Subgraphs and Community Structure of Networks

Saptarshi Ghosh
Department of CSE, IIT Kharagpur
Social Computing course, CS60017

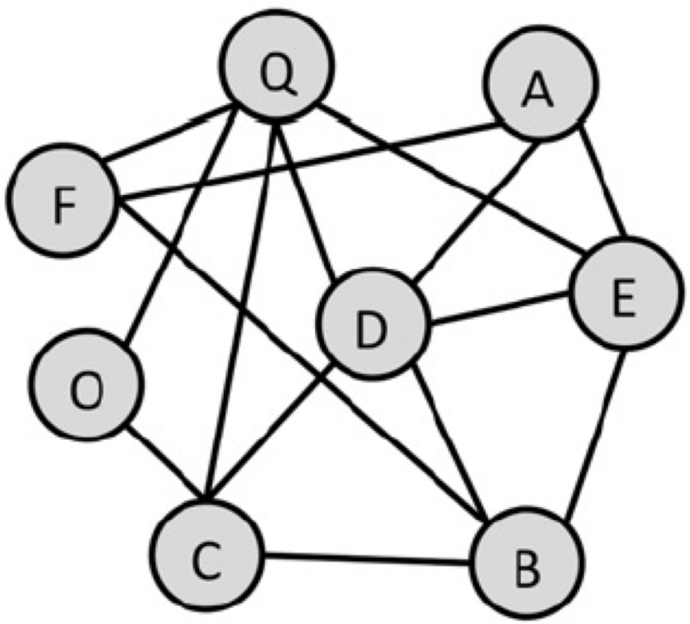
Subgraphs of interest

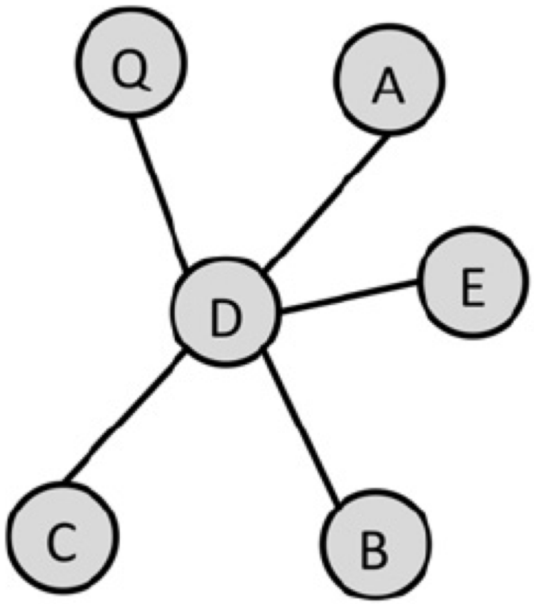
- Given a (social) network, what are some subgraphs of interest?
  - From the perspective of an individual user – Egocentric networks
  - From the perspective of the network as a whole or the network administrators – Communities or clusters

- Lots of applications of these subgraphs of interest – recommendation, summarization, …

Egocentric networks

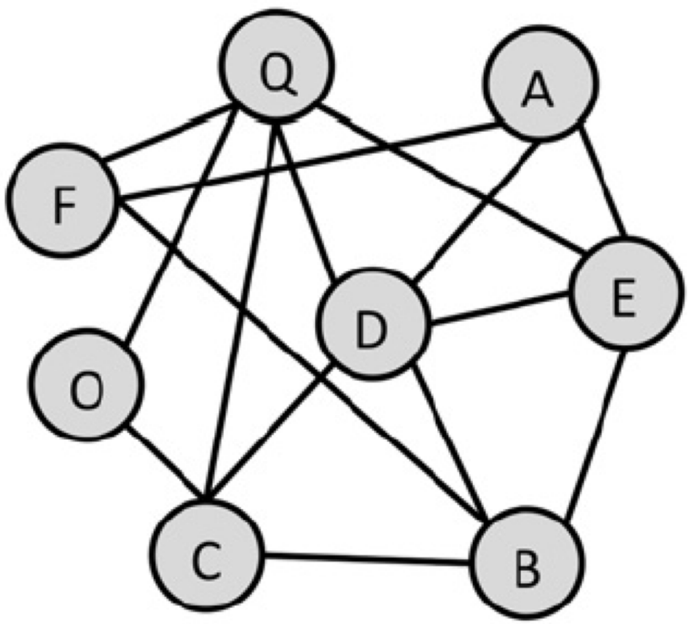
- Interesting from the perspective of a node (user)
- 1-degree egocentric network: a node and all its connections to its neighbors

