# Community Detection

Jake Carlson

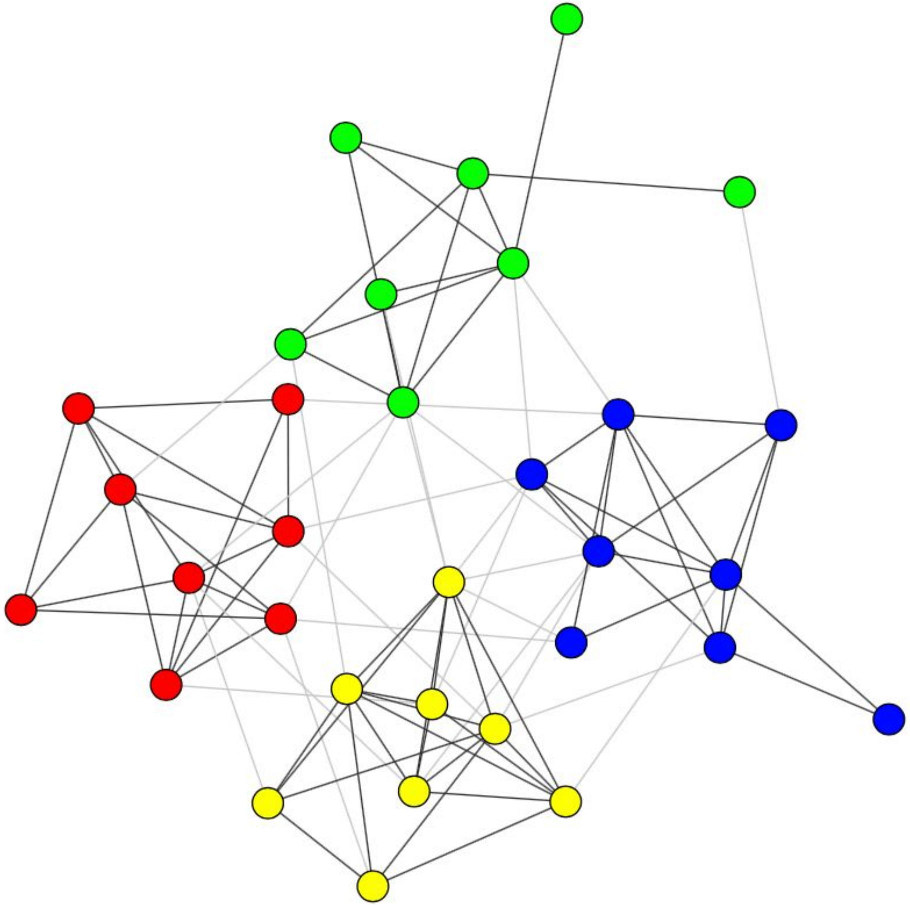
# Background

# Related Problems

- Graph partitioning
  - Divide (cut) a graph into two subgraphs
  - Repeatedly find the best cut to make
- Graph clustering
  - Group nodes based on distance measure

# Goal

- Find community structure in graphs such that there are:
  - Many connections between nodes in a community
  - Few connections between different communities



# Assessing Quality

- Quality
- Multi-criterion scores
- Single-criterion scores

# Multi-criterion Scores

- Conductance - the ratio of edge volume within the community to the edge volume outside of the community

- Expansion - the ratio between the number of edges on the boundary of a community to the number of nodes in the community

$$\phi(S) = \frac{c_S}{min(Vol(S), Vol(V \setminus S))}$$

$$c_S = |(u, v) : u \in S, v \notin S|$$

$$Vol(S) = \sum_{u \in S} d(u)$$

$$f(S) = \frac{c_S}{n_S}$$

# Single-criterion Score

- Modularity - difference between the number of edges in the community to the expected number of edges in a random graph where nodes have the same degree

