Community Structure in Networks
Observations

- Small diameter, Edge clustering
- Patterns of signed edge creation
- Viral Marketing, Blogosphere, Memetracking
- Scale-Free
- Densification power law, Shrinking diameters
- Strength of weak ties, Core-periphery

Models

- Erdös-Renyi model, Small-world model
- Structural balance, Theory of status
- Independent cascade model, Game theoretic model
- Preferential attachment, Copying model
- Microscopic model of evolving networks
- Kronecker Graphs

Algorithms

- Decentralized search
- Models for predicting edge signs
- Influence maximization, Outbreak detection, LIM
- PageRank, Hubs and authorities
- Link prediction, Supervised random walks
- Community detection: Girvan-Newman, Modularity
Networks & Communities

- We often think of networks “looking” like this:

- What lead to such conceptual picture?
Networks: Flow of Information

- How information flows through the network?
  - What structurally distinct roles do nodes play?
  - What roles do different links (short vs. long) play?
- How people find out about new jobs?
  - Mark Granovetter, part of his PhD in 1960s
  - People find the information through personal contacts
- But: Contacts were often acquaintances rather than close friends
  - This is surprising: One would expect your friends to help you out more than casual acquaintances
- Why is it that acquaintances are most helpful?
Two perspectives on friendships:

- **Structural**: Friendships span different parts of the network
- **Interpersonal**: Friendship between two people is either strong or weak

**Structural role: Triadic Closure**

If two people in a network have a friend in common, then there is an increased likelihood they will become friends themselves.

Which edge is more likely a-b or a-c?
Granovetter’s Explanation

- **Granovetter makes a connection between social and structural role of an edge**

- **First point: Structure**
  - Structurally embedded edges are also socially strong
  - Long-range edges spanning different parts of the network are socially weak

- **Second point: Information**
  - Long-range edges allow you to gather information from different parts of the network and get a job
  - Structurally embedded edges are heavily redundant in terms of information access
Triadic Closure

- Triadic closure == High clustering coefficient

Reasons for triadic closure:
- If $B$ and $C$ have a friend $A$ in common, then:
  - $B$ is more likely to meet $C$
    - (since they both spend time with $A$)
  - $B$ and $C$ trust each other
    - (since they have a friend in common)
  - $A$ has incentive to bring $B$ and $C$ together
    - (as it is hard for $A$ to maintain two disjoint relationships)

- Empirical study by Bearman and Moody:
  - Teenage girls with low clustering coefficient are more likely to contemplate suicide
Define: **Bridge edge**
- If removed, it disconnects the graph

Define: **Local bridge**
- Edge of Span > 2
  - (Span of an edge is the distance of the edge endpoints if the edge is deleted. Local bridges with long span are like real bridges)

Define: Two types of edges:
- **Strong** (friend), **Weak** (acquaintance)

Define: **Strong triadic closure**:
- Two strong ties imply a third edge

**Fact:** If strong triadic closure is satisfied then **local bridges are weak ties**!
**Claim:** If node $A$ satisfies **Strong Triadic Closure** and is involved in at least two strong ties, then any **local bridge** adjacent to $A$ must be a **weak tie**.

**Proof by contradiction:**
- $A$ satisfies **Strong Triadic Closure** and has 2 strong ties
- Let $A - B$ be **local bridge** and a **strong** tie
- Then $B - C$ must exist because of **Strong Triadic Closure**
- But then $A - B$ is **not a bridge**! (since $B-C$ must be connected due to Strong Triadic Closure property)
For many years the Granovetter’s theory was not tested

But, today we have large who-talks-to-whom graphs:
- Email, Messenger, Cell phones, Facebook

Onnela et al. 2007:
- Cell-phone network of 20% of country’s population
- Edge strength: # phone calls
Neighborhood Overlap

- **Edge overlap:**

\[
O_{ij} = \frac{N(i) \cap N(j)}{N(i) \cup N(j)}
\]

- \(N(i)\) ... a set of neighbors of node \(i\)

- **Overlap = 0** when an edge is a local bridge
Cell-phone network

Observation:

- Highly used links have high overlap!