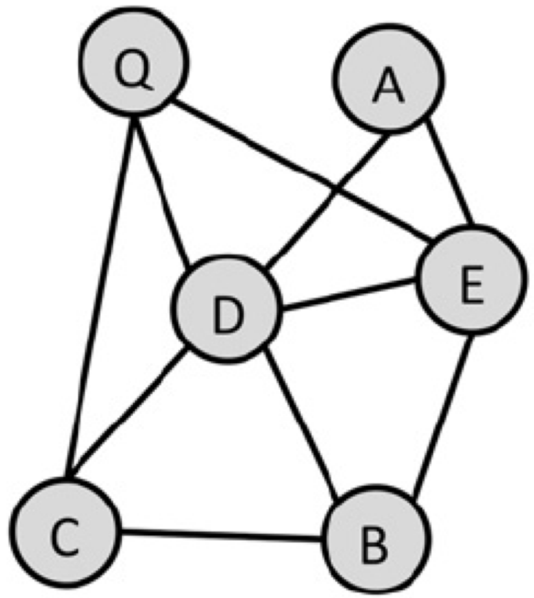




Egocentric networks

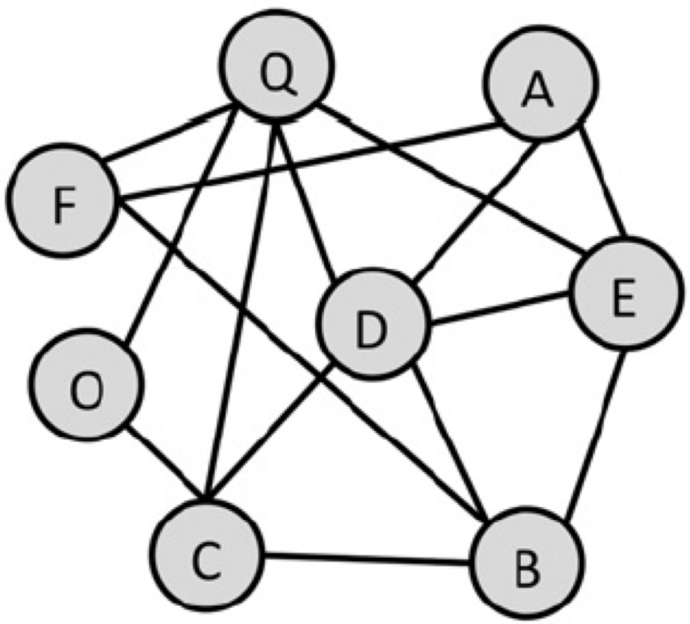
- 1.5-degree egocentric network: a node, all its connections to its neighbors, and the connections among the neighbors

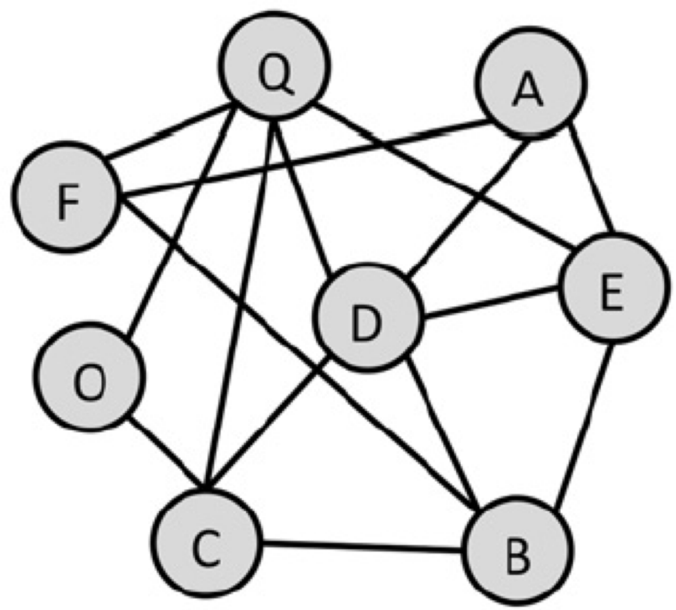




Egocentric networks

- 2-degree egocentric network: a node, all its neighbors, all neighbors of neighbors, and the connections among all these nodes





Communities

- Community or network cluster
  - Typically a group of nodes having more and / or better interactions among its members, than between its members and the rest of the network
  - No unique formal definition

- Community Detection (CD) -- automatically detecting communities in a network
- Challenging
  - Communities are not well-defined
  - Number of communities in a network is not known

Different types of CD algorithms

- Detection of disjoint communities
  - Each community is a partition of the network
- Detection of overlapping communities
  - A node can be members of multiple communities

- CD algorithms that rely only on network structure
- CD algorithms that rely on network structure and content (e.g., content posted by users)

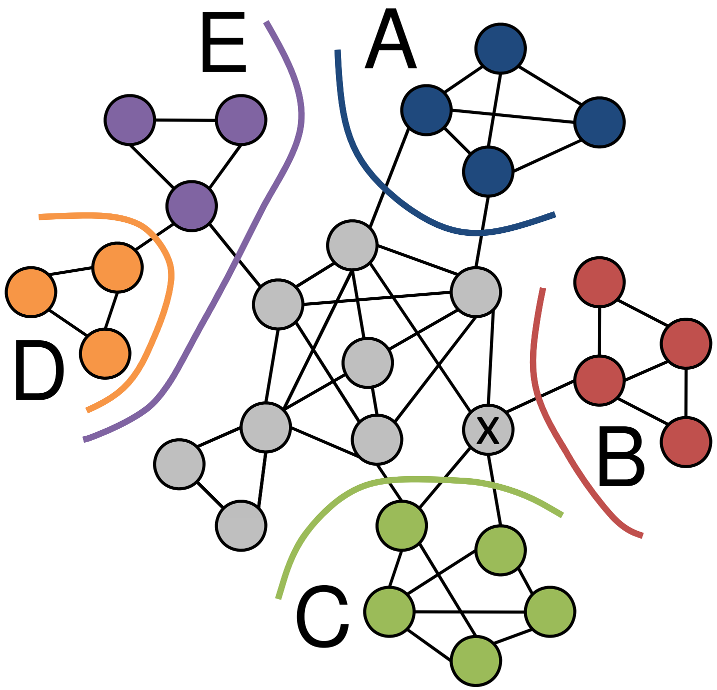
Our focus

- We are primarily focusing on
  - Algorithms that rely only on the network structure
  - Algorithms for detection of disjoint communities

- A case-study at the end will discuss detection of overlapping topical communities on Twitter, utilizing both network and content

What is the output of a CD algorithm?