$$f(S) = \frac{1}{4m}(m_S - E(m_S))$$

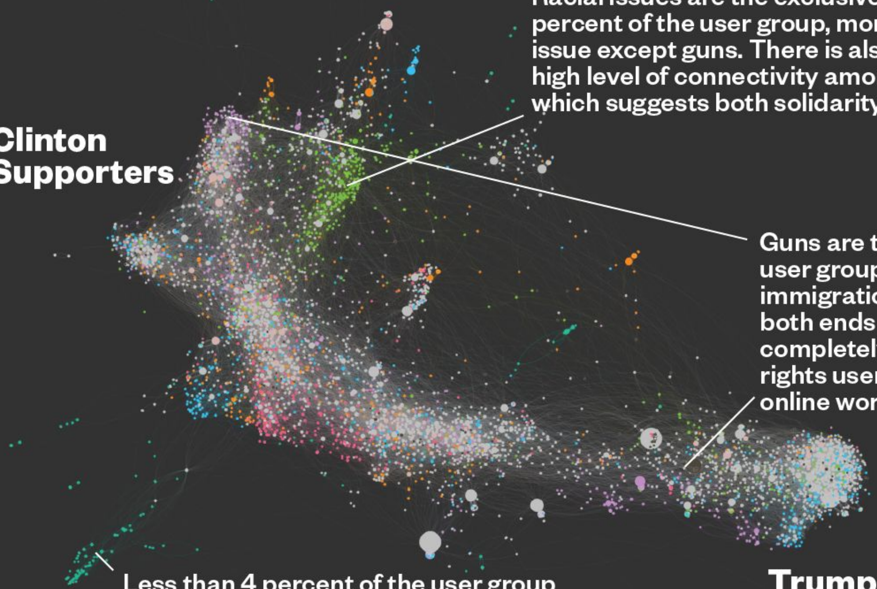
# Motivation

# Motivation

- Widely available network data
- Business motivation to find communities:
  - Better target ads
  - Recommend friends
  - Recommend new communities
- Understand phenomena in the world

# Which issues are talked about the most on Twitter

| | | |
|---|---|---|
| 🟪 | Guns | 10.05% |
| 🟩 | Racial Issues | 8.65% |
| 🟦 | Immigration | 8.65% |
| ⬛ | Terrorism | 8.3% |
| 🟧 | Jobs | 6.84% |
| 🟥 | Economy | 5.44% |
| 🟩 | Education | 3.23% |
| ⬜ | Combination of issues | |



Racial issues are the exclusive focus of 8 percent of the user group, more than any other issue except guns. There is also an extremely high level of connectivity among these users which suggests both solidarity and insularity.

Guns are the sole focus of over 10 percent of the user group, the most of any issue. Similar to immigration, there are clusters of "guns" users on both ends of the spectrum and they are almost completely disconnected from each other. Gun rights users and gun control users live in separate online worlds.

Less than 4 percent of the user group talk solely about education and those who do are very disconnected from most of the political conversation.

Source: The Electome | The Laboratory for Social Machines at the MIT Media Lab

# Methods

# Methods

- Spectral Partitioning
- Edge Betweenness
- Walktrap

# Method 1 - Spectral Partitioning

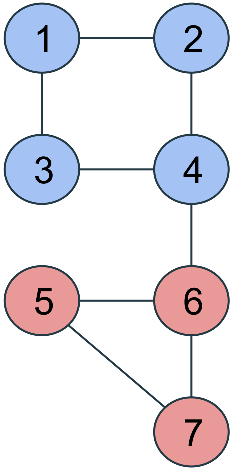
- Miroslav Fiedler’s theory of spectral graph partitioning
  - Build Laplacian matrix for the graph: $L = D - A$
  - Calculate eigenvectors and eigenvalues
  - Partition using the eigenvector with the second smallest eigenvalue
- Not intuitive why this works…

D

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 2 |   |   |   |   |   |   |
| 2 |   | 2 |   |   |   |   |   |
| 3 |   |   | 2 |   |   |   |   |
| 4 |   |   |   | 3 |   |   |   |
| 5 |   |   |   |   | 2 |   |   |
| 6 |   |   |   |   |   | 3 |   |
| 7 |   |   |   |   |   |   | 2 |

A

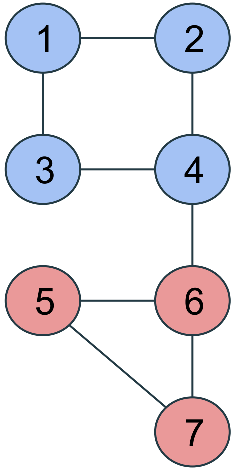
|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 |   | 1 | 1 |   |   |   |   |
| 2 | 1 |   |   | 1 |   |   |   |
| 3 | 1 |   |   | 1 |   |   |   |
| 4 |   | 1 | 1 |   |   | 1 |   |
| 5 |   |   |   |   |   | 1 | 1 |
| 6 |   |   |   | 1 | 1 |   | 1 |
| 7 |   |   |   |   | 1 | 1 |   |



