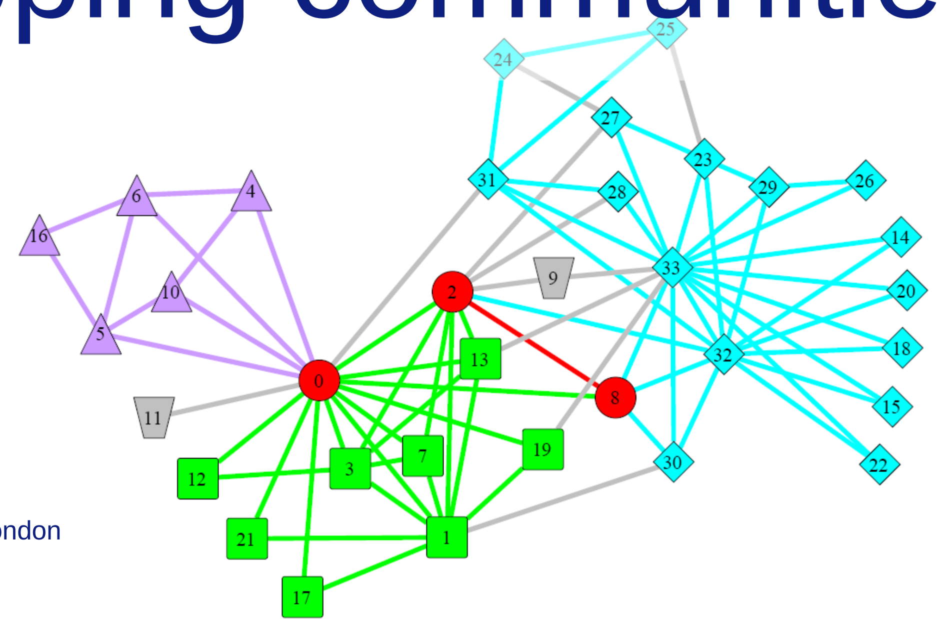
Imperial College London

Making Sense of Data - Theory and Practice
NCAF Summer meeting
Guildford, 12th July 2010

Tim Evans
Theoretical Physics

Give cliques a chance: line graphs, clique graphs and overlapping communities



- Graphs, Data and our Vertex Centric Viewpoint
- Clique Graphs
- Overlapping Communities
- Line Graphs and Overlapping Communities
- Conclusions

Network/Graph Notation

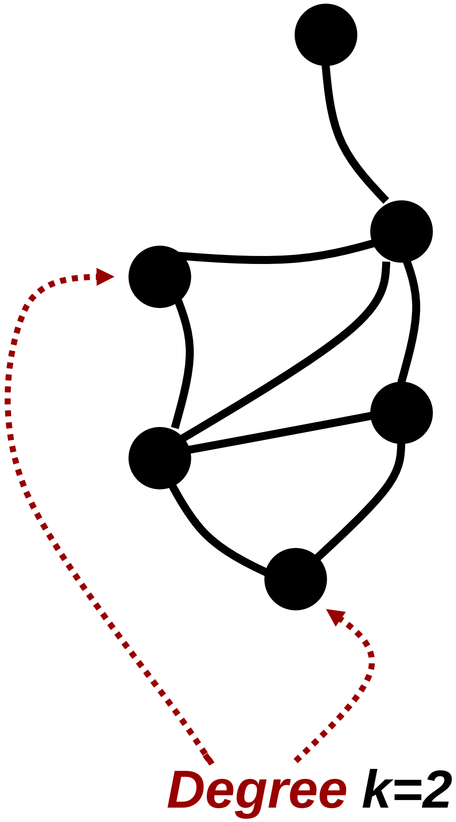
I will focus on **Simple Graphs**

(no values or directions on edges, no values for vertices)

