14: Clique Finding

Machine Learning and Real-world Data (MLRD)

Paula Buttery (based on slides by Simone Teufel)

Ticks 11 and 12: focus on betweenness centrality

- Tick 11: implementation of **betweenness centrality**.
- This let you find “gatekeeper” nodes in the Facebook network.
- Tick 12: uses betweenness to find clusters in networks.

Quick run through of Task 12

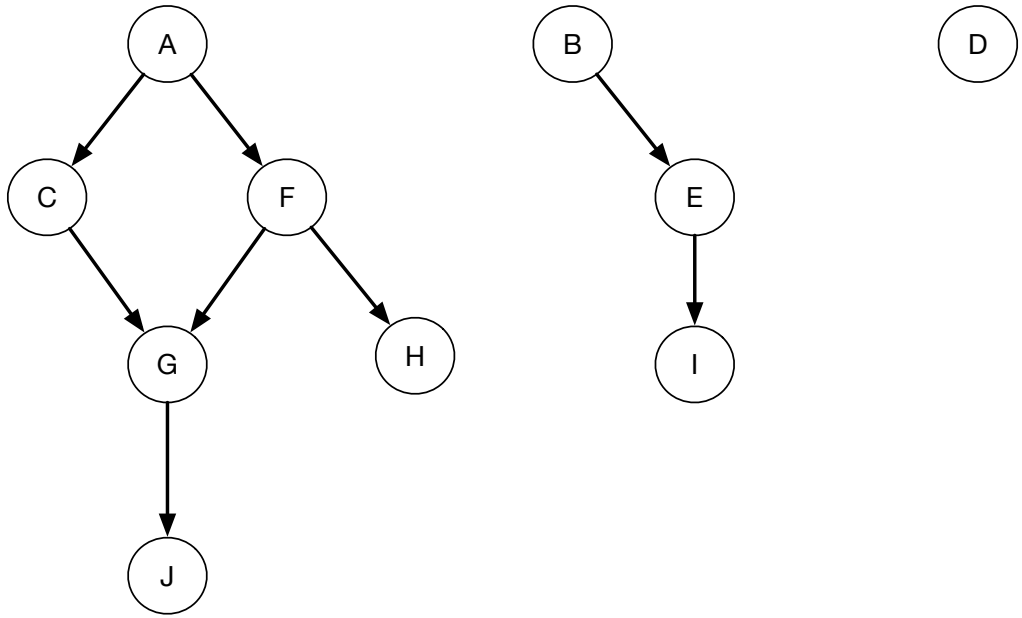
Three main tasks:

1. Determine connected components in the graph.
2. Change the Brandes code for betweenness centrality (from nodes to edges).
3. Implement the Newman-Girvan to discover clusters in the network provided.

1. Determining connected components

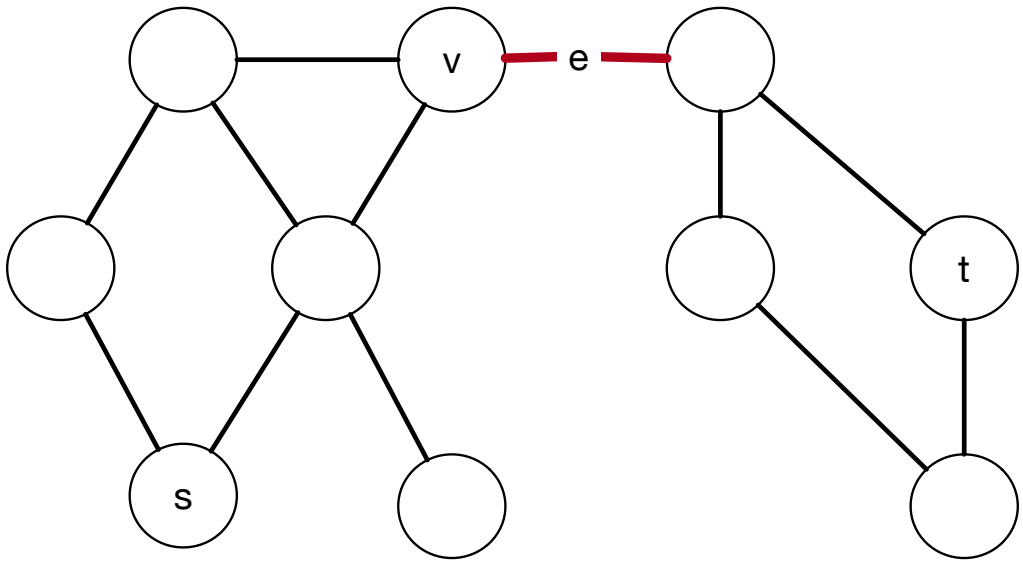
The task's graph is disconnected: there are five connected components. To find connected components:

- Depth-first search, start at an arbitrary node and mark the other nodes you reach.
- Repeat with unvisited nodes, until all are visited.



2. Change Brandes code for **edge** betweenness

- Previously: $\sigma(s, t|v)$ — the number of shortest paths between s and t going through node v.
- Now: $\sigma(s, t|e)$ — the number of shortest paths between s and t going through edge e.