Legend:

- True: The data
- Permuted strengths: Keep the network structure but randomly reassign edge strengths
Real edge strengths in mobile call graph

- Strong ties are more embedded (have higher overlap)
Same network, same set of edge strengths but now **strengths are randomly shuffled**
Link Removal by Strength

- Removing links by **strength (#calls)**
  - Low to high
  - High to low

Conceptual picture of network structure

Low disconnects the network sooner
Link Removal by Overlap

- Removing links based on **overlap**
  - Low to high
  - High to low

Low disconnects the network sooner.
Granovetter’s theory leads to the following conceptual picture of networks

- Strong ties
- Weak ties
Small Detour:
Structural Holes
Small Detour: Structural Holes

[Diagram of a social network with nodes labeled A, B, C, Robert, and James, illustrating structural holes.]
Structural Holes provide ego with access to novel information, power, freedom.
The “network constraint” measure [Burt]:
- To what extent are person’s contacts redundant
  - **Low**: disconnected contacts
  - **High**: contacts that are close or strongly tied

\[ p_{uv} = \frac{1}{d_u} \]

\[ c_i = \sum_j c_{ij} = \sum_j \left[ p_{ij} + \sum_k \left( p_{ik} p_{kj} \right) \right]^2 \]

\[ p_{uv} \ldots \text{prop. of } u\text{’s “energy” invested in relationship with } v \]
Example: Robert vs. James

Network constraint:
- James: $c_J = 0.309$
- Robert: $c_R = 0.148$

Constraint: To what extent are person’s contacts redundant
- Low: disconnected contacts
- High: contacts that are close or strongly tied
Spanning Holes Matters

A. Probability of Poor Evaluation (3.3 logit t-test)
   - Probability of Outstanding Evaluation (-2.3 logit t-test)

B. Network Constraint
   - Many Structural Holes
   - Few
   (manager C above, mean C in team below)

C. Early Promotion
   \[ Y = 2.035 - 0.074(C) \]
   \[ r = -0.40 \]
   \[ t = -5.4 \]
   \[ P < .001 \]

D. Recognition of TQM Team Achievements
   \[ Y = 7.525 - 0.210(C) \]
   \[ r = -0.79 \]
   \[ t = -4.3 \]
   \[ P < .001 \]

E. Large Bonus
   \[ Y = 0.438 - 0.022(C) \]
   \[ r = -0.30 \]
   \[ t = -3.7 \]
   \[ P < .001 \]

F. Probability of Small Bonus
   \[ Y = 2.102 - 0.077(C) \]
   \[ r = -0.44 \]
   \[ t = -3.7 \]
   \[ P < .001 \]

French Salary

[Ron Burt]
Network Communities
Granovetter’s theory (and common sense) suggest that networks are composed of tightly connected sets of nodes.

- **Network communities:**
  - Sets of nodes with *lots* of connections *inside* and *few* to *outside* (the rest of the network)
How to automatically find such densely connected groups of nodes?

Ideally such automatically detected clusters would then correspond to real groups

For example:
Zachary’s Karate club network:
- Observe social ties and rivalries in a university karate club
- During his observation, conflicts led the group to split
- Split could be explained by a minimum cut in the network
Find micro-markets by partitioning the “query x advertiser” graph:
Can we identify node groups? (communities, modules, clusters)

Nodes: Teams
Edges: Games played
NCAA Football Network

Nodes: Teams
Edges: Games played

NCAA conferences

Jure Leskovec, Stanford
Can we identify social communities?

Nodes: Users
Edges: Friendships
Facebook Ego-network

Nodes: Users
Edges: Friendships

Social communities

High school

Company

Stanford (Basketball)

Stanford (Squash)
Protein-Protein Interactions

Can we identify functional modules?

Nodes: Proteins
Edges: Interactions
Protein-Protein Interactions

Nodes: Proteins
Edges: Interactions

Functional modules
Community Detection

How to find communities?