- ***N*** = *number of vertices in graph*
- ***E*** = *number of edges in graph*
- ***k*** = *degree of a vertex*
- ***<k>*** = *average degree* = ***2E/N***
- Degree Distribution

  n(k) = *number of vertices with degree* ***k***

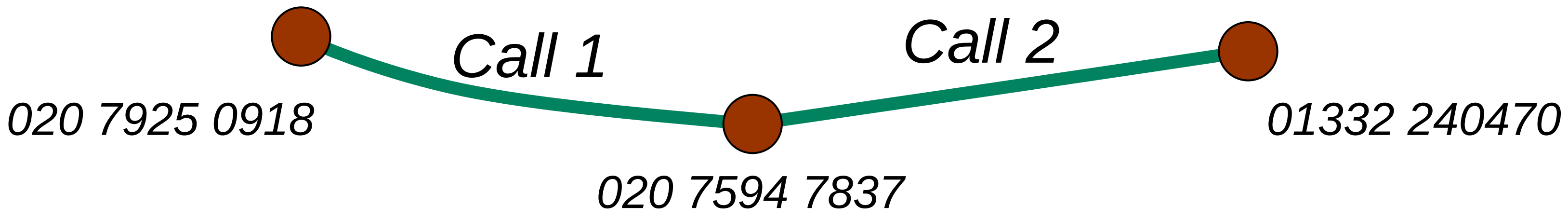
  p(k) = ***n(k)/N*** = *normalised distribution*



Graphs and "Making Sense of Data"

- Many data sets can be represented by a graph
  e.g. vertex = distinct value in a database field
  edge = two values in same record

Originator	Receiver	Time	Duration
020 7594 7837	020 7925 0918	2010.07.09:11.10	360
020 7594 7837	01332 240470	2010.07.09:11.17	415



Graphs as Data Reduction

Disadvantages

- Loss of Information
  - User Account Information, …
- No unique graph representation
  - Edge directions, edge weights, …

Advantages

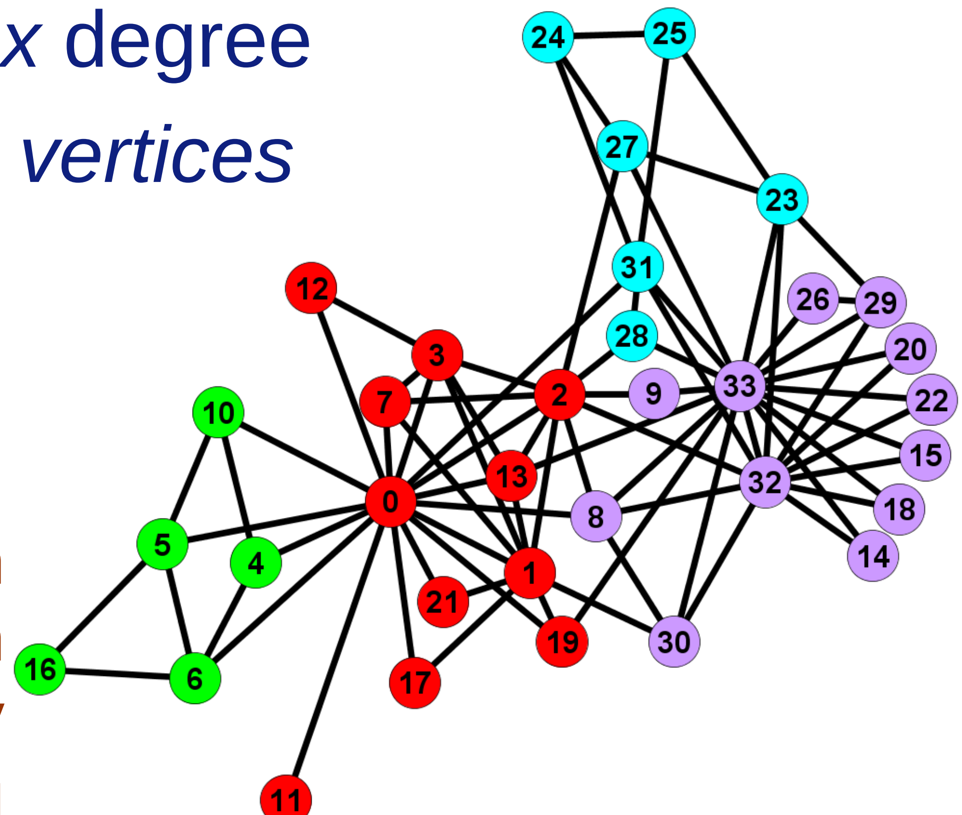
- Graphs are simpler
- Universal representations
  - Comparison to other networks, standard libraries, …