- A community structure – a set of communities
  - Communities in this set may be disjoint partitions or overlapping



How to evaluate a CD algorithm?

- Assume a known community structure $X = \{x_1, x_2, \ldots, x_I\}$
- An algorithm finds a community structure $Y = \{y_1, y_2, \ldots, y_J\}$
- How close is Y to X? Note: |X| may be different from |Y|
- Several existing measures
  - Purity
  - Rand index
  - Normalized Mutual Information (NMI) [has been extended to overlapping communities]
- Additional reference:
  - Generalized Measures for the Evaluation of Community Detection Methods, by Labatut (https://arxiv.org/abs/1303.5441)

AN EARLY COMMUNITY DETECTION ALGORITHM

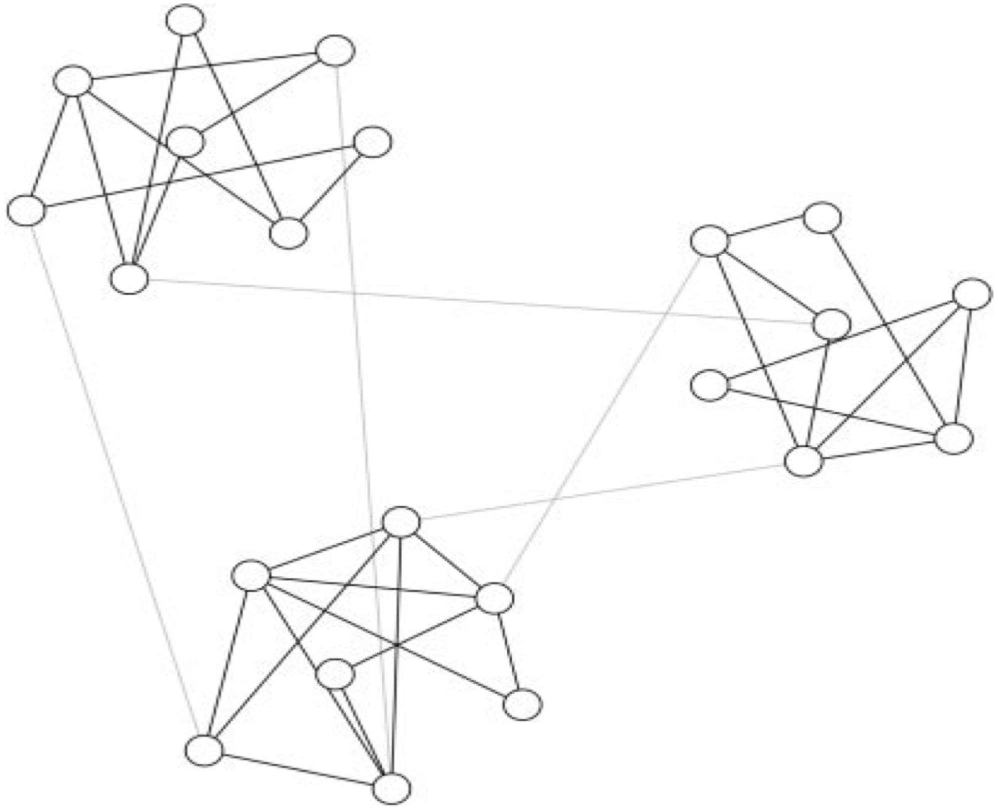
Community structure in social and biological networks
PNAS, 2002

Algorithm by Girvan & Newman

- Focus on edges that are most “between” communities
  - Edge betweenness of an edge *e* : fraction of shortest paths between all pairs of vertices, which run through *e*
  - Edges between communities are likely to have high edge betweenness centrality
- Idea of this algorithm
  - Progressively remove edges having high betweenness centrality, to separate communities from one another

Algorithm by Girvan & Newman

- Focus on edges that are most “between” communities



Girvan-Newman algorithm

1. Compute betweenness centrality for all edges
2. Remove the edge with highest betweenness centrality
3. Re-compute betweenness centrality for all edges affected by the removal
4. Repeat steps 2 and 3 until no edges remain

