LECTURE 5:COMMUNITY STRUCTURE IN NETWORKS

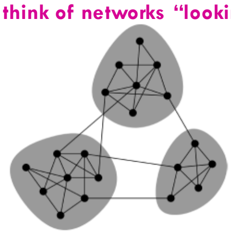
COMP4641: Social Information Network Analysis and Engineering
Wednesday February 25th 2015

Announcement

- Project
  - 4 March: Group list and tentative title/topic
  - 15 March: Project proposal
  - 3 April: Project milestone report
  - 3 May: Final report
  - Study week?: Final presentation

Networks & Communities

- **We often think of networks "looking" like this:**



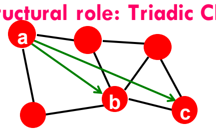
- **What lead to such conceptual picture?**

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?**

Granovetter's Answer

- **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**

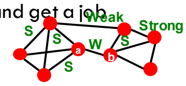


Which edge is more likely a-b or a-c?

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

Granovetter's Explanation

- **Granovetter makes a connection between social and structural role of an edge**
- **First point:**
  - Structurally embedded edges are also socially strong
  - Edges spanning different parts of the network are socially weak
- **Second point:**
  - The 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

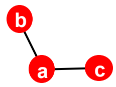


Triadic Closure

- **Triadic closure == High clustering coefficient**

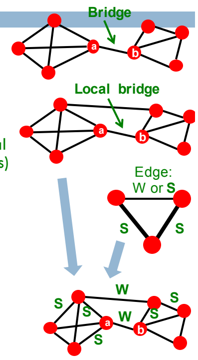
Reasons for triadic closure:

- 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



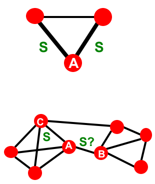
Granovetter's Explanation

- 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!**



Local Bridges and Weak ties

- **Claim:** If node satisfies **Strong Triadic Closure** and is involved in at least **two strong ties,** then any **local bridge** adjacent to must be a **weak tie.**
- **Proof by contradiction:**
  - satisfies **Strong Triadic Closure**
  - Let be local bridge and a strong tie
  - Then must exist because of **Strong Triadic Closure**
  - But then is **not a bridge!**



Tie strength in real data

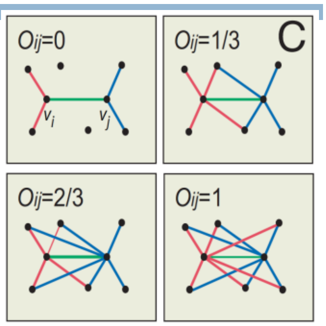
- **For many years the Granovetter's theory was not tested**
- But, today we have large who-talks-to-whom graphs:
  - Email, Messenger, Cell phones, Facebook
- **Onnela et al. 2007:**
  - Cell-phone network of 20% of country's population
  - **Edge strength:** # phone calls

Neighborhood Overlap

- **Edge overlap:**

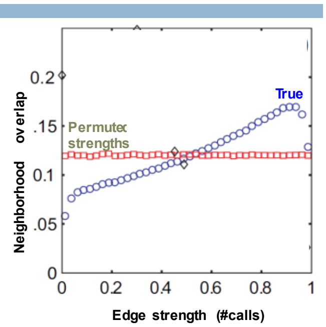
$$O_{ij} = \frac{N(i) \cap N(j)}{N(i) \cup N(j)}$$

  - $N(i)$ … a set of neighbors of node i
- Overlap = 0 when an edge is a **local bridge**



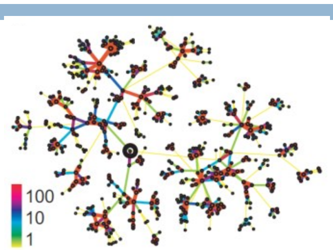
Phones: Edge Overlap vs. Strength

- **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 Network, Real Tie Strengths

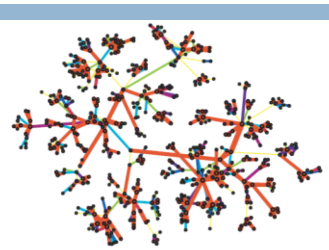
13



- **Real edge strengths in mobile call graph**
  - Strong ties are more embedded (have higher overlap)