$\Rightarrow$ Making Sense of Data

Vertex Centric Viewpoint

The focus in the literature is often on the vertices:-

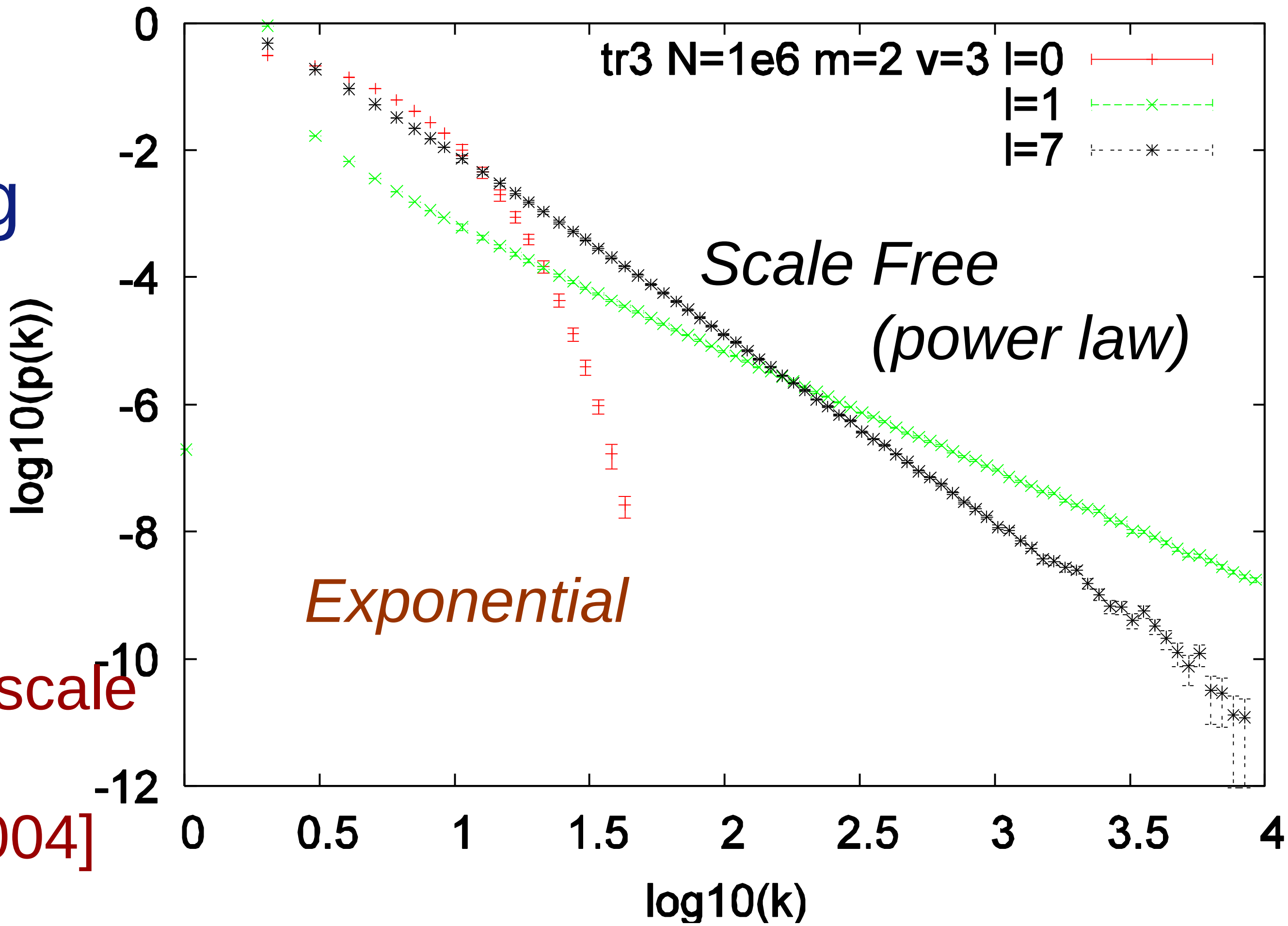
- Distributions of *vertex* degree
- Cluster coefficient of *vertices*
- *Vertex* partitions as communities
- …..