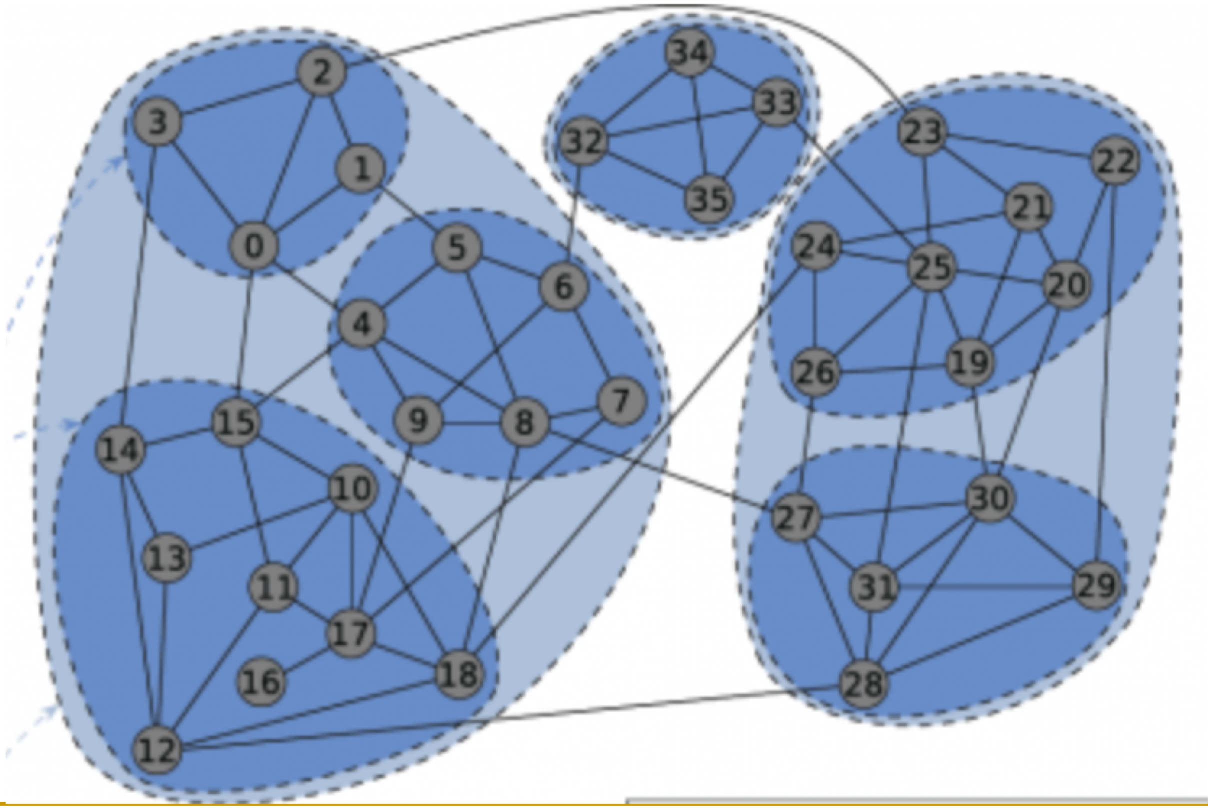
What will be the output of this algorithm?

NOT a single community structure (a set of communities)

Rather, this algorithm outputs many possible community structures. We have to choose one of the community structures.

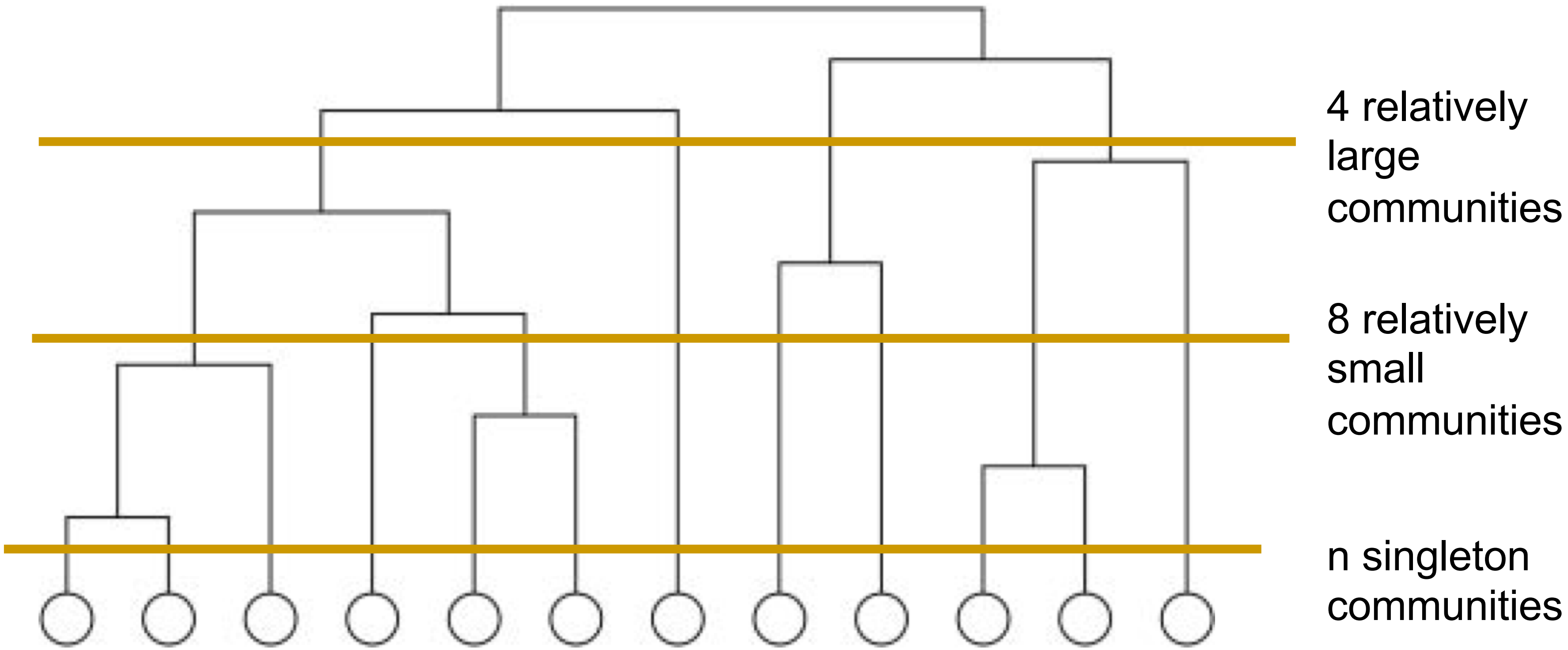
What is a good community structure?