We will work with **undirected** (unweighted) networks
Method 1: Strength of Weak Ties

- **Edge betweenness**: Number of shortest paths passing over the edge
- **Intuition**:

![Image of network with edge betweenness and edge strengths](image)

**Edge strengths (call volume) in real network**

**Edge betweenness in real network**

- $b=16$
- $b=7.5$
Divisive hierarchical clustering based on the notion of edge **betweenness:**

Number of shortest paths passing through the edge

**Girvan-Newman Algorithm:**
- Undirected unweighted networks

**Repeat until no edges are left:**
- Calculate betweenness of edges
- Remove edges with highest betweenness

Connected components are communities

Gives a hierarchical decomposition of the network
Girvan-Newman: Example

Need to re-compute betweenness at every step
Girvan-Newman: Example

Step 1:

Step 2:

Hierarchical network decomposition:

Step 3:
Communities in physics collaborations
Girvan-Newman: Results

- Zachary’s Karate club:
  Hierarchical decomposition
We need to resolve 2 questions

1. How to compute betweenness?
2. How to select the number of clusters?
How to Compute Betweenness?

- Want to compute betweenness of paths starting at node $A$

- Breath first search starting from $A$:

```
A
  /  \
 /    \  \
B --- C --- I
  |    |    |
  E    D    G

A
  /  \
 /    \  \
B --- C --- D
  |    |    |
  E    G    I

A
  /  \
 /    \  \
B --- C --- D
  |    |    |
  E    G    I

A
  /  \
 /    \  \
B --- C --- D
  |    |    |
  E    G    I

A
  /  \
 /    \  \
B --- C --- D
  |    |    |
  E    G    I
```
Count the number of shortest paths from $A$ to all other nodes of the network:
How to Compute Betweenness?

- **Compute betweenness by working up the tree:** If there are multiple paths count them fractionally.

The algorithm:
- Add edge flows:
  - node flow = \(1 + \sum \text{child edges}\)
  - split the flow up based on the parent value
- Repeat the BFS procedure for each starting node \(U\)

---

1. **Add edge flows:**
   - node flow = \(1 + \sum \text{child edges}\)
   - split the flow up based on the parent value
2. **Repeat the BFS procedure for each starting node \(U\)**

---

Diagram:
- **A**
  - **B** (1)
  - **C** (1)
  - **D** (1)
  - **E** (1)
  - **F** (2)
  - **G** (1)
  - **H** (2)
  - **I** (3)
  - **J** (3)
  - **K** (6)

Paths:
- 1+1 paths to H (Split evenly)
- 1+0.5 paths to J (Split 1:2)
- 1 path to K (Split evenly)
We need to resolve 2 questions

1. How to compute betweenness?
2. How to select the number of clusters?
Communities: sets of tightly connected nodes

Define: Modularity $Q$

- A measure of how well a network is partitioned into communities
- Given a partitioning of the network into groups $s \in S$:

$$Q \propto \sum_{s \in S} [ (# \text{edges within group } s) - \text{(expected # edges within group } s) ]$$

Need a null model!
Given real $G$ on $n$ nodes and $m$ edges, construct rewired network $G'$

- Same degree distribution but random connections
- Consider $G'$ as a multigraph

The expected number of edges between nodes $i$ and $j$ of degrees $k_i$ and $k_j$ equals to: $k_i \cdot \frac{k_j}{2m} = \frac{k_i k_j}{2m}$

The expected number of edges in (multigraph) $G'$:

$= \frac{1}{2} \sum_{i \in N} \sum_{j \in N} \frac{k_i k_j}{2m} = \frac{1}{2} \cdot \frac{1}{2m} \sum_{i \in N} k_i \left( \sum_{j \in N} k_j \right) = \frac{1}{4m} 2m \cdot 2m = m$

Note: $\sum_{u \in N} k_u = 2m$
Modularity of partitioning $S$ of graph $G$:

- $Q \propto \sum_{s \in S} \left[ (\text{# edges within group } s) - (\text{expected # edges within group } s) \right]$

- $Q(G, S) = \frac{1}{2m} \sum_{s \in S} \sum_{i \in s} \sum_{j \in s} \left( A_{ij} - \frac{k_i k_j}{2m} \right)$

Normalizing cost.: $-1 < Q < 1$

Modularity values take range $[-1, 1]$

- It is positive if the number of edges within groups exceeds the expected number
- $0.3 < Q < 0.7$ means significant community structure
Modularity is useful for selecting the number of clusters:

Next time: Why not optimize Modularity directly?