Vertex partition of Karate club graph with optimal modularity [Agarwal & Kempe 2007]

Vertex Centric – Degree Distributions ***p(k)***

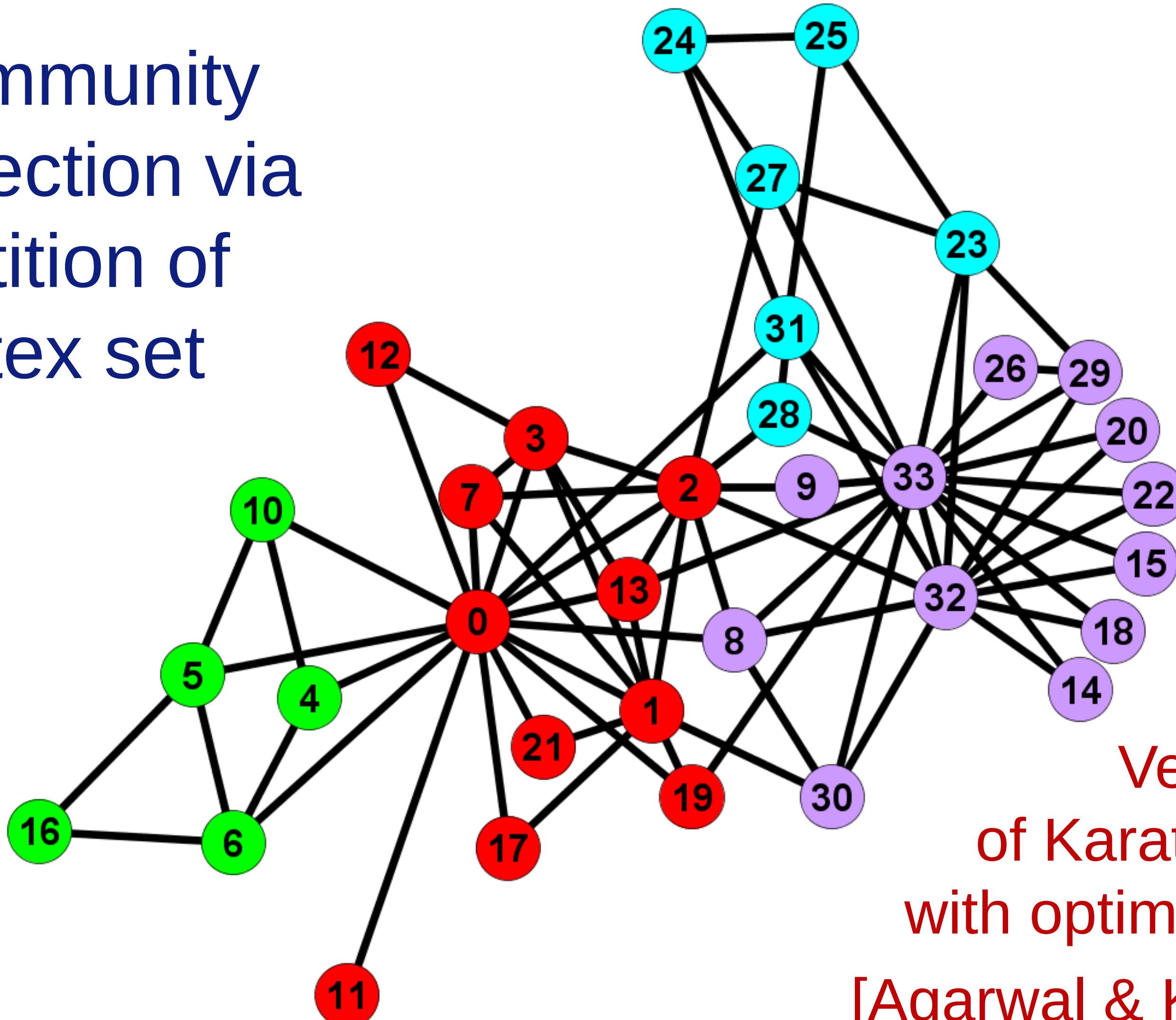
p(k) **=** probability of vertex having degree ***k***



Random walk on vertices produces scale free graphs
[TSE, Saramäki 2004]