2. Change Brandes code for **edge** betweenness

Add edge betweenness $c_B[(v,w)]$ in the bottom-up phase:

▼ **accumulation** // — back-propagation of dependencies
  for $v \in V$ **do** $\delta[v] \leftarrow 0$
  while S *not empty* **do**
    pop $w \leftarrow S$
    for $v \in Pred[w]$ **do** $\delta[v] \leftarrow \delta[v] + \frac{\sigma[v]}{\sigma[w]} \cdot (1 + \delta[w])$
    if $w \neq s$ **then** $c_B[w] \leftarrow c_B[w] + \delta[w]$

Edge betweenness

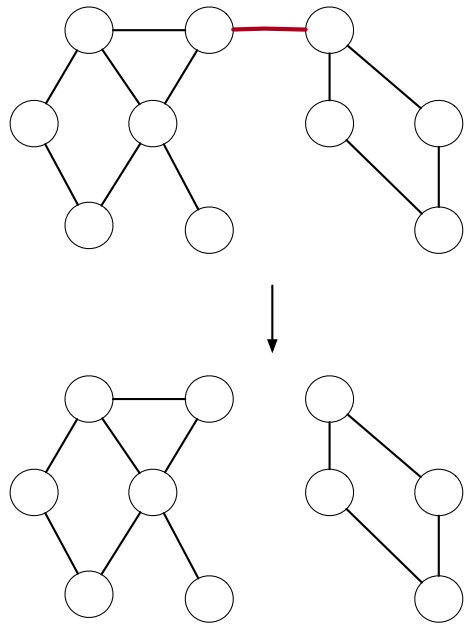
output: betweenness $c_B[q]$ for $q \in V \cup E$ (initialized to 0)

▼ **accumulation**
  for $v \in V$ **do** $\delta[v] \leftarrow 0$
  while S *not empty* **do**
    pop $w \leftarrow S$
    for $v \in Pred[w]$ **do**
      $c \leftarrow \frac{\sigma[v]}{\sigma[w]} \cdot (1 + \delta[w])$
      $c_B[(v,w)] \leftarrow c_B[(v,w)] + c$
      $\delta[v] \leftarrow \delta[v] + c$
    if $w \neq s$ **then** $c_B[w] \leftarrow c_B[w] + \delta[w]$

3. Implement Newman-Girvan to form clusters

While number of connected subgraphs $<$ specified number of clusters (and there are still edges):

1. calculate edge betweenness for every edge in the graph
2. remove edge(s) with highest betweenness
3. recalculate number of connected components



Note:

- Treatment of tied edges: either remove all (do this for task 12) or choose one randomly.

Thinking more about clustering...

Clustering vs. classification:

- Clustering: automatically grouping data according to some notion of closeness or similarity.
- Classification (e.g., sentiment classification): assigning data items to predefined classes.
- Clustering: groupings can emerge from data, unsupervised.
- Can cluster anything as long as there's a notion of similarity between items.

There are many ways to cluster...

Hard vs. soft:

- Hard clustering: each data point either belongs to a cluster completely or it doesn't.
- Soft clustering: data points are scored for likelihood of being in a cluster.
- Most famous technique for hard clustering is k-means: it's a general technique with a variant for graphs (k is number of clusters).

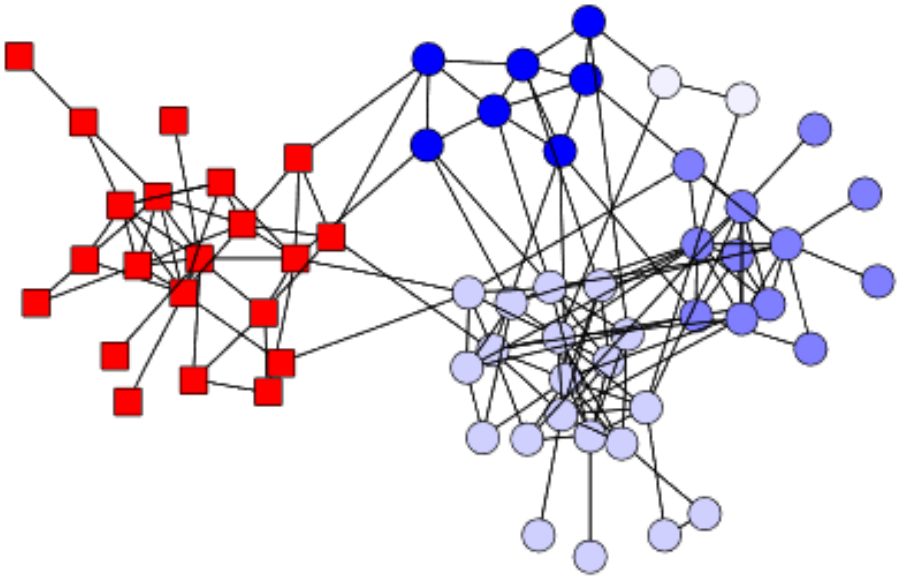
Top-down vs. bottom up:

- Agglomerative clustering joins nodes together.
- Divisive clustering splits nodes apart.
  - Newman-Girvan method — divisive clustering where criterion for breaking links is edge betweenness centrality.

