Subgraphs and Community Structure of Networks (part 2)

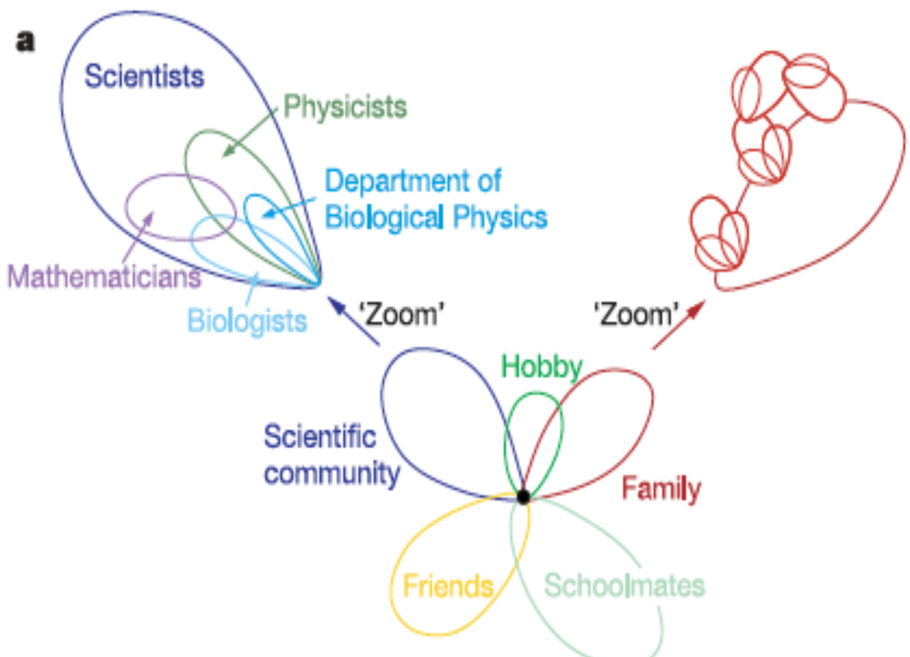
Mainack Mondal

CS 60017
Autumn 2021

OVERLAPPING COMMUNITY DETECTION

Overlapping communities

- Nodes in real networks are often parts of multiple overlapping communities



Two algorithms

- Clique Percolation Method
  - Uncovering the overlapping community structure of complex networks in nature and society, Palla et al., Nature Letters, vol. 435, 2005

- Link communities
  - Link communities reveal multiscale complexity in networks, Ahn et al., Nature Letters, vol. 466, 2010

Clique Percolation Method (CPM)

- Concept:
  - Internal edges of communities likely to be part of cliques
  - Inter-community edges unlikely to be part of cliques

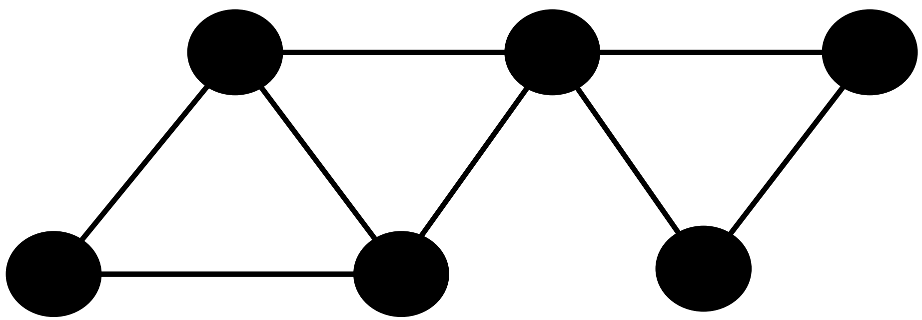
- Adjacent *k*-cliques: two *k*-cliques are adjacent if they share *k-1* nodes

Some material on CPM borrowed from slides by Eugene Lim

k-Clique Communities

- Adjacent k-cliques
  - Two k-cliques are adjacent when they share **k-1** nodes

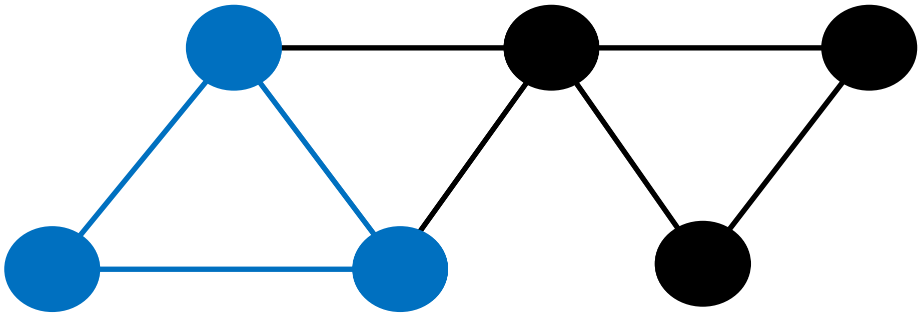
k = 3



k-Clique Communities

- Adjacent k-cliques
  - Two k-cliques are adjacent when they share **k-1** nodes

k = 3

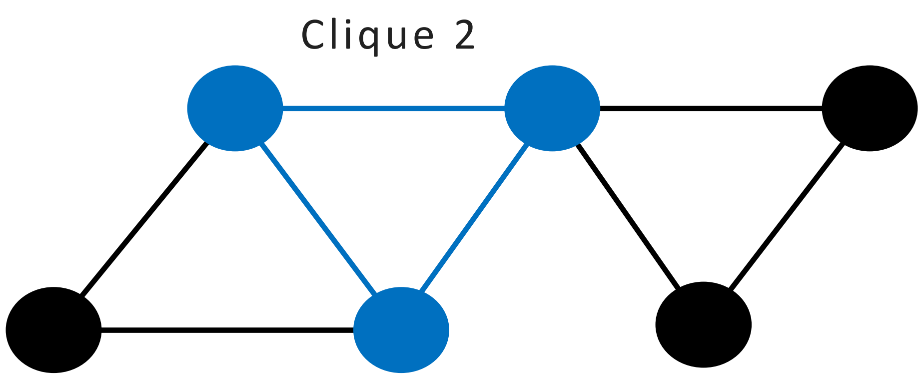


Clique 1

k-Clique Communities

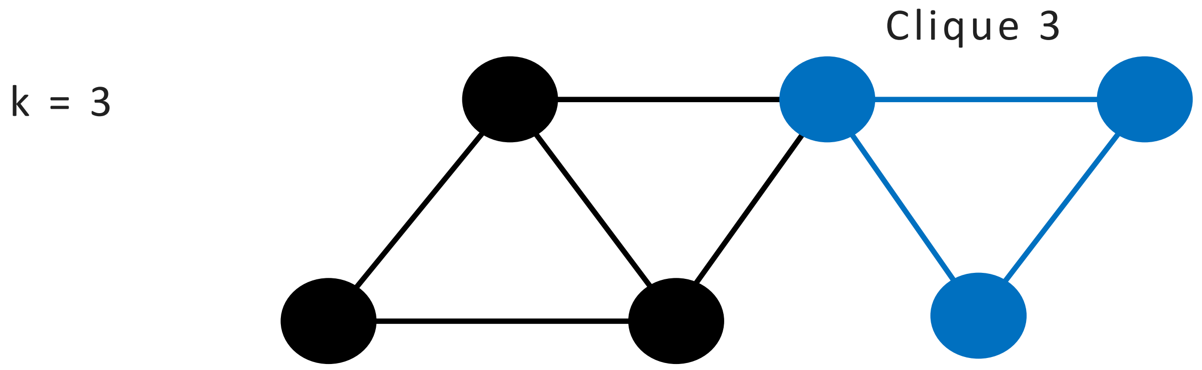
- Adjacent k-cliques
  - Two k-cliques are adjacent when they share **<u>k-1</u>** nodes

