution. Finally, if M_c is negative, then the nodes of C_c do not form a community.

Modularity: Greedy Algorithm (Newman)

Maximal Modularity Hypothesis

The partition with the maximum modularity offers the best community structure, where modularity is given by

$$M = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

Iteratively joins pairs of communities if the move increases the partition's modularity. The algorithm follows these steps:

1. Assign each node to a community of its own, starting with *N* communities of single nodes.
2. Inspect each community pair connected by at least one link and compute the modularity difference *ΔM* obtained if we merge them. Identify the community pair for which *ΔM* is the largest and merge them. Note that modularity is always calculated for the full network.
3. Repeat Step 2 until all nodes merge into a single community, recording *M* for each step.
4. Select the partition for which *M* is maximal.

Hierarchy-Centric Community Detection

Goal: build a hierarchical structure of communities based on network topology.

Allow the analysis of a network at different resolutions.

Representative approaches:

- Divisive Hierarchical Clustering (top-down)
- Agglomerative Hierarchical clustering (bottom-up)

Divisive Hierarchical Clustering (top down)

Divisive clustering

- Partition nodes into several sets.
- Each set is further divided into smaller ones.
- Network-centric partition can be applied for the partition.

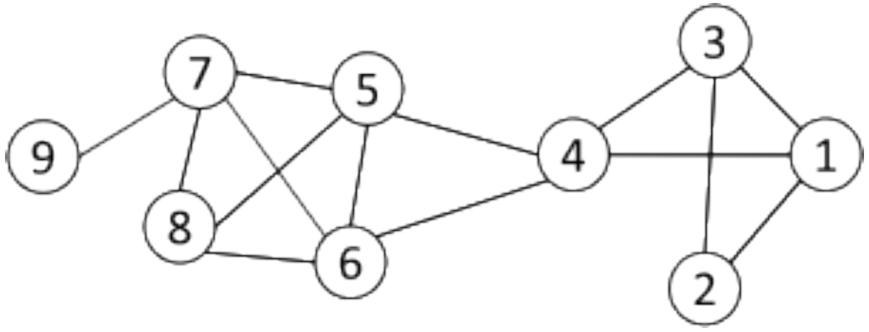
One particular example: **recursively remove the "weakest" tie**

1. Find the edge with the least strength.
2. Remove the edge and update the corresponding strength of each edge.

Recursively apply the two steps above until a network is decomposed into desired number of connected components.

Each component forms a community.

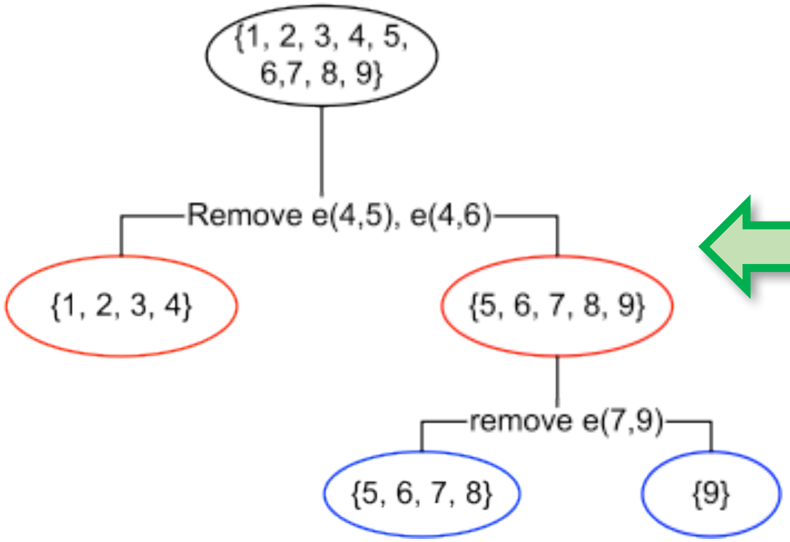
Divisive clustering based on edge betweenness



Initial betweenness value

Table 3.3: Edge Betweenness

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0



After remove e(4,5), the betweenness of **e(4, 6) becomes 20**, which is the highest;

After remove e(4,6), the edge **e(7,9)** has the highest betweenness value **4**, and should be removed.