# Building the Laplacian

L

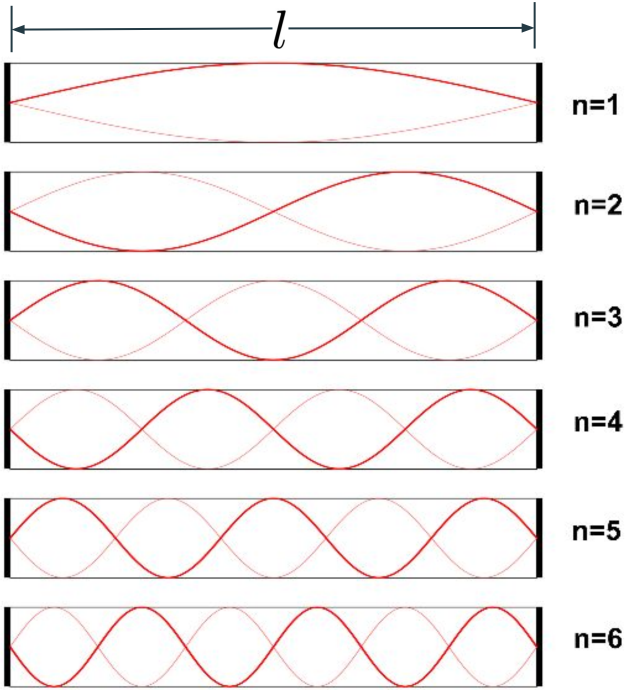
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 2 | -1 | -1 | | | | |
| 2 | -1 | 2 | | -1 | | | |
| 3 | -1 | | 2 | -1 | | | |
| 4 | | -1 | -1 | 3 | | -1 | |
| 5 | | | | | 2 | -1 | -1 |
| 6 | | | | -1 | -1 | 3 | -1 |
| 7 | | | | | -1 | -1 | 2 |



# Building the Laplacian

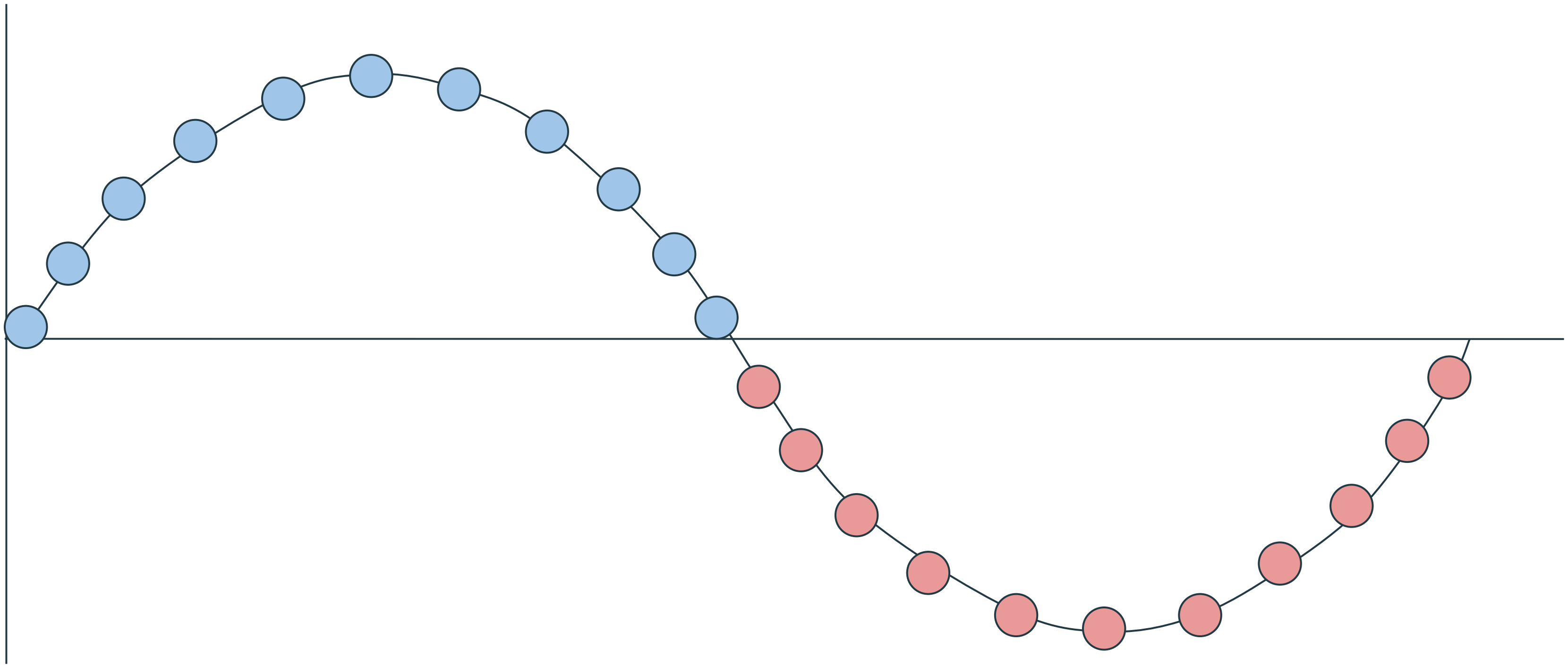
- If we say the points on the string are nodes in a graph, we have a chain graph
- The Laplacian describes the motion of the individual points
- The eigenvectors and eigenvalues of the Laplacian are tied to the frequency of these waves