- Community structure of a graph is hierarchical, with smaller communities nested within larger ones



Dendrogram

- Hierarchical community structure represented as a hierarchical clustering tree: dendrogram
- A "slice" through the tree at any level gives a certain community structure



What is a good community structure?

- At which level to slice the dendrogram?
  - A few large communities, or many small communities?
  - Often depends on the end application

- Need an objective function to measure the "goodness" of a community structure

OBJECTIVE FUNCTIONS FOR COMMUNITY DETECTION

Empirical Comparison of Algorithms for Network Community Detection, Leskovec et al., WWW 2010

Objective functions for CD

- Community or network cluster (recap)
  - Typically a group of nodes having more and / or better interactions among its members, than between its members and the rest of the network

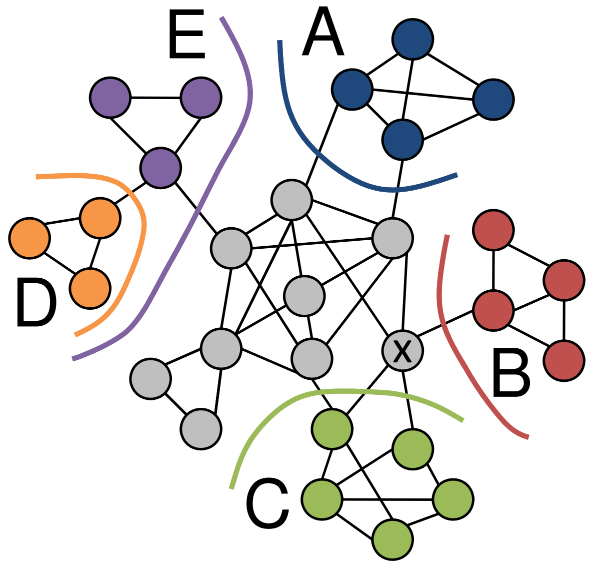
- Two criteria of interest for measuring how well a particular set *S* of nodes represents a community
  - Number of edges among the nodes within *S*
  - Number of edges between nodes in *S* and rest of network

Two types of objective functions

- Multi-criterion scores
  - Consider both the criteria for measuring quality of set S of nodes
  - Lower values of f(S) signify a more community-like set S
  - Examples: expansion, internal density, cut ratio, conductance, …
- Single-criterion scores
  - Consider only one of the criteria, usually the number of edges among the nodes within S
  - Example: Modularity

Notations

- $G = (V, E)$ is the network.
- $n = |V|$ = number of nodes
- $m = |E|$ = number of edges
- $d(u) = k_u$ = degree of node u



- S: set of nodes
- n_s = number of nodes in S
- m_s = number of edges within S (both nodes in S)
- c_s = number of edges on the boundary of S

Expansion

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

- Number of edges per node in S, that points outside the set S

n_s = number of nodes in S
m_s = number of edges within S (both nodes in S)
c_s = number of edges on the boundary of S

Internal density

$$f(S) = 1 - \frac{m_S}{n_S(n_S - 1)/2}$$

- Internal edge density of the set S

n_s = number of nodes in *S*
m_s = number of edges within *S* (both nodes in *S*)
c_s = number of edges on the boundary of S

Cut Ratio

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

- Fraction of all possible edges leaving the set S

n_s = number of nodes in S
m_s = number of edges within S (both nodes in S)
c_s = number of edges on the boundary of S

Conductance

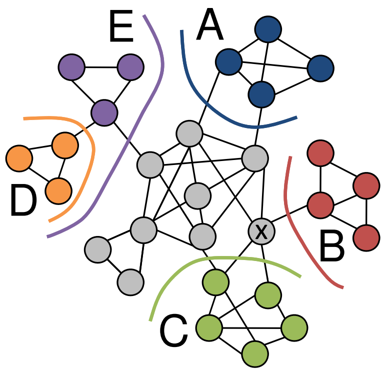
$$f(S) = \frac{c_S}{2m_S + c_S}$$

- Fraction of total edge volume of S that points outside the cluster
- Edge volume = sum of node-degrees

n_s = number of nodes in S
m_s = number of edges within S (both nodes in S)
c_s = number of edges on the boundary of S

How to use these objective functions?