k = 3



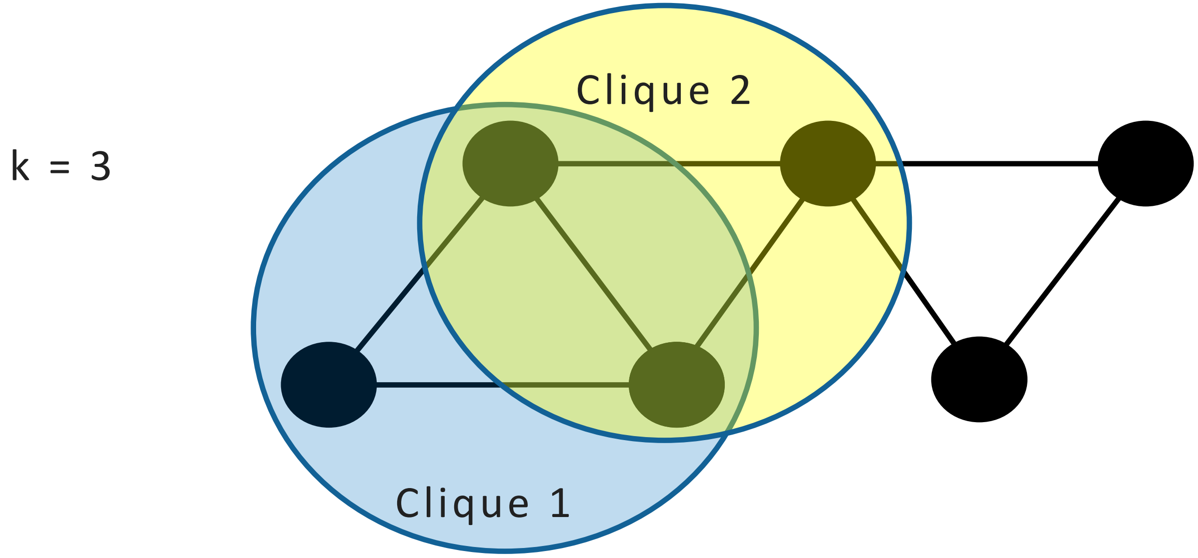
k-Clique Communities

- Adjacent k-cliques
  - Two k-cliques are adjacent when they share **k-1** nodes



k-Clique Communities

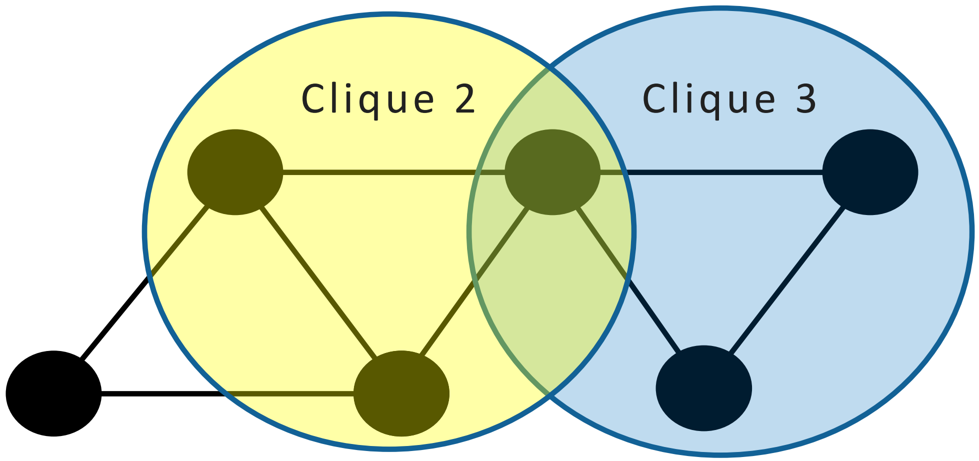
- Adjacent k-cliques
  - Two k-cliques are adjacent when they share **k-1** nodes



k-Clique Communities

- Adjacent k-cliques
  - Two k-cliques are adjacent when they share **k-1** nodes

k = 3



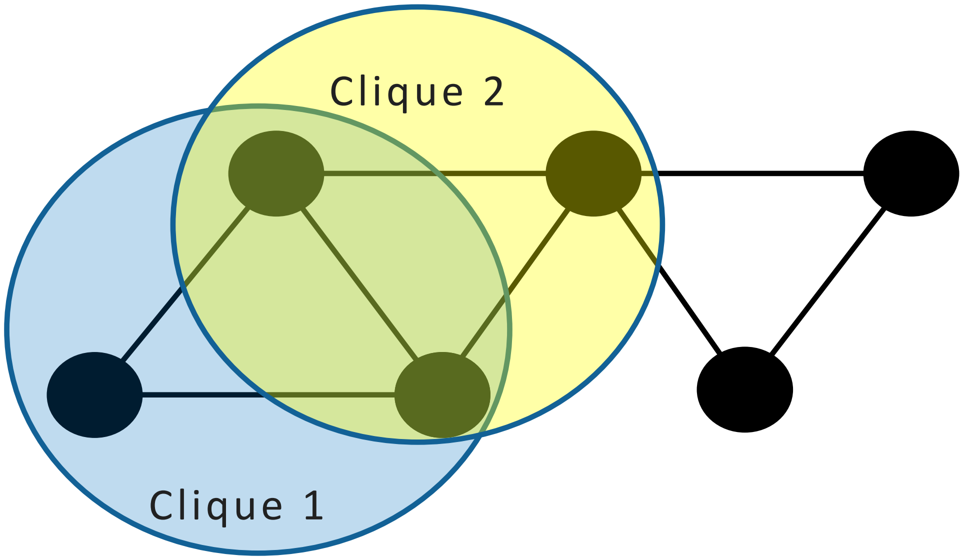
k-Clique Communities

- k-clique community
  - Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques

k-Clique Communities

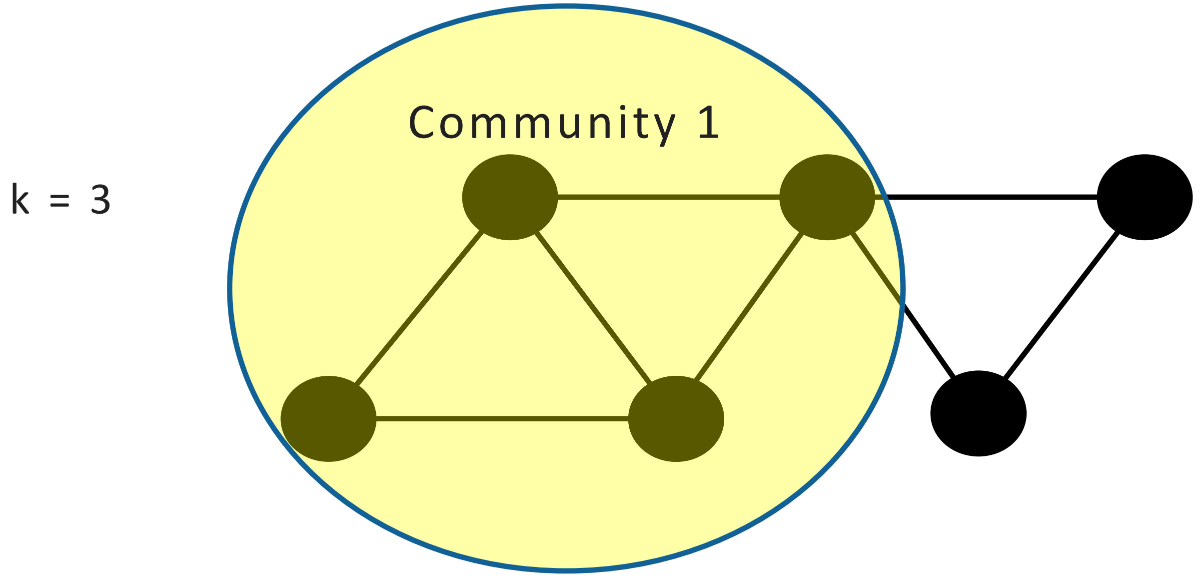
- k-clique community
  - Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques

k = 3



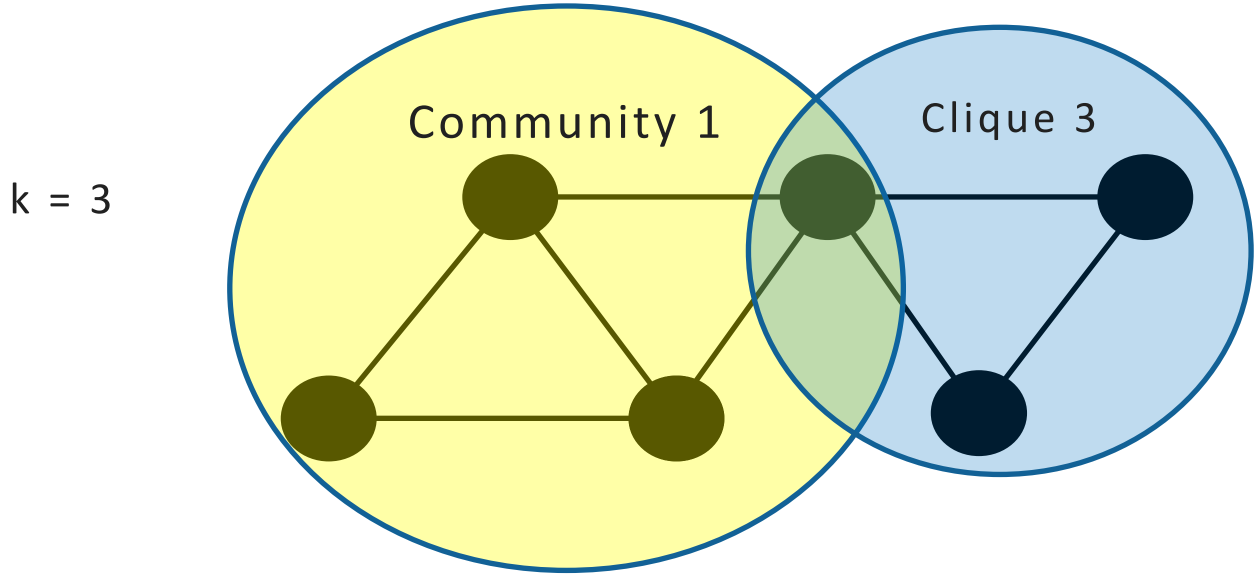
k-Clique Communities

- k-clique community
  - Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



k-Clique Communities

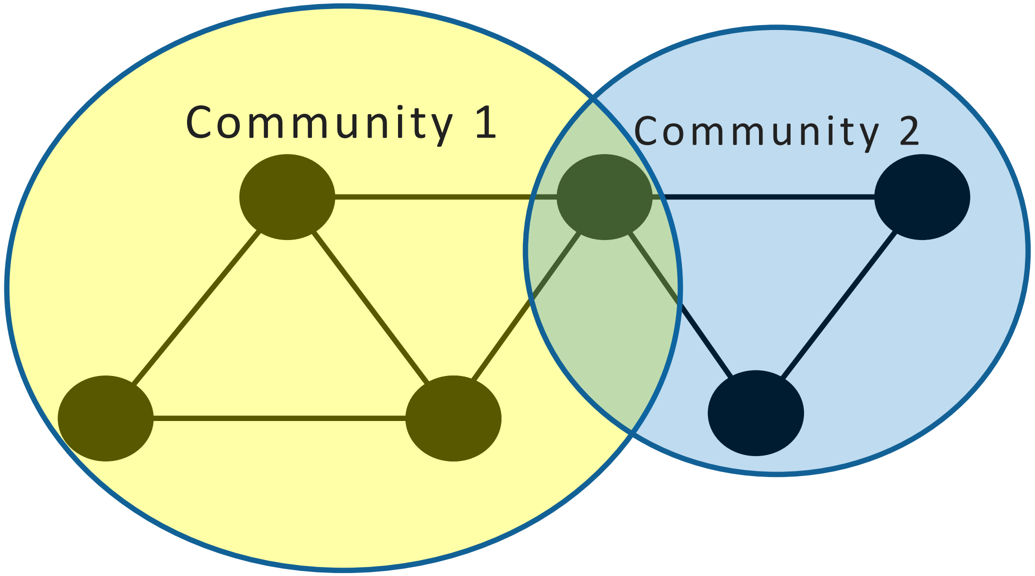
- k-clique community
  - Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques



k-Clique Communities

- k-clique community
  - Union of all k-cliques that can be reached from each other through a series of adjacent k-cliques

k = 3



Algorithm

- Locate maximal cliques
- Convert from cliques to k-clique communities

Locate Maximal Cliques

- Largest possible clique size can be determined from degrees of vertices
- Starting from this size, find all cliques, then reduce size by 1 and repeat

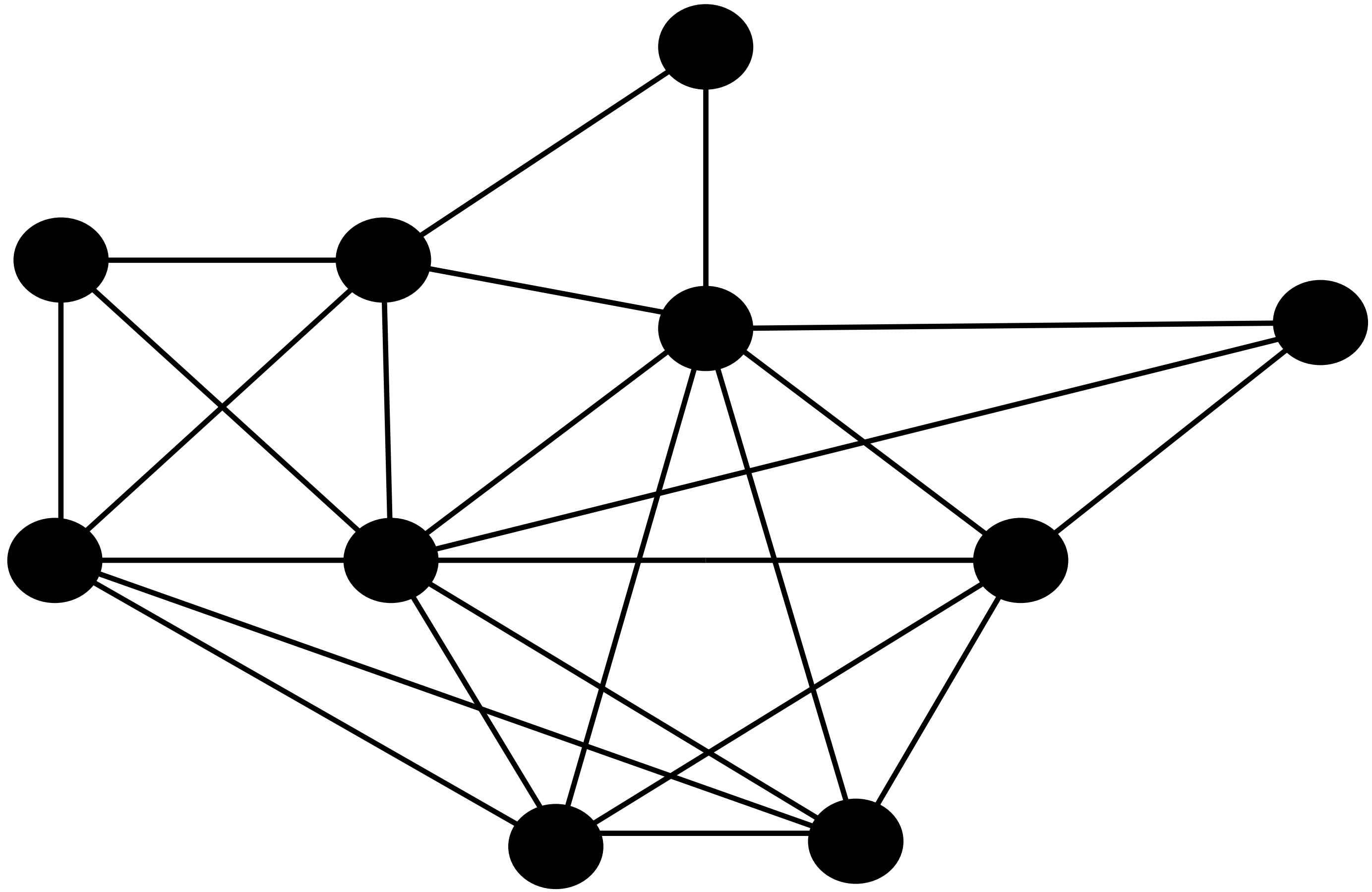
Finding all cliques: brute-force

- Set A initially contains vertex v, Set B contains neighbors of v
- Transfer one vertex w from B to A
- Remove vertices that are not neighbors of w from B
- Repeat until A reaches desired size
- If fail, step back and try other possibilities