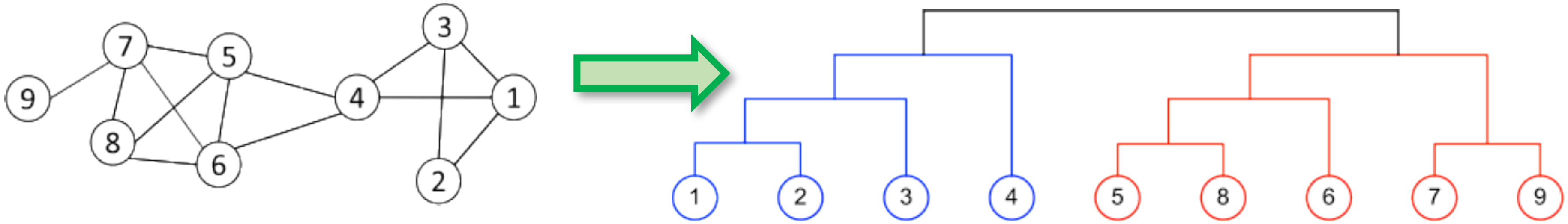
Idea: progressively removing edges with the highest betweenness.

Agglomerative Hierarchical Clustering (bottom up)

Initialize each node as a community.

Merge communities successively into larger communities following a certain criterion.

- For example, based on modularity increase.



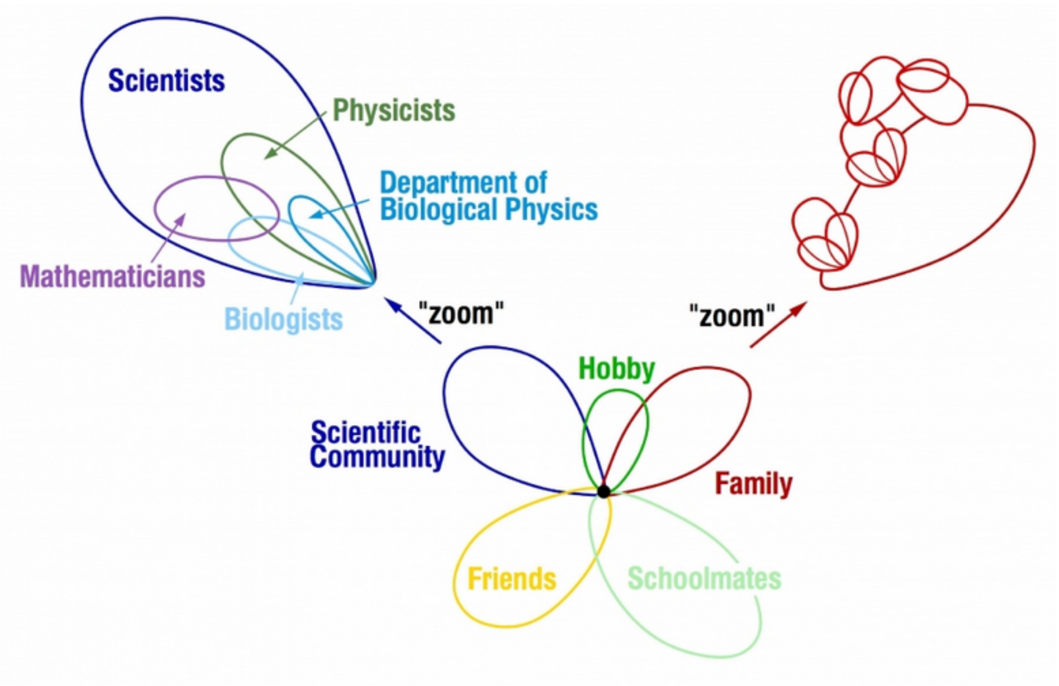
Dendrogram according to Agglomerative Clustering based on Modularity