Real Net, Permuted Tie Strengths

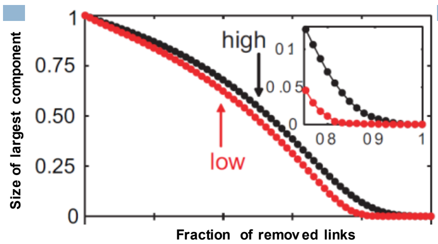
14



- **Same network, same set of edge strengths** but now **strengths are randomly shuffled**

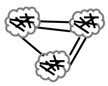
Link Removal by Strength

15



Low disconnects the network sooner

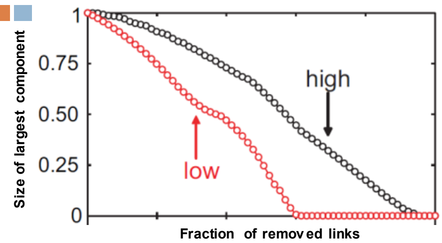
- Removing links by **strength (#calls)**
  - Low to high
  - High to low



Conceptual picture of network structure

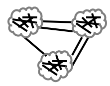
Link Removal by Overlap

16



Low disconnects the network sooner

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

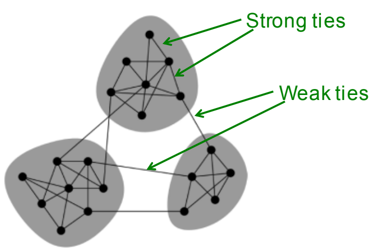


Conceptual picture of network structure

Conceptual Picture of Networks

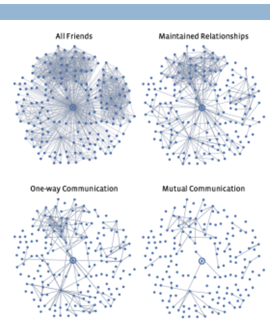
17

- **Granovetter's theory leads to the following conceptual picture of networks**



Facebook User's Tie Strength

18



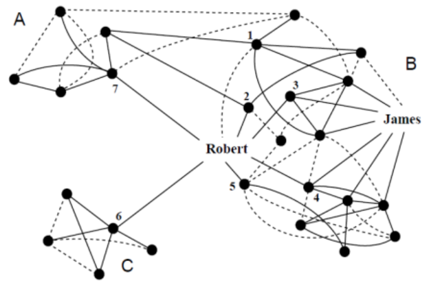
Four different views of a Facebook User's network neighborhood, show the structure of links corresponding respectively to all declared friendships, maintained relationships, one-way communication, and reciprocal communication

19

SMALL DETOUR: STRUCTURAL HOLES

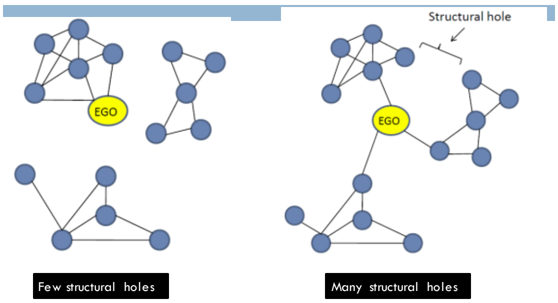
Small Detour: Structural Holes

20



Structural Holes

21

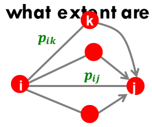


Structural Holes provide ego with access to novel information, power, freedom

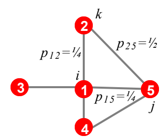
Structural Holes: Network Constraint

22

- The "network constraint" measure [Burt]:
  - To what extent are person's contacts redundant



$p_{uv} = 1/du$



- **Low**: disconnected contacts
- **High**: contacts that are close or strongly tied

p_{uv}

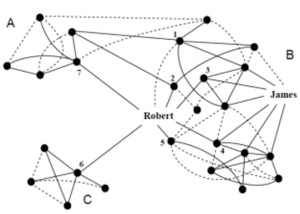
	1	2	3	4	5
1	.00	.25	.25	.25	.25
2	.50	.00	.00	.00	.50
3	1.0	.00	.00	.00	.00
4	.50	.00	.00	.00	.50
5	.33	.33	.00	.33	.00