Algorithm

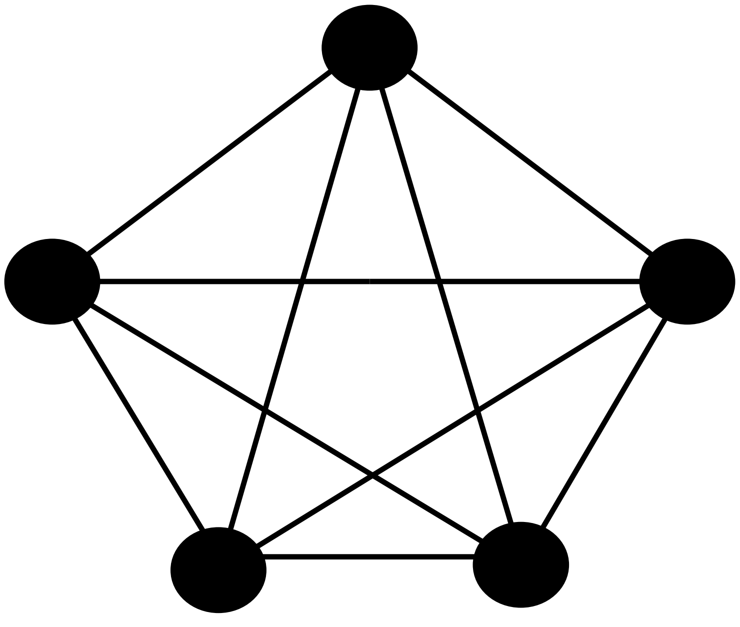
- Locate maximal cliques
- Convert from cliques to k-clique communities