- These objective functions measure how good a subset of nodes is, as a community
- Given a community structure Y = $\{y_1, y_2, \ldots, y_J\}$
  - Use an objective function to measure goodness of every community (subset of nodes) y_i
  - Measure the goodness of Y as a function (e.g., weighted linear combination) of the goodness of all y_i



Modularity-based measures

- A set of nodes is a good community if the number of edges within the set is significantly more than what can be expected by random chance

- Modularity $Q = 1/K * (m_s - E(m_s))$
  - Number of edges m_s within set S, minus expected number of edges $E(m_s)$ within the set S
  - K is a constant, used for normalization

Expected number of edges

- Null model: Erdos-Renyi random network having the same node degree sequence as given network

- Randomized realization of a given network, realized in practice using Configuration Model
  - Cut each edge of the given network into two half-edges or stubs
  - Randomly connect each stub to any stub
  - Expected to have no community structure

Definition of Modularity Q

- For two particular nodes *i* and *j* :
  - Number of edges existing between the nodes: A_{ij}
  - Degrees: k_i and k_j
  - Probability that a particular stub of node *i* connects to some stub of node *j*: $p_{ij} = k_j / 2m$
  - Expected number of links between *i* and *j*: $k_i k_j / 2m$

- Do the nodes *i* and *j* have more edges than expected by random chance?

$$A_{ij} - k_i k_j / 2m$$

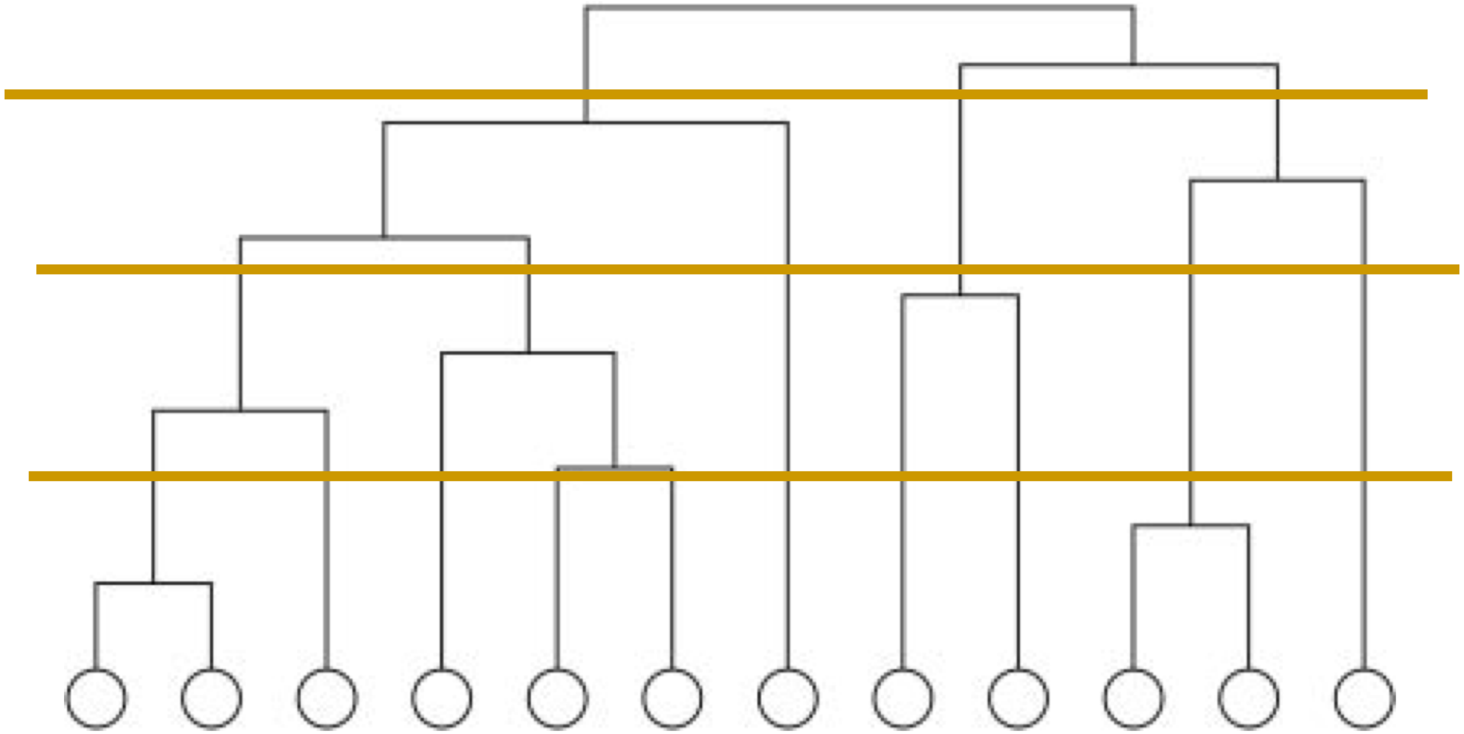
Q for a given community structure

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

- The delta function is 1 if both nodes *i* and *j* are in the same community ($C_i = C_j$), 0 otherwise
- Consider a network with two communities c1, c2
  - Q is the fraction of edges within c1 or c2, minus the expected number of edges within c1 and c2 for a random graph with the same node degree sequence as the given network
- More details: "Modularity and community structure in networks" by Newman (PNAS 2006)