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

p_{uv} … prop. of u's "energy" invested in relationship with v

Example: Robert vs. James

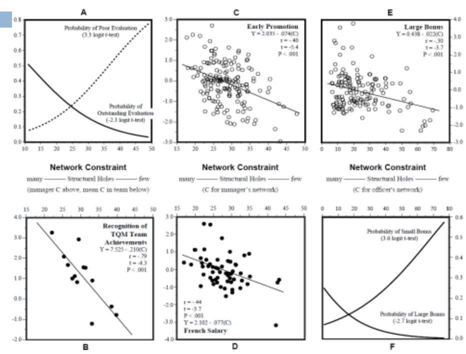
23



- **Constraint:** To what extent are person's contacts redundant
  - Low: disconnected contacts
  - High: contacts that are close or strongly tied
- **Network constraint:**
  - James:
  - Robert:

Spanning the Holes Matters

24



25

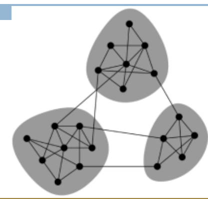
NETWORK COMMUNITIES

02-Mar-15

Network Communities

26

- Granovetter's theory (and common sense) suggest that networks are composed of **tightly connected sets of nodes**



Communities, clusters, groups, modules

- **Network communities:**
  - Sets of nodes with **lots** of connections **inside** and **few** to **outside** (the rest of the network)

Finding Network Communities

27

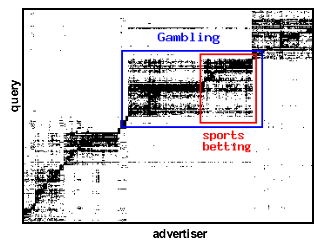
- **How to automatically find such densely connected groups of nodes?**
- Ideally such automatically detected clusters would then correspond to real groups
- For example:

Communities, clusters, groups, modules

Micro-Markets in Sponsored Search

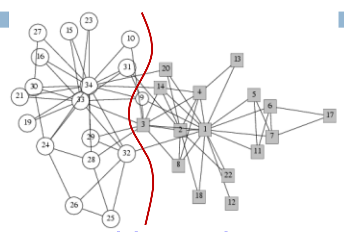
28

Find micro-markets by partitioning the "query x advertiser" graph:



Social Network Data

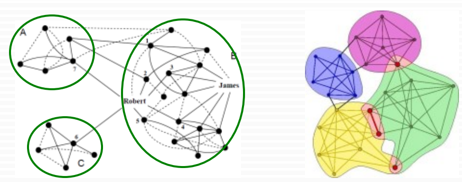
29



- **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

30 Community Detection

How to find communities?

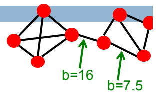


We will work with **undirected** (unweighted) networks

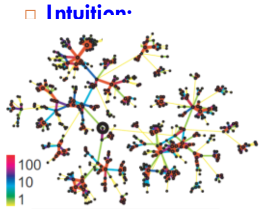
Method 1: Strength of Weak Ties

31

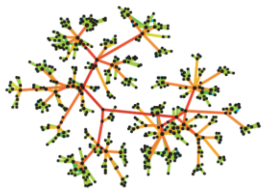
- **Edge betweenness:** Number of shortest paths passing over the edge