Cliques to k-Clique Communities

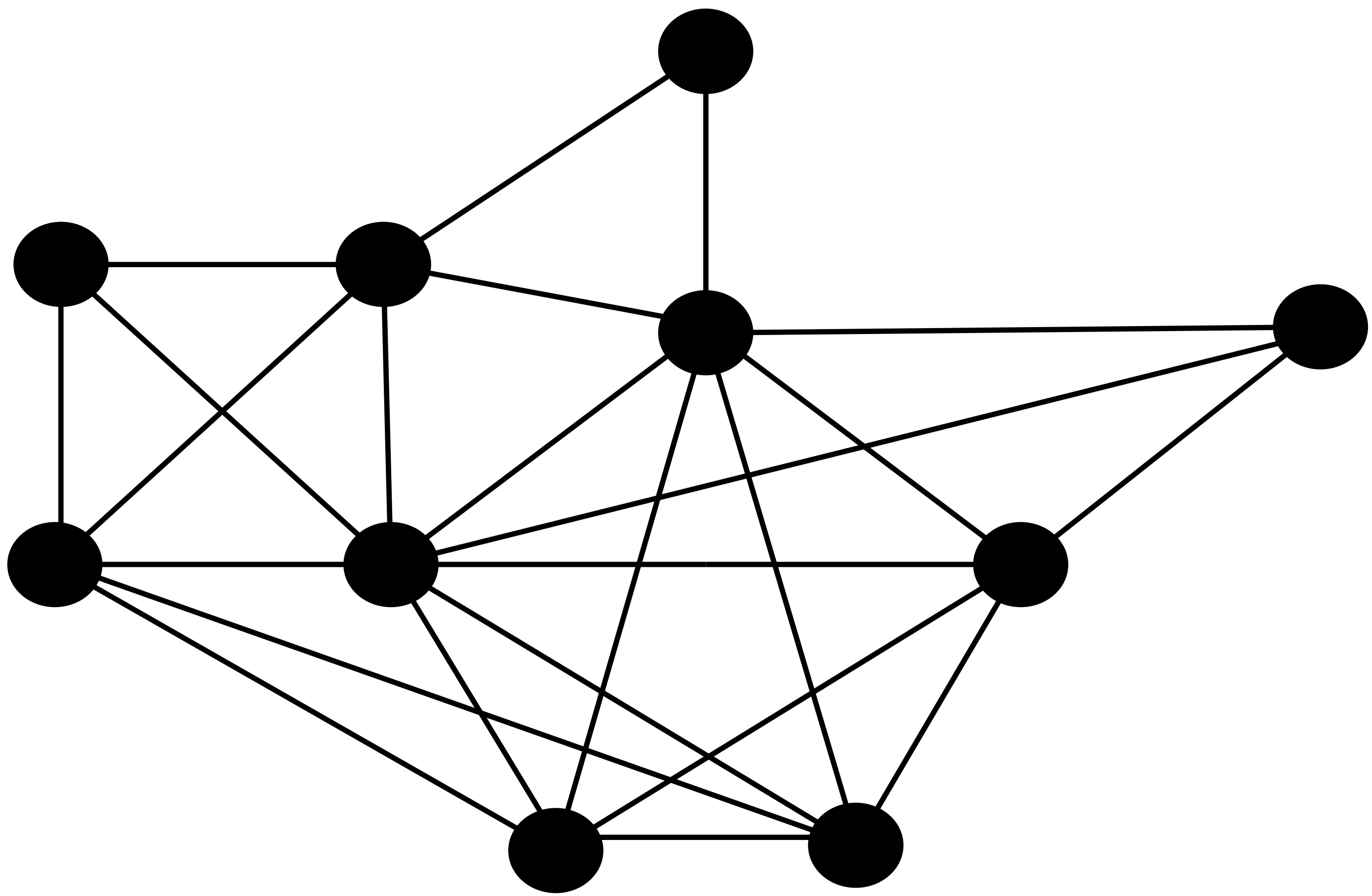


Cliques to k-Clique Communities

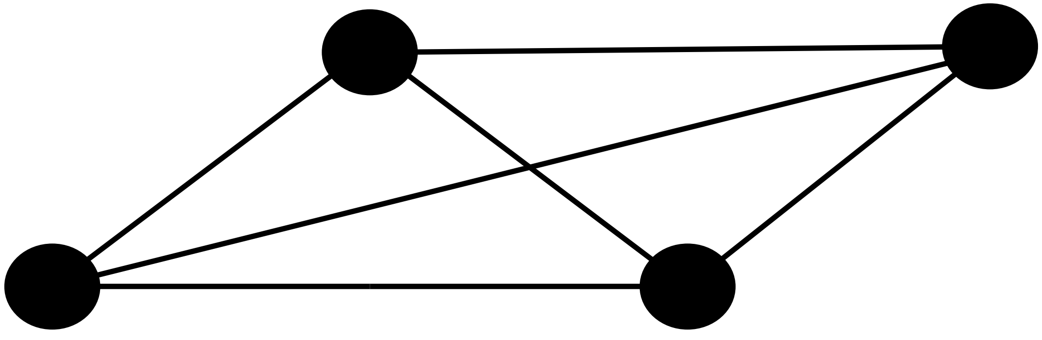
Clique 1: 5-clique



Cliques to k-Clique Communities

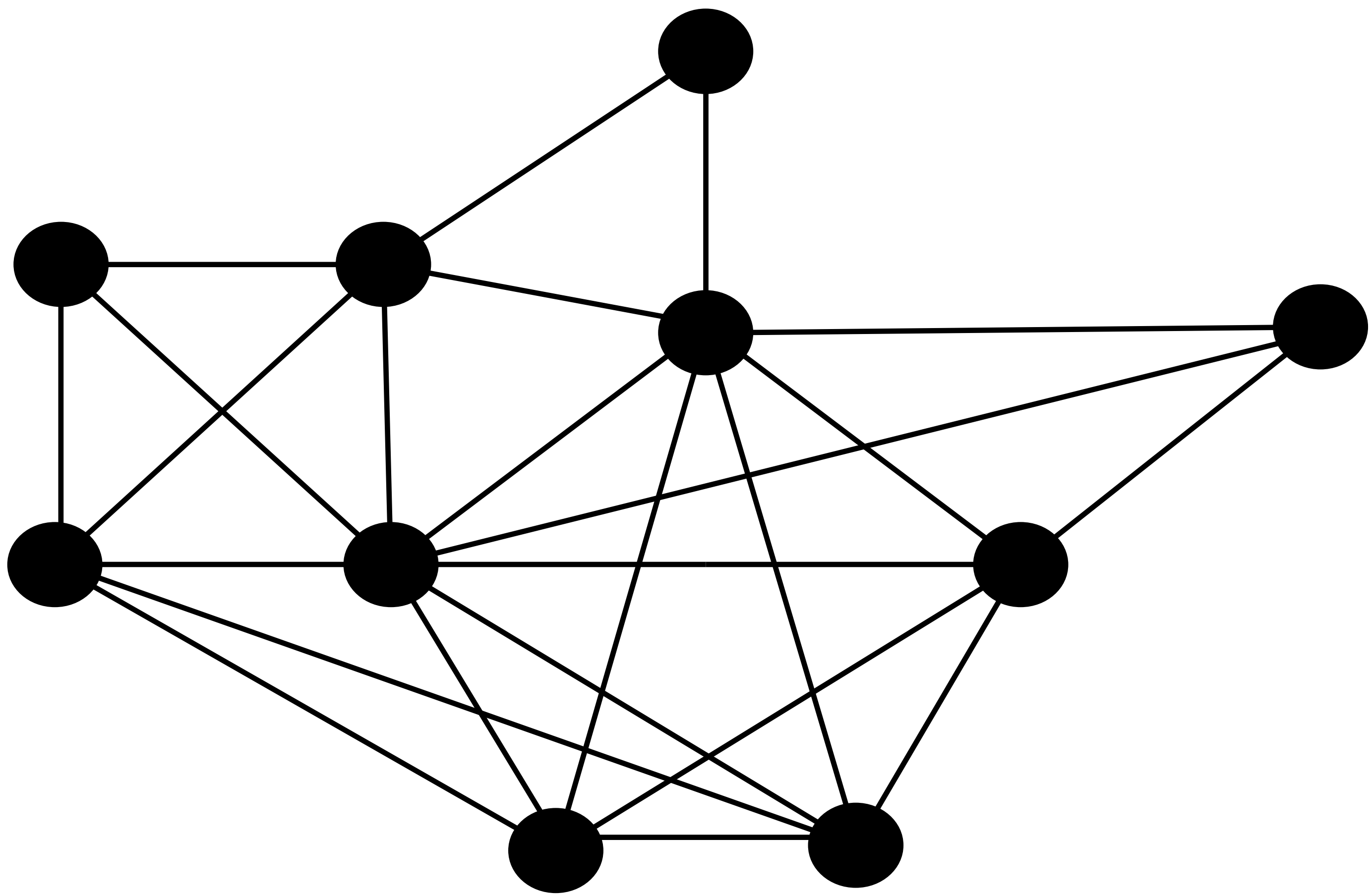


Cliques to k-Clique Communities



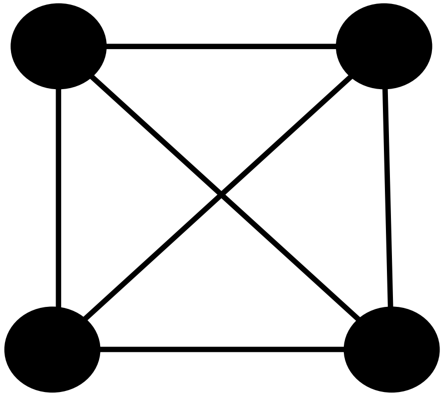
Clique 2: 4-clique

Cliques to k-Clique Communities

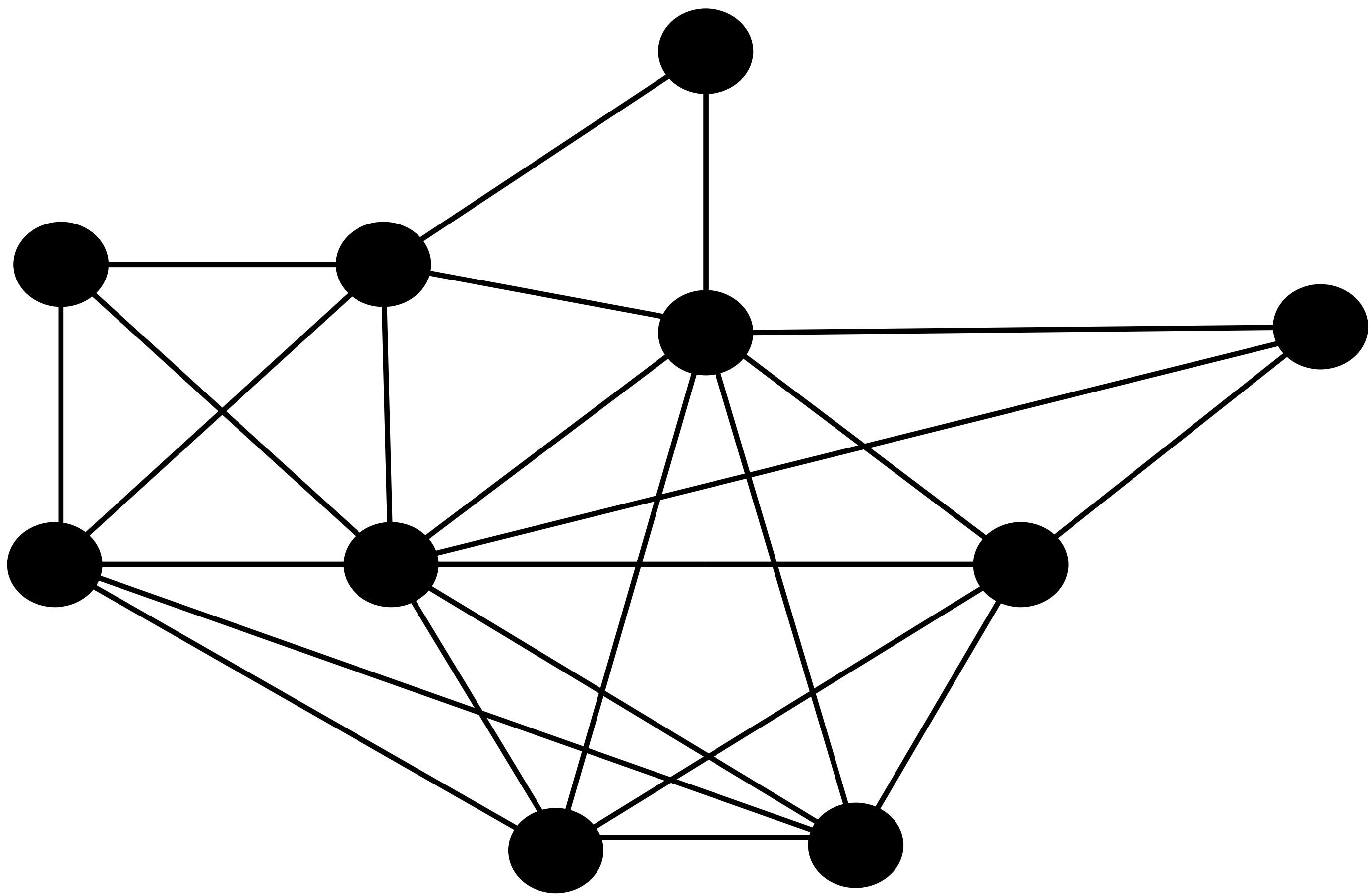


Cliques to k-Clique Communities

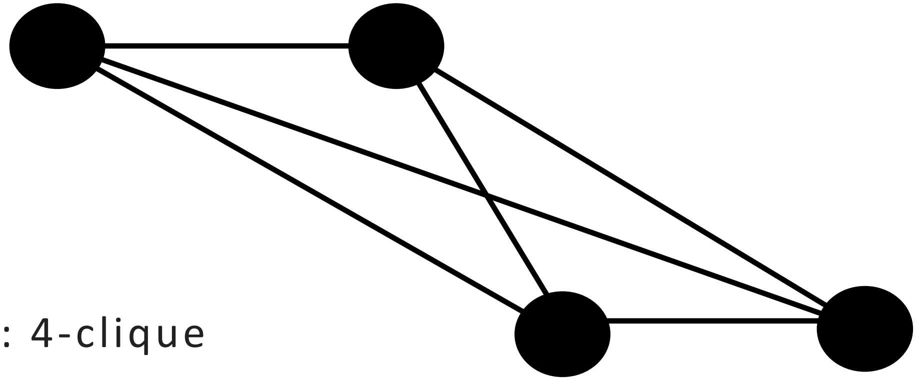
Clique 3: 4-clique



Cliques to k-Clique Communities

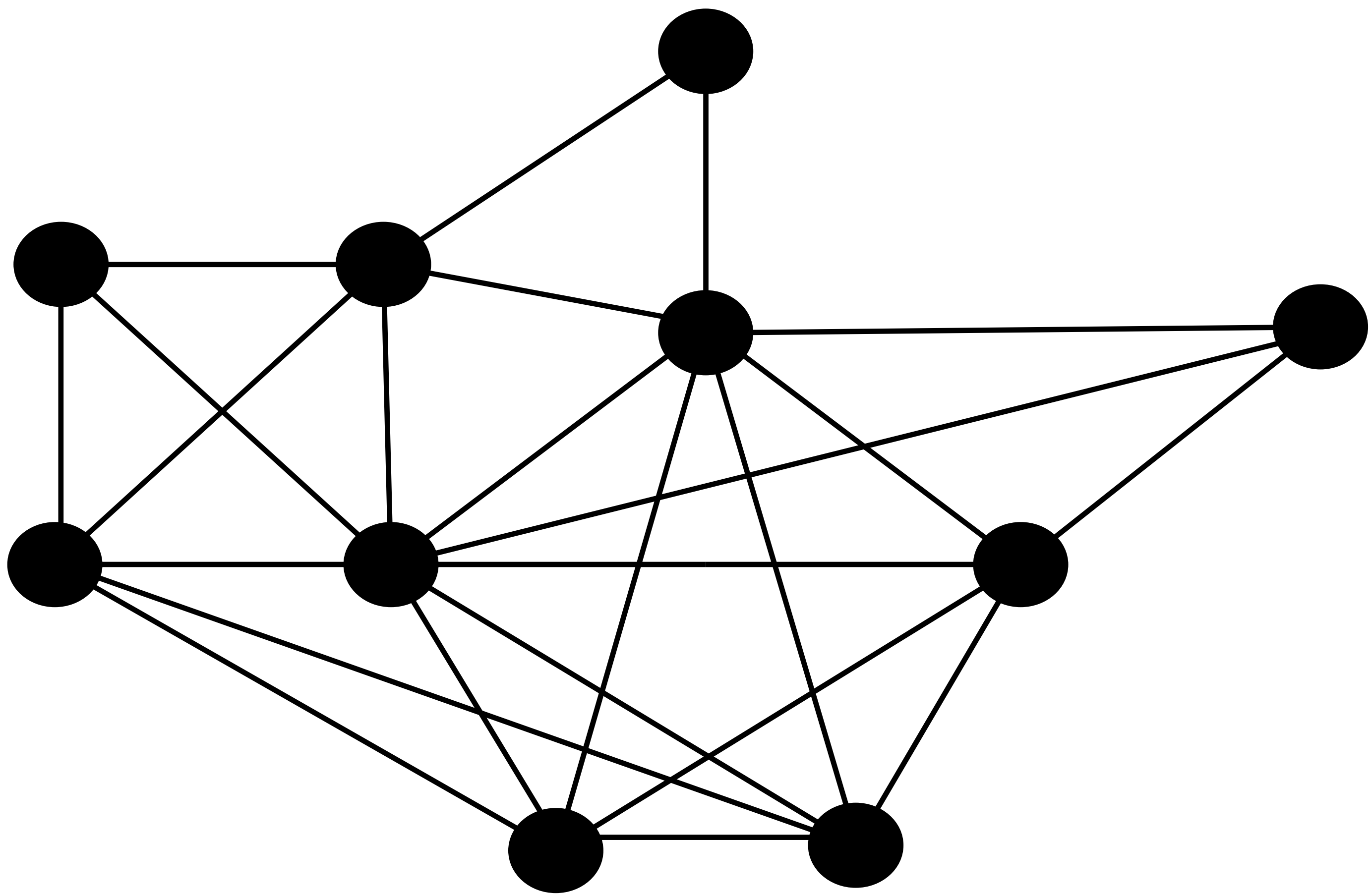


Cliques to k-Clique Communities

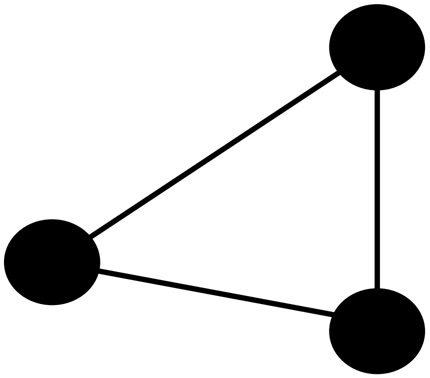


Clique 4: 4-clique

Cliques to k-Clique Communities

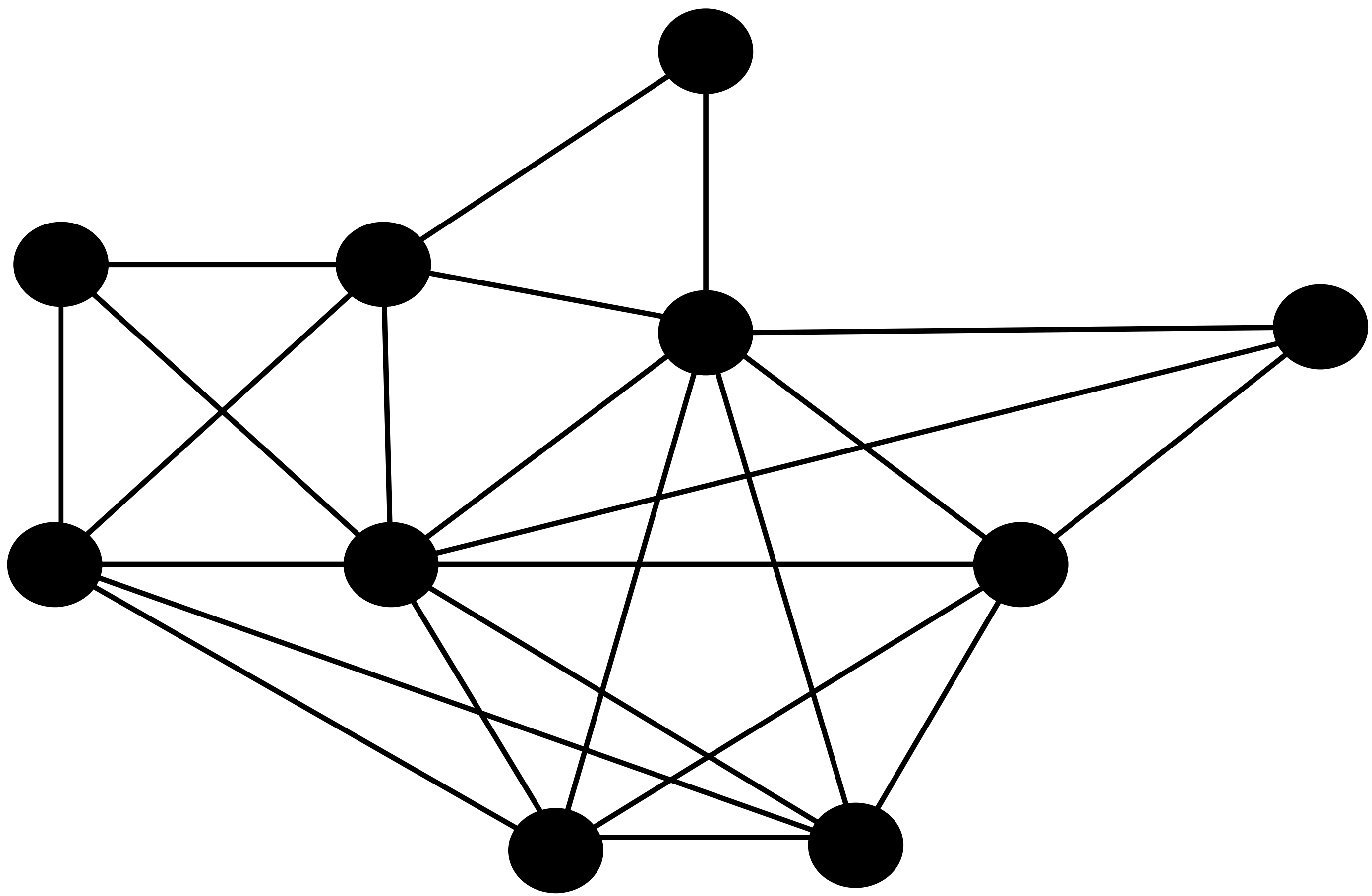


Cliques to k-Clique Communities



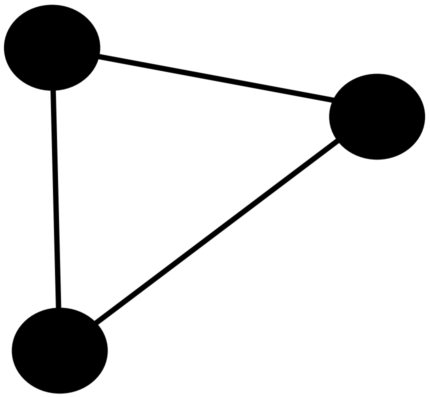
Clique 5: 3-clique

Cliques to k-Clique Communities



Cliques to k-Clique Communities

Clique 6: 3-clique



Cliques to k-Clique Communities

Clique-Clique overlap matrix

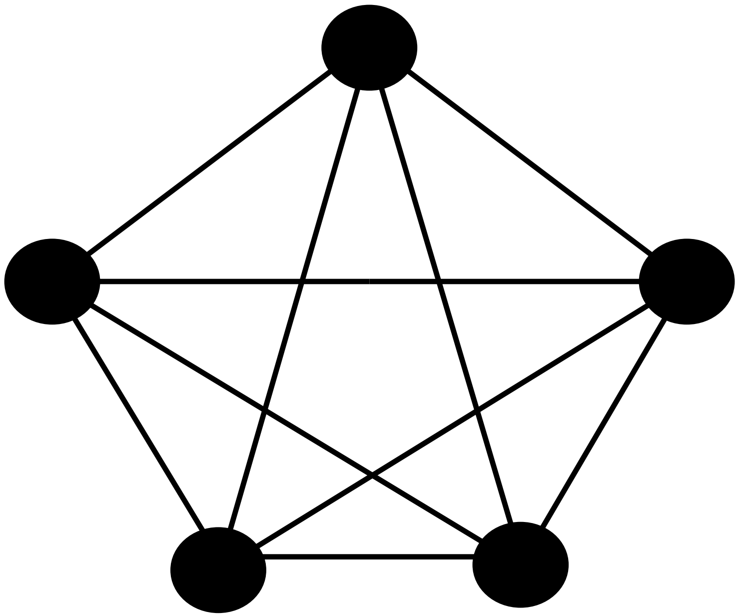