Overlapping Communities

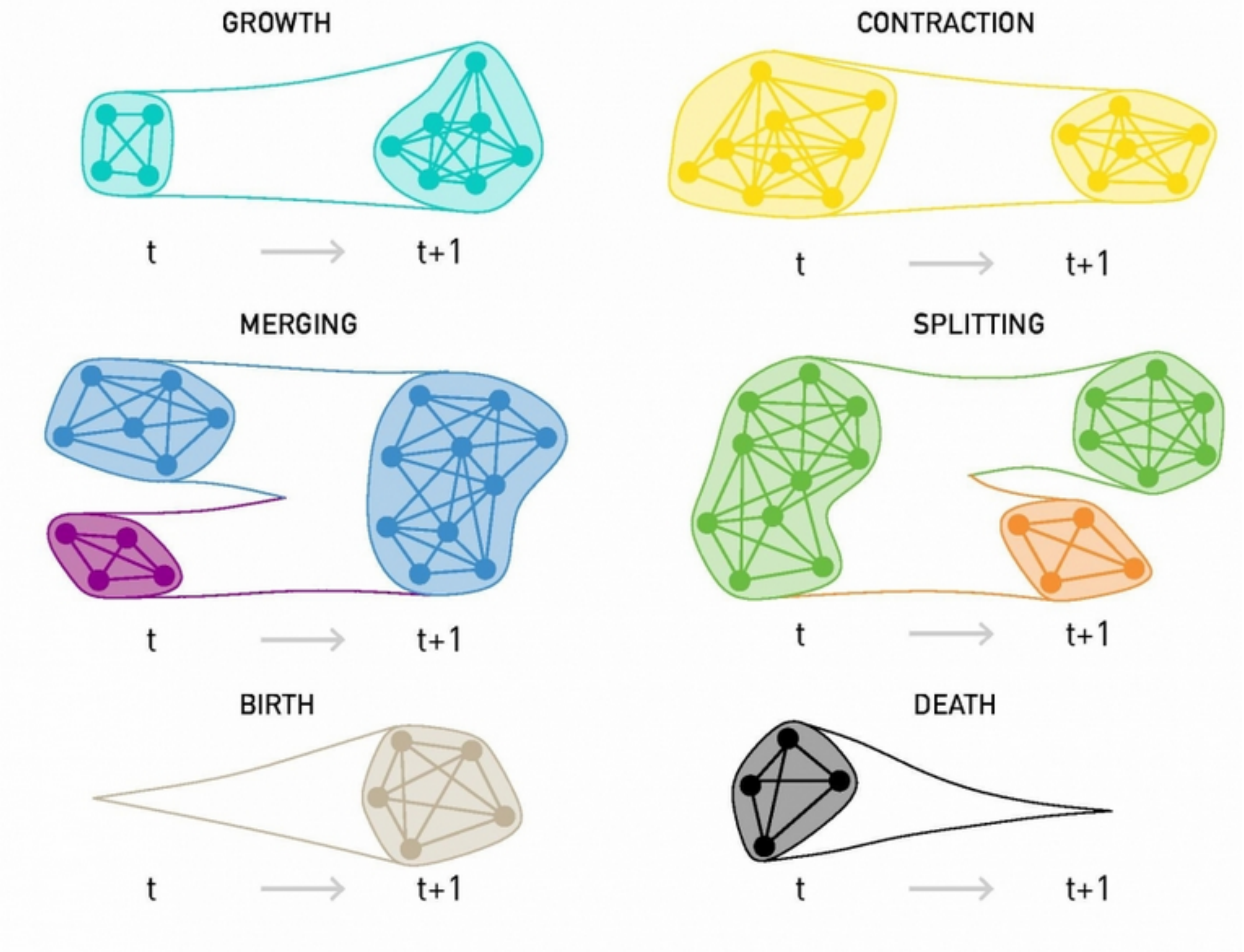
A node is rarely confined to a single community.

A scientist may belong to multiple communities depending upon his professional interests. Yet, she also belongs to a community consisting of family members and relatives and perhaps another community of individuals sharing his hobby.

Overlapping communities are not limited to social systems: The same genes are often implicated in multiple diseases, an indication that disease modules of different disorders overlap.



Community Evolution



Questions? Thoughts? Visit the course's online discussion forum.