- **Intuition:**



Edge strengths (call volume) in real network



Edge betweenness in real network

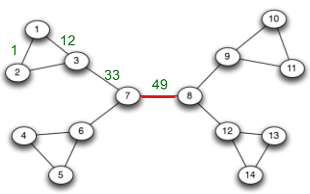
Method 1: Girvan-Newman

32

- 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

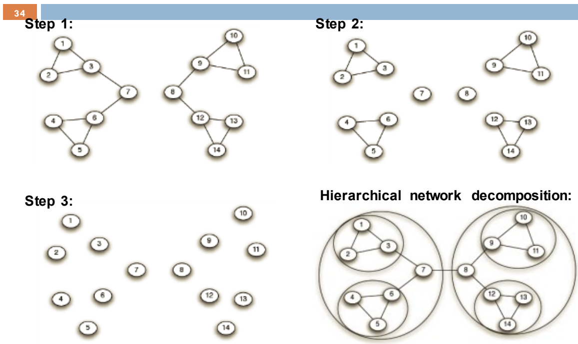
33



Need to re-compute betweenness at every step

Girvan-Newman: Example

34



Girvan-Newman: Results

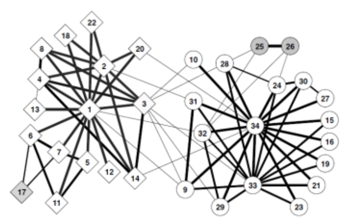
35

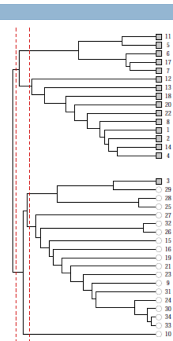
Communities in physics collaborations

Girvan-Newman: Results

36

- **Zachary's Karate club:** Hierarchical decomposition

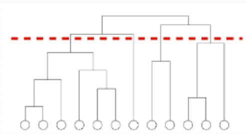




We need to resolve 2 questions

37

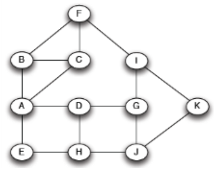
1. How to compute betweenness?
2. How to select the number of clusters?



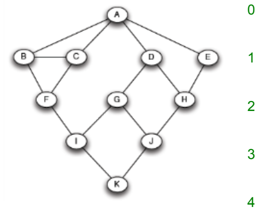
How to Compute Betweenness?

38

- Want to compute betweenness of paths starting at node A



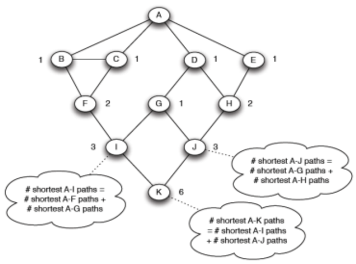
- Breath first search starting from A:



How to Compute Betweenness?

39

- Count the number of shortest paths from A to all other nodes of the network:

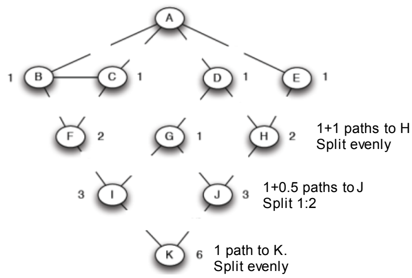


How to Compute Betweenness?

40

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

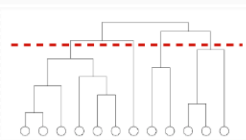
The algorithm:
- Add edge **flows**:
  - node flow = 1+∑child edges
  - split the flow up based on the parent value
- Repeat the BFS procedure for each starting node U



We need to resolve 2 questions

42

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

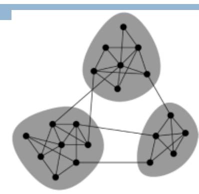


Network Communities

43

- **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 $s \in S$:

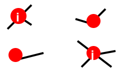
$$Q \propto \sum_{s \in S} [\,(\#\text{ edges within group } s) - \underbrace{(\text{expected}\#\text{ edges within group } s)}_{\text{Need a null model!}}\,]$$



Null Model: Configuration Model

44

- **Given real on nodes and edges, construct rewired network**
  - Same degree distribution but random connections
  - Consider as a multigraph
  - **The expected number of edge 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

45

- **Modularity of partitioning S of graph G:**
  - $Q \propto \sum_{s \in S}$ [(# edges within group s) – (expected # edges within group s)]
  - $Q(G,S) = \underbrace{\frac{1}{2m}}_{\text{Normalizing cost.: -1<Q<1}} \sum_{s \in S}\sum_{i \in s}\sum_{j \in s}\left(A_{ij} - \frac{k_i k_j}{2m}\right)$

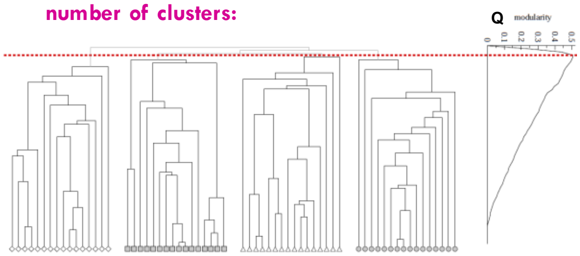
A_{ij} = 1 if i→j, 0 else

- **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: Number of clusters

46

- **Modularity is useful for selecting the number of clusters:**



Why not optimize modularity directly?