	1	2	3	4	5	6
1	5					
2		4				
3			4			
4				4		
5					3	
6						3

Cliques to k-Clique Communities

Clique-Clique overlap matrix

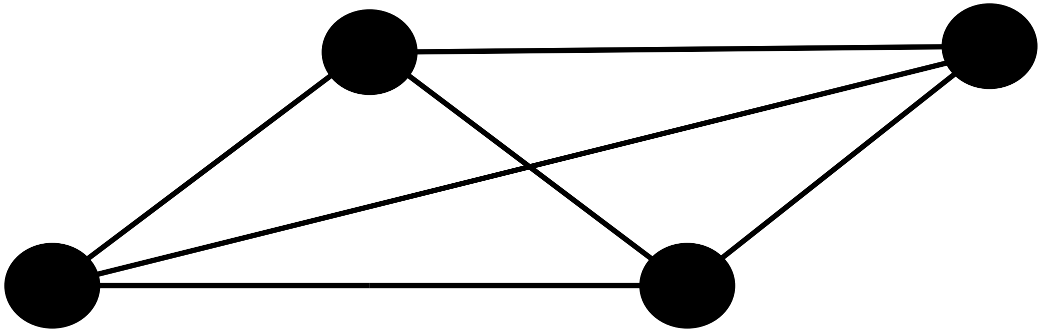
	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	3	2
6	2	2	2	1	2	3

Cliques to k-Clique Communities



Clique 1: 5-clique

Cliques to k-Clique Communities



Clique 2: 4-clique

Cliques to k-Clique Communities

Clique-Clique overlap matrix

	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	3	2
6	2	2	2	1	2	3

Intuition of the algorithm

- First find all cliques of size k in the graph
- Then create graph where nodes are cliques of size k
- Add edges if two nodes (cliques) share k-1 common nodes
- Each connected component is a community

Cliques to k-Clique Communities

- For a given value of k, k-clique communities:
  - Connected clique components in which neighboring cliques linked to each other by at least k-1 common nodes

- How to find k-clique communities from the clique-clique overlap matrix?
  - Erase every diagonal element smaller than k
  - Erase every off-diagonal element smaller than k-1
  - Replace remaining elements by 1
  - Carry out a component analysis of this matrix

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	3	2
6	2	2	2	1	2	3

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	3	2
6	2	2	2	1	2	3

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	0	2
6	2	2	2	1	2	0

Delete/ replace by 0 if less than k

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	0	2
6	2	2	2	1	2	0

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	1	3	1	2
2	3	4	1	1	1	2
3	1	1	4	2	1	2
4	3	1	2	4	0	1
5	1	1	1	0	0	2
6	2	2	2	1	2	0

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	0	3	0	0
2	3	4	0	0	0	0
3	0	0	4	0	0	0
4	3	0	0	4	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Delete/ replace with 0 if less than k-1

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	5	3	0	3	0	0
2	3	4	0	0	0	0
3	0	0	4	0	0	0
4	3	0	0	4	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	1	1	0	1	0	0
2	1	1	0	0	0	0
3	0	0	1	0	0	0
4	1	0	0	1	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Change all non-zeros to 1