$$\lambda = \frac{2l}{n}$$



https://www.researchgate.net/figure/Standing-waves-in-a-box_fig66_324820292

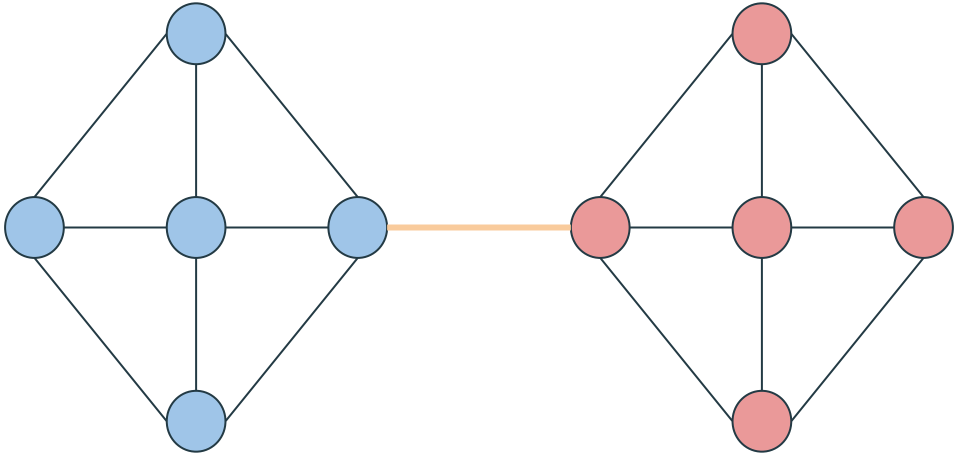
# A Physical Example - Standing Waves



# A Physical Example - Standing Waves

# Method 2 - Edge Betweenness

- Find edges that are the most ‘between’ communities
    - Extended Freeman’s betweenness centrality from nodes to edges
- Repeatedly remove the most ‘between’ edge, save the order
    - In reverse order, these are the edges that are most central to communities
- Build dendrogram by replaying merges and cut where modularity is optimal