Vertex Centric – Vertex Partitions

Community detection via partition of vertex set



Vertex partition of Karate club graph with optimal modularity [Agarwal & Kempe 2007]

Word Count of Network Review [Evans '04]

Stem	Rank	Count	Stem	Rank	Count
network	1	254	number	11	58
vertic	2	107	distanc	12	48
edg	3	86	model	13	47
random	3	86	connect	14	46
graph	5	81	data	15	40
degre	6	78	link	16	38
power	7	68	world	16	38
lattic	8	67	hub	33	25
law	9	65	point	38	23
vertex	10	61	site	40	22

Stop words removed then stemming

Word Count - Approximate Ratios

Network Words	Vertex Words	Edge Words
6	3	2

Word count shows focus on vertices

How can we compensate for vertex bias?

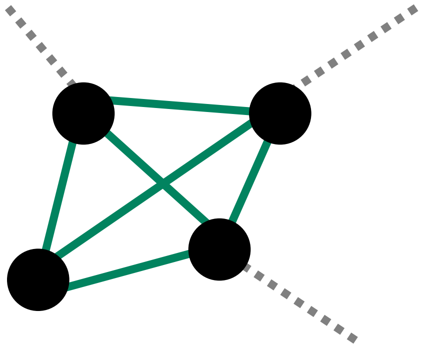
One Answer:-

1. Represent structures in the original graph ***A*** (edges, cliques, motifs, …) as vertices in a new graph
2. Analyse new graph as usual

$\Rightarrow$ Vertex bias on new graph = emphasis on other structures in original graph

- Graphs, Data and our Vertex Centric Viewpoint
- Clique Graphs
- Overlapping Communities
- Line Graphs and Overlapping Communities
- Conclusions

Cliques – complete subgraphs



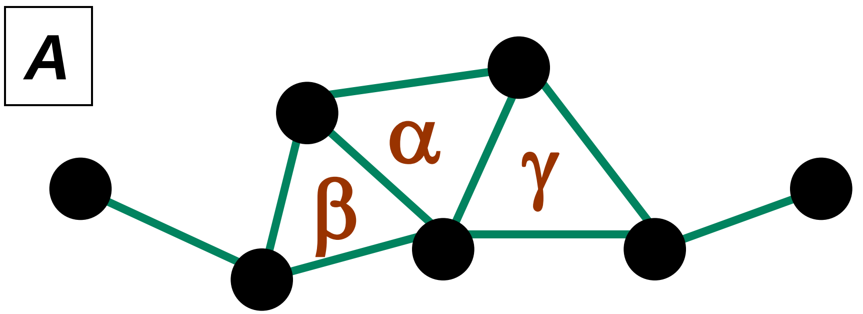
- Name originates from representation of cliques of people in social science
  [Luce and Perry ‘49]
- Theory of Triadic Closure
  [Granovetter, ‘73]
- Maximal Clique and Clique Cover problems
  [Bron-Kerbosch algorithm, ‘73]
- Clustering/Community detection
  [Freeman ’92, ‘96; Palla et al, ‘05; Yan & Gregory ‘10]
- Structure Analysis
  [Samudrala & Moult, ‘98; Takemote et al. ‘07]

Give Cliques a Chance – a clique centric viewpoint

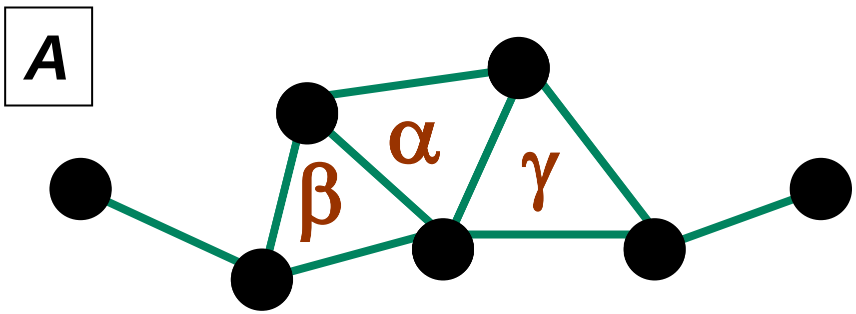
Shift our viewpoint from vertices to cliques using a **Clique Graph**