Using modularity for CD

- Approach 1: use Modularity to decide at which level to slice the dendrogram



Using modularity for CD

- Approach 1: use Modularity to decide at which level to slice the dendrogram

- Approach 2: Optimize for modularity itself
  - Exhaustive maximization is NP-hard
  - Heuristics and approximations used
  - Several algorithms have been developed for optimizing Modularity

Most popular Q optimization algorithm

- Louvain algorithm:
  - https://perso.uclouvain.be/vincent.blondel/research/louvain.html
- Optimization in two steps
  - Step 1: look for small communities - optimizing Q locally
  - Step 2: aggregate nodes in the same community and build a new network whose nodes are the communities
  - Repeat iteratively until a maximum of modularity is attained and a hierarchy of communities is produced
  - Time: approx $O(n \log n)$

Additional reference

- Many subsequent works have suggested improvements for maximizing modularity
  - Reducing time complexity
  - Normalizing with number of edges to minimize bias towards larger communities
  - …
- Read “Community detection in graphs” by Fortunato, Physics Reports, 2010.

CASE STUDY: DIFFERENT TYPES OF GROUPS IN A SOCIAL NETWORK

Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale, Bhattacharya et al., ACM CSCW 2014

Different methods to identify groups

- Identifying groups based on network structure – community detection algorithms (what we have discussed till now)

- How about identifying groups in a social network based on content, e.g., text or profile attributes of users?

Identified topical groups in Twitter

Topical Groups = Experts + Seekers

Experts: Users who have expertise on the topic (List-based method

Seekers: Users who are interested in the topic (who follow several experts on a topic)

@BarackObama
Expert on Politics

@BarackObama
Seeker on Basketball

Identifying topical groups at scale

- Crawled data for first 38 million users in Twitter
- 88 Million lists, 1.5 Billion social links
- Identified 36 thousand topical groups

Diversity: Topics and Group Size

No. of seekers	Number of experts					
	< 100	100 – 500	500 – 1K	1K – 5K	5K – 10K	> 10K
< 1K	**(5416)** *geology*, *karate*, *malaria*, *neurology*, *tsunami*, psychiatry, radiology, pediatrics, dermatology, dentistry	**(132)** volleyball, philosophers, tarot, perfume, florists, copywriters, taxi, esperanto				
1K – 5K	**(915)** *biology*, *chemistry*, *swimmers*, astrophysics, multimedia, semiconductor, renewable-energy, breast-cancer, judaism	**(428)** *painters*, *astrology*, *sociology*, *geography*, *forensics*, anthropology, genealogy, archaeology, gluten, diabetes, neuroscience	**(17)** architects, insurance, second-life, police, progressives, creativity			
5K – 10K	**(166)** *malware*, gnu, robot, chicago-sports, gospel-music, space-exploration, wall-street	**(202)** horror, agriculture, atheism, attorneys, furniture, art-galleries, ubuntu	**(34)** *psychology*, poetry, catholic, hospitals, autism, jazz	**(2)** coffee, dealers		
10K – 50K	**(174)** ipod, ipad, virus, Liverpool-FC, choreographers, heavy-metal, backstreet-boys, world-cup,	**(312)** *olympics*, *physics*, *theology*, *earthquake*, opera, makeup, Adobe, wrestlers, typography, american-idol	**(146)** *tennis*, *linux*, *astronomy*, yoga, animation, manga, doctors, realtors, wildlife, rugby, forex, php, java,	**(67)** *law*, *history*, *beer*, *golf*, librarians, theatre, military, poker, conservatives, vegan		
50K–100K	**(7)** bbc-radio, UK-celebs, christian-leaders, superstars	**(61)** *hackers*, *programmers*, bicycle, GOP, fantasy-football, NCAA, wwe, sci-fi	**(35)** *medicine*, *cyclists*, investors, recipes, NHL, xbox, triathlon, Google	**(37)** hotels, museums, hockey, architecture, charities, weather, space		
> 100K	**(3)** headlines, brits	**(49)** pop-culture, gospel, BBC, reality-tv, bollywood	**(58)** *religion*, actresses, gadgets, graphic-design, directors, lifestyle, gossip, commentators, youtube	**(140)** *books*, *government*, *comedy*, *environment*, *baseball*, *soccer*, *hollywood*, *iphone*, *economics*, money	**(25)** *fashion*, *education*, *wine*, photography, radio, restaurants, science, SEO	**(17)** *music*, *tech*, *business*, *politics*, *food*, *sports*, *celebs*, *health*, media, bloggers, travel, writers