Real world data: Newman-Girvan on Dolphin data

Community structure of bottlenose dolphins at Doubtful Sound:

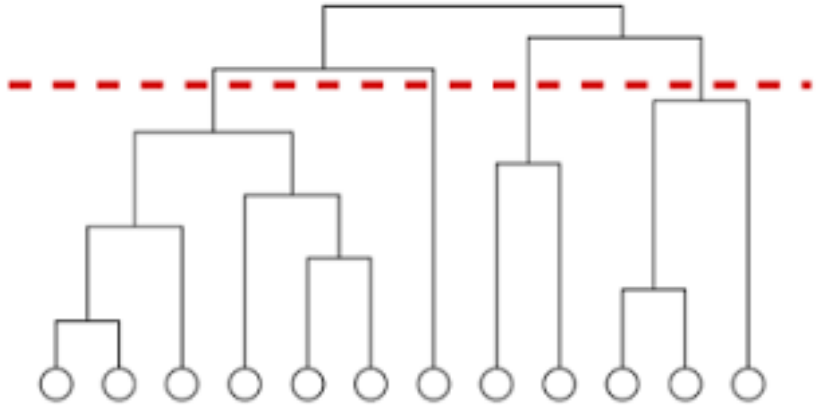
- squares vs circles: first split.
- shades of blue: 4 further splits.



Links between dolphin pairs established by observation of statistically significant frequent association.

Note: longer edges between vertices in different communities only to make the community groupings clearer. Newman and Girvan (2004)

How many clusters?



A cross-section of a dendrogram tree gives the clusters at a given number of splits.

- Newman-Girvan define modularity as the quality of a particular division of a network
- Modularity is 0 if the number of within-community edges is no better than random (the maximum is 1 indicating strong community structure)
- Can use peak modularity to choose number of clusters.

The modularity for the bottlenose dolphin split is Q = 0.52

Newman and Girvan (2004)

Real world data: evaluating dolphin clusters

The split into two groups appears to correspond to a known division of the dolphin community [38]. Lusseau reports that for a period of about two years during observation of the dolphins they separated into two groups along the lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. When some of these individuals later reappeared, the two halves of the network joined together once more. As Lusseau points out, developments of this kind illustrate that the dolphin network is not merely a scientific curiosity but, like human social networks, is closely tied to the evolution of the community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals: the largest subgroup consists almost of entirely of females and the others almost entirely of males, and it is conjectured that the split between the male groups is governed by matrilineage (D. Lusseau, personal communication) Newman and Girvan (2004)

Real world data: evaluating dolphin clusters

The split into two groups appears to correspond to a known division of the dolphin community [38]. Lusseau reports that for a period of about two years during observation of the dolphins they separated into two groups along the lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. When some of these individuals later reappeared, the two halves of the network joined together once more. As Lusseau points out, developments of this kind illustrate that the dolphin network is not merely a scientific curiosity but, like human social networks, is closely tied to the evolution of the community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals: the largest subgroup consists almost of entirely of females and the others almost entirely of males, and it is conjectured that the split between the male groups is governed by matrilineage (D. Lusseau, personal communication) Newman and Girvan (2004)

Real world data: evaluating dolphin clusters

The split into two groups appears to correspond to a known division of the dolphin community [38]. Lusseau reports that for a period of about two years during observation of the dolphins they separated into two groups along the lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. When some of these individuals later reappeared, the two halves of the network joined together once more. As Lusseau points out, developments of this kind illustrate that the dolphin network is not merely a scientific curiosity but, like human social networks, is closely tied to the evolution of the community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals: the largest subgroup consists almost of entirely of females and the others almost entirely of males, and it is conjectured that the split between the male groups is governed by matrilineage (D. Lusseau, personal communication) Newman and Girvan (2004)

Real world data: evaluating dolphin clusters

The split into two groups appears to correspond to a known division of the dolphin community [38]. Lusseau reports that for a period of about two years during observation of the dolphins they separated into two groups along the lines found by our analysis, apparently because of the disappearance of individuals on the boundary between the groups. When some of these individuals later reappeared, the two halves of the network joined together once more. As Lusseau points out, developments of this kind illustrate that the dolphin network is not merely a scientific curiosity but, like human social networks, is closely tied to the evolution of the community. The subgroupings within the larger half of the network also seem to correspond to real divisions among the animals: the largest subgroup consists almost of entirely of females and the others almost entirely of males, and it is conjectured that the split between the male groups is governed by matrilineage (D. Lusseau, personal communication) Newman and Girvan (2004)

How to evaluate clusters?

Intrinsic evaluation—evaluate the clusters directly.

- Evaluate against reference clusters (there doesn't need to be a 1-to-1 mapping).
- Compare to small set of reference labels.
- Sample random pairs from the data set and human annotate whether they should be in the same cluster.

Extrinsic evaluation—evaluate the clusters through task performance.

- Practical evaluation: use the system to do a task and evaluate that task.

Don’t forget to pick up your pen!