- New Vertices = original cliques
- New Edges = overlap of cliques

Basic idea also used in social science [Everett and Borgatti, ‘98]



Example



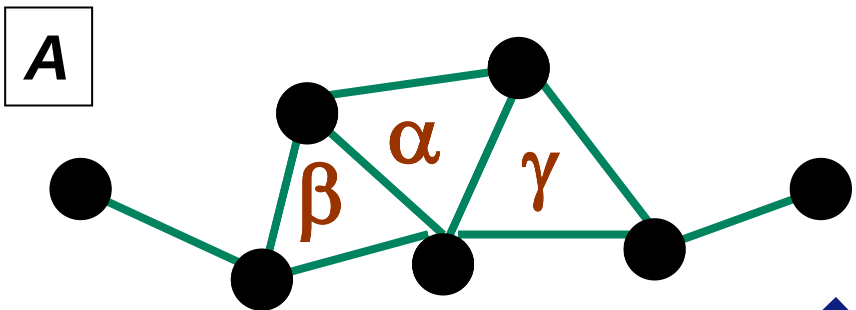
- three cliques of order 3 = triangles, triads
  - labelled α,β,γ.

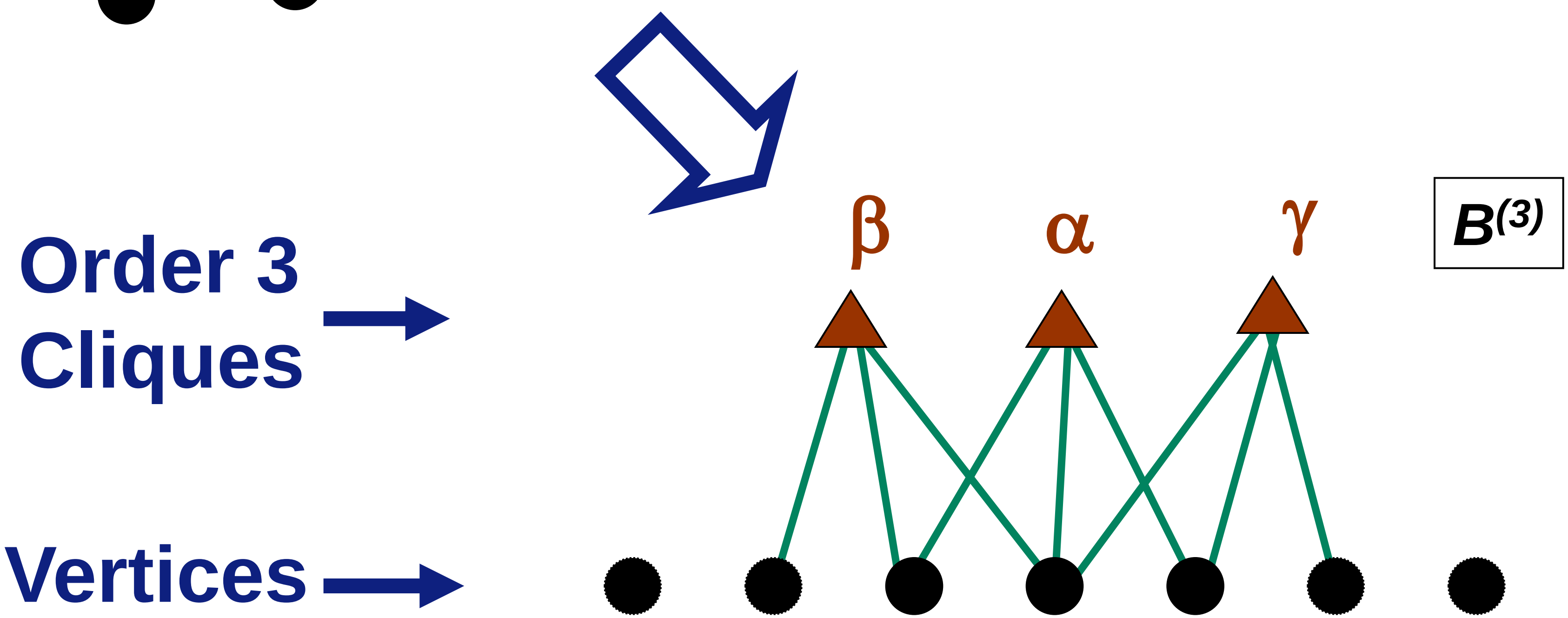
- nine cliques of order 2 = edges

I will only use cliques of one order unlike most Social Science work.