A Small Number of Very Popular Groups

No. of seekers	Number of experts					
	< 100	100 – 500	500 – 1K	1K – 5K	5K – 10K	> 10K
< 1K	**(5416)** *geology*, *karate*, *malaria*, *neurology*, *tsunami*, psychiatry, radiolo… pediatrics… dermat…	**(132)** volleyball, philosophers, tarot, perfume, florists, copywriters, taxi, esperanto…				
1K – 5K	**(915)** … *istry*, … astroph… media, … renewal… breast-c…					
5K – 10K	**(166)** … robot, … gospel-… explora…					
10K – 50K	**(174)** … virus, … choreog… metal, … world-c…					
50K–100K	**(7)** b… celebs, … leaders, superstars	… GOP, fantasy-football, NCAA, wwe, sci-fi	… xbox, triathlon, Google	**(37)** hotels, museums, hockey, architecture, charities, weather, space		
> 100K	**(3)** headlines, brits	**(49)** pop-culture, gospel, BBC, reality-tv, bollywood	**(58)** *religion*, actresses, gadgets, graphic-design, directors, lifestyle, gossip, commentators, youtube	**(140)** *books*, *government*, *comedy*, *environment*, *baseball*, *soccer*, *hollywood*, *iphone*, *economics*, money	**(25)** *fashion*, *education*, *wine*, photography, radio, restaurants, science, SEO	**(17)** *music*, *tech*, *business*, *politics*, *food*, *sports*, *celebs*, *health*, media, bloggers, travel, writers

(37) hotels, museums, hockey, architecture, charities, weather, space		
(140) *books*, *government*, *comedy*, *environment*, *baseball*, *soccer*, *hollywood*, *iphone*, *economics*, money	**(25)** *fashion*, *education*, *wine*, photography, radio, restaurants, science, SEO	**(17)** *music*, *tech*, *business*, *politics*, *food*, *sports*, *celebs*, *health*, media, bloggers, travel, writers

Thousands of Specialized Niche Groups

No. of seekers	Number of experts					
	< 100	100 – 500	500 – 1K	1K – 5K	5K – 10K	> 10K
< 1K	(5416) *geology*, *karate*, *malaria*, *neurology*, *tsunami*, psy... radiology, pe... dermatology, de...	(132) volleyball, philosophers, tarot, ...				
1K – 5K	(915) *biology*, ... *istry*, *swi*... astrophysics, media, semico... renewable-energ... breast-cancer, ju...					
5K – 10K	(166) *malware*... robot, chicago... gospel-music, exploration, wal...					
10K – 50K	(174) ipod, ... virus, Liverp... choreographers, metal, backstre... world-cup,					
50K–100K	(7) bbc-radio, ... celebs, ch... leaders, supersta...	... NCAA, wwe, sci-fi		... ties, weather, space		
> 100K	(3) headlines, brits	(49) pop-culture, gospel, BBC, reality-tv, bollywood	(58) *religion*, actresses, gadgets, graphic-design, directors, lifestyle, gossip, commentators, youtube	(140) *books*, *government*, *comedy*, *environment*, *baseball*, *soccer*, *hollywood*, *iphone*, *economics*, money	(25) *fashion*, *education*, *wine*, photography, radio, restaurants, science, SEO	(17) *music*, *tech*, *business*, *politics*, *food*, *sports*, *celebs*, *health*, media, bloggers, travel, writers

(5416) *geology*, *karate*, *malaria*, *neurology*, *tsunami*, psychiatry, radiology, pediatrics, dermatology, dentistry	**(132)** volleyball, philosophers, tarot, perfume, florists, copywriters, taxi, esperanto
(915) *biology*, *chemistry*, *swimmers*, astrophysics, multi-media, semiconductor, renewable-energy, breast-cancer, judaism	**(428)** *painters*, *astrology*, *sociology*, *geography*, *forensics*, anthropology, genealogy, archaeology, gluten, diabetes, neuroscience

Breaking the Twitter stereotype

- Twitter stereotype
  - Popular news on few topics such as sports, entertainment, politics, technology
  - Celebrity gossip, current news, and chatter

- Breaking the stereotype
  - Majority of the population discuss few popular topics, but
  - Smaller groups interested in thousands of niche, specialized topics

Detecting topical groups

- We followed content-based approach to identify topical groups

- Could community detection algorithms be used to detect topical groups?
  - Applied BGLL / Louvain algorithm on the Twitter social network to identify communities
  - Louvain largely unable to detect topical groups, especially the smaller ones (on niche topics)

Why do groups/communities form in a social network?

- "Common Identity and Bond Theory"
  - Prentice et. al. "Asymmetries in Attachments to Groups and to Their Members: Distinguishing Between Common-Identity and Common-Bond Groups", Personality and Social Psychology Bulletin, 1994

- Identity based groups

- Bond based groups

Common Identity and Bond Theory

Identity Based Groups	Bond Based Groups
Low Reciprocity	High Reciprocity
Low Personal Interactions	High Personal Interactions
High Topicality of discussions	Low Topicality of discussions

Examples:
Fans at a football match,
Attendees at a conference

Examples:
Family, personal friends

Detecting topical groups

- Louvain largely unable to detect topical groups, especially the smaller ones (on niche topics)

- Communities detected by Louvain fare better on structural measures like cut-ratio, conductance

- Topical groups do not have good structural quality
  - Poor values for standard community quality metrics such as cut-ratio and conductance

Analysis of 50 topical groups

- Low reciprocity among members
- Few one-to-one interactions
- Most tweets posted by experts are related to topic
- → Topical groups are identity-based which are difficult to detect via community detection algorithms