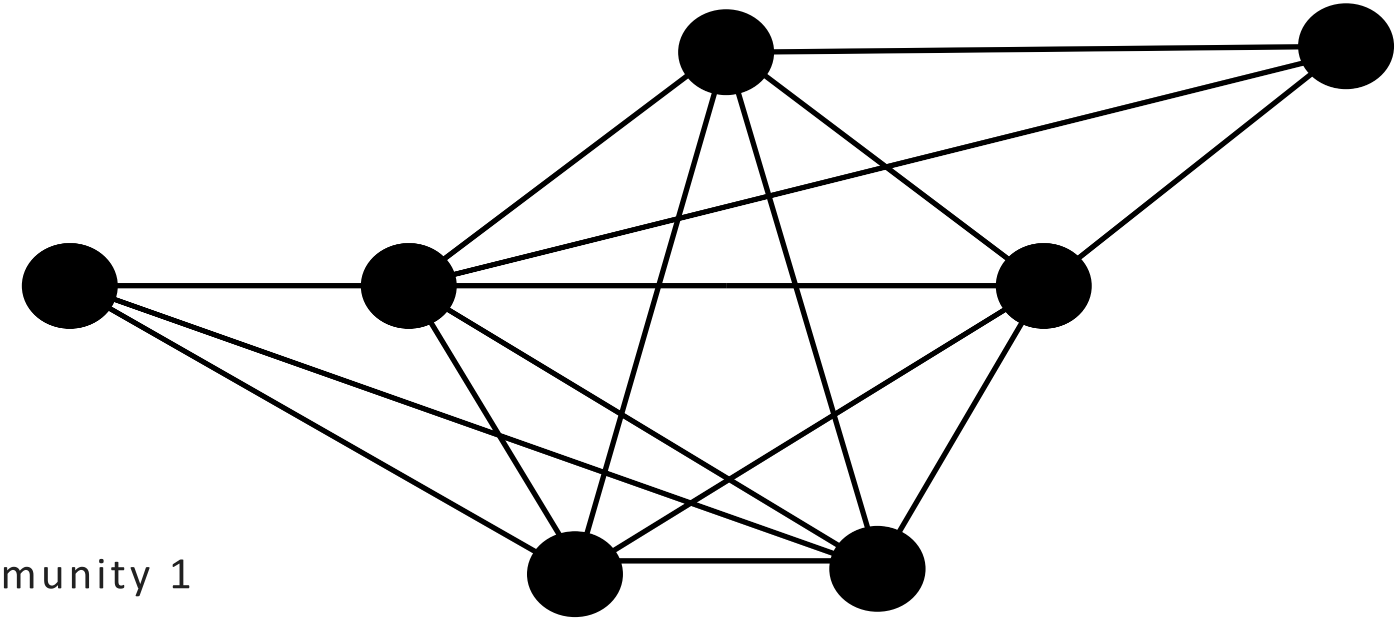
Cliques to k-Clique Communities

k=4

	1	2	3	4	5	6
1	1	1	0	1	0	0
2	1	1	0	0	0	0
3	0	0	1	0	0	0
4	1	0	0	1	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Cliques to k-Clique Communities

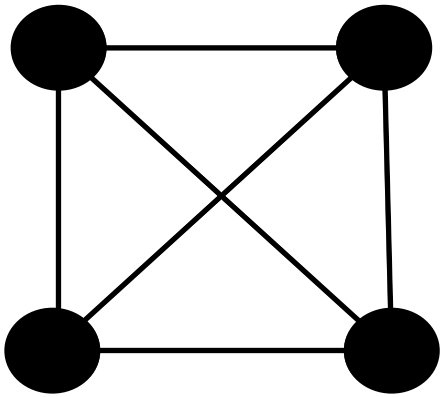
k=4



Community 1

Cliques to k-Clique Communities

k=4



Community 2

Clique Percolation Method: Analysis

- Believed to be non-polynomial
- No closed formula can be given
- However, claimed to be efficient on real systems

- Limitations
    - Fail to give meaningful covers for graph with few cliques
    - With too many cliques, might give a trivial community structure

Link communities

- A node might belong to multiple communities
  - For a person: family, co-workers, friends, …
- A link often exists for one dominant reason
  - Two people are in the same family, or are co-workers
- Link community: a set of closely inter-related links

Identifying Link communities

- Hierarchical clustering with a similarity between links to build a dendrogram
  - Each leaf of the dendrogram is a link from the original network
  - Branches of the dendrogram are link communities

- Slice the dendrogram at a suitable level
- Each link placed in a single community
- Each node inherits membership of the communities of all its links

For hierarchical clustering

- Two questions to be answered

- How to measure similarity between items (e.g., links)?

- At which level to slice the dendrogram?

Similarity measure between links

- Node i and its neighboring nodes: $n_+(i)$

- Similarity measured only between pairs of links which share a node
- Similarity between e_{ik} and e_{jk}:

$$S(e_{ik}, e_{jk}) = |n_+(i) \cap n_+(j)| / |n_+(i) \cup n_+(j)|$$

Which level to slice the dendrogram?

- Measure: Partition density D
  - Total number of links in network: M
  - $\{ P_1, P_2, \ldots, P_C \}$: partition of links into C subsets
  - P_c has n_c nodes and m_c links

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)}$$

  - Partition density is average of D_c weighted by the fraction of links present in P_c

$$D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)}$$

Going from non-overlapping to overlapping algorithms

- Simple “partition + growth” approach
  - Partition: First detect partition of the network using a *good* community detection algorithm
  - Growth: Next consider nodes in each community as seed set and add nodes which are highly connected to seed

Going from non-overlapping to overlapping algorithms

- Simple "partition + growth" approach
  - Partition: First detect partition of the network using a *good* community detection algorithm
  - Growth: Next consider nodes in each community as seed set, add nodes which are highly connected to seed

Going from non-overlapping to overlapping algorithms

- Simple “partition + growth” approach
  - Partition: First detect partition of the network using a *good* community detection algorithm
  - Growth: Next consider nodes in each community as seed set, add nodes which are highly connected to seed

- You are who you know: Inferring user profiles in online social networks, by Mislove et al (http://www.ccs.neu.edu/home/amislove/publications/Inferring-WSDM.pdf)

Definition: Conductance

- How strong is a particular community A?

- Conductance previously proposed

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

- But, biased towards large communities

Definition: Normalized conductance

- Metric: Normalized conductance C

$$C = \frac{e_{AA}}{e_{AA} + e_{AB}} - \frac{e_A e_A}{e_A e_A + e_A e_B}$$

- Fraction of A's links within A Relative to a random graph
- Range is [-1,1]
- 0 represents no stronger than random

Growth algorithm

- Given seed users, find a community by
  - Adding users
  - Stopping at some point
- At each step, add user who increases normalized conductance by the most
- Stop when no user increases normalized conductance

Partition + Growth algorithm in action

- Finding friendlists from 1-hop subgraph in Facebook
- Used Louvain's modularity-based algorithm to find partitions
- Then grow each community by normalized conductance based growth algorithm
- Provide final overlapping communities to users in an App—Friendlist Manager

- Simplifying Friendlist Management , by Liu et al, WWW Demo 2012 (https://cse.iitkgp.ac.in/~mainack/publications/Friendlist-WWW-Demo.pdf)

Partition + Growth algorithm in action

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?
- Several existing measures
  - Purity
  - Rand index
  - Normalized Mutual Information (NMI) [has been extended to overlapping communities]
- Generalized Measures for the Evaluation of Community Detection Methods, by Labatut (https://arxiv.org/abs/1303.5441)

DIFFERENT TYPES OF GROUPS IN A SOCIAL NETWORK

Different methods to identify groups

- Identifying groups based on network structure – community detection algorithms
- How about identifying groups based on content, e.g., text or profile attributes?
- Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale, Bhattacharya et al., CSCW 2014

Identified topical groups in Twitter

Topical Groups = Experts + Seekers

Experts: Users who have expertise on the topic

Seekers: Users who are interested in the 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, copy-writers, taxi, esperanto				
1K – 5K	**(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	**(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, radiolog... pediatrics, dermat...	(132) volleyball, philosophers, tarot, perfume, florists, copy-writers, 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	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[illegible] radiology, pe[illegible] dermatology, de[illegible]	**(132)** volleyball, philosophers, tarot, [illegible]				
1K – 5K	**(915)** *biology*, [illegible] *istry*, *sw*[illegible] astrophysics, [illegible] media, semico[illegible] renewable-energ[illegible] breast-cancer, ju[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
5K – 10K	**(166)** *malware*[illegible] robot, chicago[illegible] gospel-music, [illegible] exploration, wal[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
10K – 50K	**(174)** ipod, [illegible] virus, Liverp[illegible] choreographers, [illegible] metal, backstre[illegible] world-cup,	[illegible]	[illegible]	[illegible]	[illegible]	
50K–100K	**(7)** bbc-radio, [illegible] celebs, ch[illegible] leaders, supersta[illegible]	[illegible] NCAA, wwe, sci-fi	[illegible]	[illegible] 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, multimedia, 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

Why do groups form?

- “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

Analysis of 50 topical groups

- Low reciprocity among members
- Few one-to-one interactions
- Most tweets posted by experts are related to topic