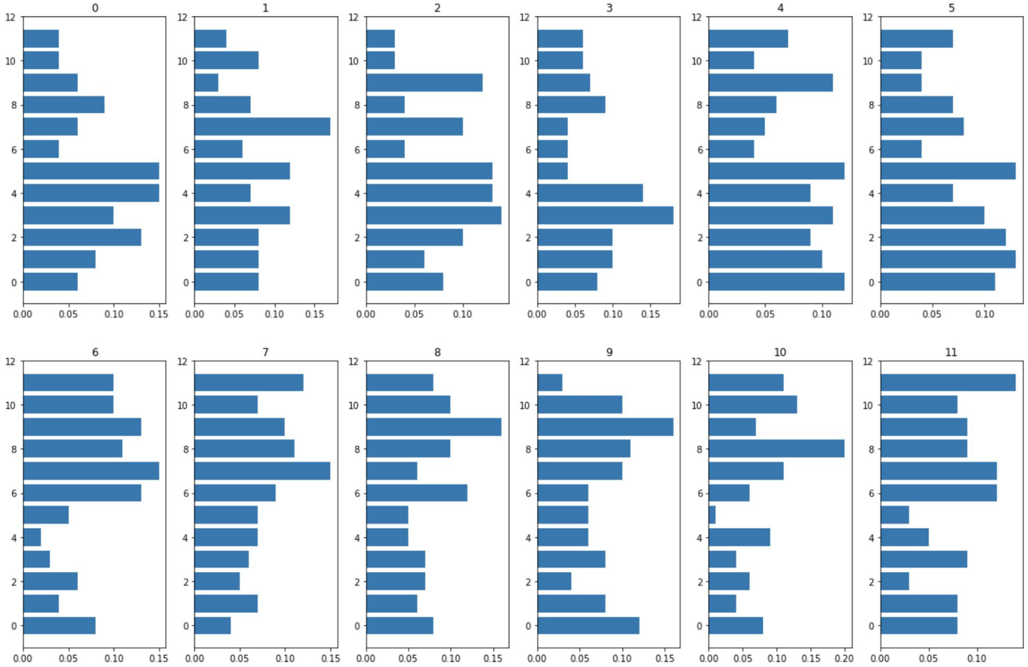
Edge Betweenness

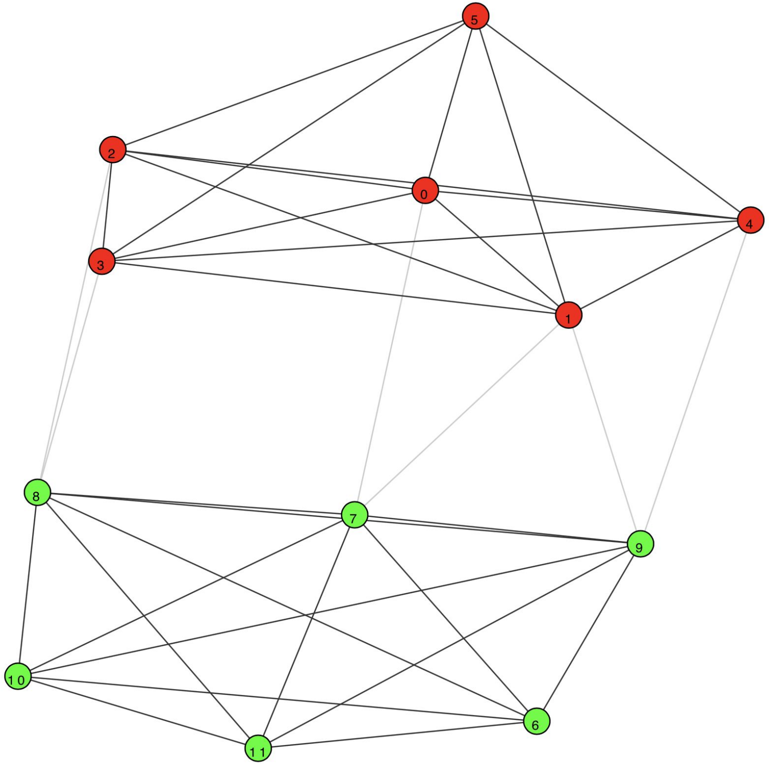
# Method 3 - Walktrap

- Random walks through graphs tend to stay in highly connected areas
  - Higher probability of choosing an edge that leads us to another community node: $P = D^{-1} A$
- Distance between vertices *i* and *j*:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^{t} - P_{jk}^{t})^2}{d(k)}} = ||D^{-\frac{1}{2}} P_{i\bullet}^{t} - D^{-\frac{1}{2}} P_{j\bullet}^{t}||$$

Pascal Pons and Matthieu Latapy, Computing communities in large networks using random walks


0
1
2
3
4
5
6
7
8
9
10
11
0
2
4
6
8
10
12
0.00
0.05
0.10
0.15
0.20


5
2
0
4
3
1
8
7
9
10
6
11

# Walktrap

# Questions