Network Theory and Dynamic Systems
Clustering and Context

Edited from lecture slides by Jerome Kunegis
who based on lecture slides by Steffen Staab
Clustering
Graph Drawing

Draw nodes in a plane so that it is meaningful.
Graph Clustering
Graph Drawing: Mean of Neighbors
Graph Drawing: Mean of Neighbors
Graph Clustering: Zachary Karate Club

http://konect.uni-koblenz.de/networks/ucidata-zachary
Karate Club Study by Wayne Zachary

Dispute leads to split into two groups

How to predict the boundary?

http://konect.uni-koblenz.de/networks/ucidata-zachary
Minimum cut problem: given two nodes S (source) and T (target), find a partition of the nodes of the graph into two groups U and V such that:

- S ∈ U and T ∈ V
- cut(U, V) is minimal

cut(U, V) is defined as the sum of the edges’ weight going from U to V.
Maximum Flow

In a directed network, each directed edge is assigned a flow, such that the sum of ingoing flows equals the sum of outgoing flows at all other nodes.

- The sum of ingoing flows minus the sum of outgoing flows at Source is \(-F\).
- The sum of ingoing flows minus the sum of outgoing flows at Target is \(+F\).

Note: those two conditions are equivalent, given the first condition.

Additionally, each directed edge is assigned a capacity, and the flow on each edge cannot be larger than that capacity.

The number \(F\) is called the flow.

Maximum flow problem: Maximize \(F\)

Max \(F\) = total weight of the edges in the minimum cut

Graph Cut: \( V = X \cup Y \)

Min cut problem:
\[
\min_{X \cup Y = V} \text{cut}(X, Y)
\]

Favors imbalanced cuts!

\[
\text{cut}(X, Y) = |\{(i,j) \mid i \in X, j \in Y\}|
\]

= 11

= 2
**Ratio Cuts**

\[
\text{RatioCut}(X, Y) = \text{cut}(X, Y) \left( \frac{1}{|X|} + \frac{1}{|Y|} \right)
\]

\[
\min_{X \cup Y = V} \text{RatioCut}(X, Y)
\]

Takes into account cluster size!

\[
= 11 \times \left( \frac{1}{16} + \frac{1}{18} \right) \approx 1.30
\]

\[
= 2(\frac{1}{1} + \frac{1}{33}) \approx 2.06
\]
Represent a Cut as a Vector

Given $X \cup Y = V$

Let $u \in \mathbb{R}^{|V|}$ with

$$u_i = + \sqrt{|Y| / |X|} \text{ when } i \in X$$
$$u_i = - \sqrt{|X| / |Y|} \text{ when } i \in Y$$

Then:

$$u^T L u = 2 |V| \text{ RatioCut}(X, Y)$$

with $L = D - A$ Unnormalized Laplacian

Let $U$ be the set of vectors of the form defined above.

$$\min_{u \in \mathbb{R}^{|V|}} u^T L u$$

s.t. $u \in U$, $u \perp 1$

Cf: page 9 of ref [1]
Relaxation: „Forget“ the Constraint on $u$

Problem (a)

$$\min_{u \in \mathbb{R}^{|V|}} u^T L u$$
subject to $u \in U$, $u \perp 1$

Problem (b)

$$\min_{u \in \mathbb{R}^{|V|}} u^T L u \quad \Rightarrow$$
subject to $u \perp 1$

$u$ is the eigenvector of $L$ with smallest nonzero eigenvalue
Check out

Ref [1]:

Ulrike von Luxburg
A Tutorial on Spectral Clustering
Technical Report No. TR-149
Max Planck Institute for Biological Cybernetics
1 August 2006
http://www.cs.cmu.edu/~aarti/Class/10701/readings/Luxburg06_TR.pdf
Context
Homophily

Your friends tend to be like you in terms of

• Culture
• Languages
• Age
• Places they live
• Occupations
• Wealth
• Hobbies
• Religions
• Opinions
• etc.

Note: « tend »
Mechanisms

The underlying mechanism is

• Selection of friends
  • Active choice to select people with similar characteristics as friends
  • Implicit selection
    • E.g. living in same neighborhood
• Selection of behavior
  • Change your behavior to match that of friends (if possible)
Homophily in High School Network

Clustering by grade

Clustering by « race »

Color of node: « race » of student

names as friend

(Race, School Integration, and Friendship Segregation in America, American J. of Sociology 107 679–716, 2001)
Homophily Measures

Assuming probability $p$ for one property value (e.g. gender = male) and probability $(1 - p)$ for another value of the same property (e.g. gender = female)

Indication for **homophily**:
• Probability for cross-value links significantly lower than $2p(1 - p)$?
  • E.g. if boys are mostly friends with boys

Indication for **inverse homophily**
• Probability for cross-value links significantly higher than $2p(1 - p)$
  • E.g. for heterosexual romantic relationships with respect to gender

Otherwise: Indication for **no homophily**
Statistical Test of Homophily: Smokers and Non-smokers

\[ n = 5 + 7 = 12 \]
\[ m = 16 + 5 = 21 \]
Null Hypothesis

Null hypothesis: The 21 edges are distributed randomly among all possible node pairs.

Number of possible edges:
- Homophilic edges: $10 + 21 = 31$
- Non-homophilic edges: 35

Total number of possible edges: $\binom{5 + 7}{2} = 31 + 35 = 66$
Probabilities under the Null Hypothesis

Number of ways to place 21 edges among 66 possible positions: \( \binom{66}{21} \)

Number of ways to place 21 edges among 66 possible positions, such that \( k \) are among the 31 possible homophilic positions:

\[
\binom{31}{k} \binom{35}{21-k}
\]

Probability that exactly \( k \) edges are homophilic:

\[
f(k) = \frac{\binom{31}{k} \binom{35}{21-k}}{\binom{66}{21}}
\]

Probability that at least 16 out of 21 edges are homophilic:

\[
p = \sum_{k=16}^{21} f(k) \approx 0.0013
\]

\( p \) is small enough (< 5%) that we can reject the null hypothesis. Hence the network is homophilic, i.e., friends between smokers and non-smokers are less likely than friendships between similar persons.
Selection vs Social Influence

Select people as friends who behave like you

Align your behavior and preferences to those of your friends

Selection

Social influence / socialization
Selection vs Social Influence

Choosing other smokers as friends

Start smoking because your friends do

Align your behavior and preferences to those of your friends

Select people as friends who behave like you
Representing Context

So far: labeled unimodal networks, i.e., networks with one type of nodes, typically persons, labeled by their properties.

Now: bimodal networks, may be used for representing persons (one type of nodes) and for representing contexts (second type of nodes).
Context as a Bipartite Network

- The network is not labeled
- All information is coded as edges
- Network has two types of edges: **friendship** and **behavior**
- The behavior network is bipartite
Homophily as Clustering

Homophily expressed in terms of network properties:

- Triangles with two behavior edges and one friendship edges are prevalent

If this triangle happens more often than by chance, the network is homophilic

If this triangle happens less often than by chance, the network is inversely homophilic
Homophily as Triangle Closing

« Choose another smoker as friend »
(focal closure)

« Start smoking because your friend does »
(social influence)
Alternative representation

(a) Sample Semantic Web Graph

(b) Tensor Representation

Thank you for your attention!