Clique-Vertex Bipartite Graph Construction

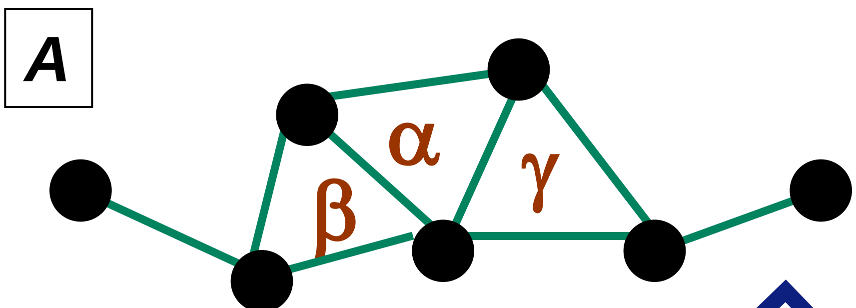
Adjacency matrix $\boldsymbol{B^{(n)}_{i\alpha} = 1}$
if vertex $\boldsymbol{i}$ is in clique $\boldsymbol{\alpha}$
of fixed order $\boldsymbol{n}$

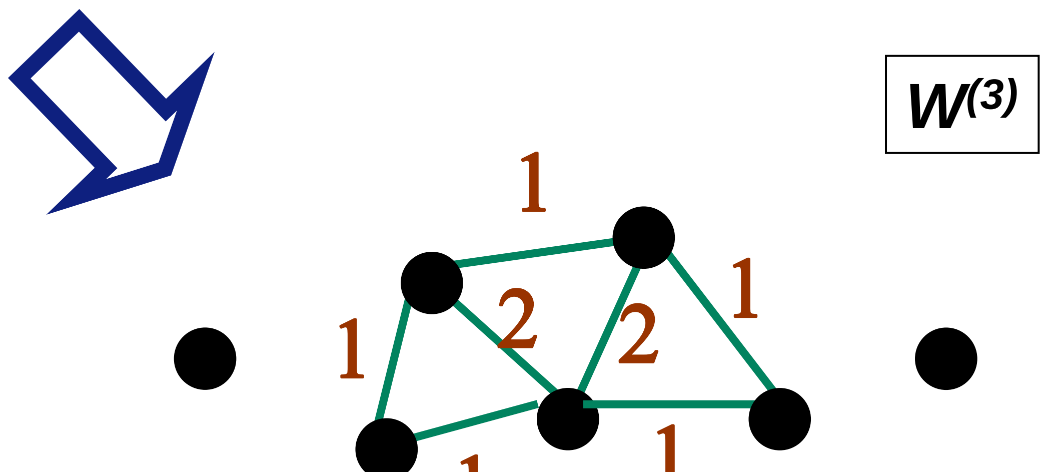




Usual Clique Overlap Graph $W^{(n)}$

- Vertices unchanged
- Edge weights equal to the number of ***n***-cliques containing that edge in in the original graph ***A*.**





Unweighted $W^{(3)}$ is isomorphic to an *n*-regular hypergraph

Clique Overlap Graph $W^{(n)}$

- In Social Science clique overlap is usually studied with cliques of several orders (e.g. subset of maximal cliques)

$$W_{ij} = \Sigma_n W^{(n)}_{ij}$$

- $W^{(n)}$ without weights is isomorphic to an ***n***-uniform hypergraph

Drawback of $W^{(3)}$ is that the vertices remain the same

$\Rightarrow$ same old vertex centric viewpoint

Definition of the Basic Clique Graph $C^{(n)}$

Project bipartite graph onto clique vertices.

⇒Weight of an edge in the basic Clique Graph $C^{(n)}$ records the number of vertices common to two ***n***-cliques in in the original graph ***A*.**

$$C_{\alpha\beta}^{(n)} = \sum_{i} B_{i\alpha}^{(n)} B_{i\beta}^{(n)} \left(1 - \delta_{\alpha\beta}\right)$$

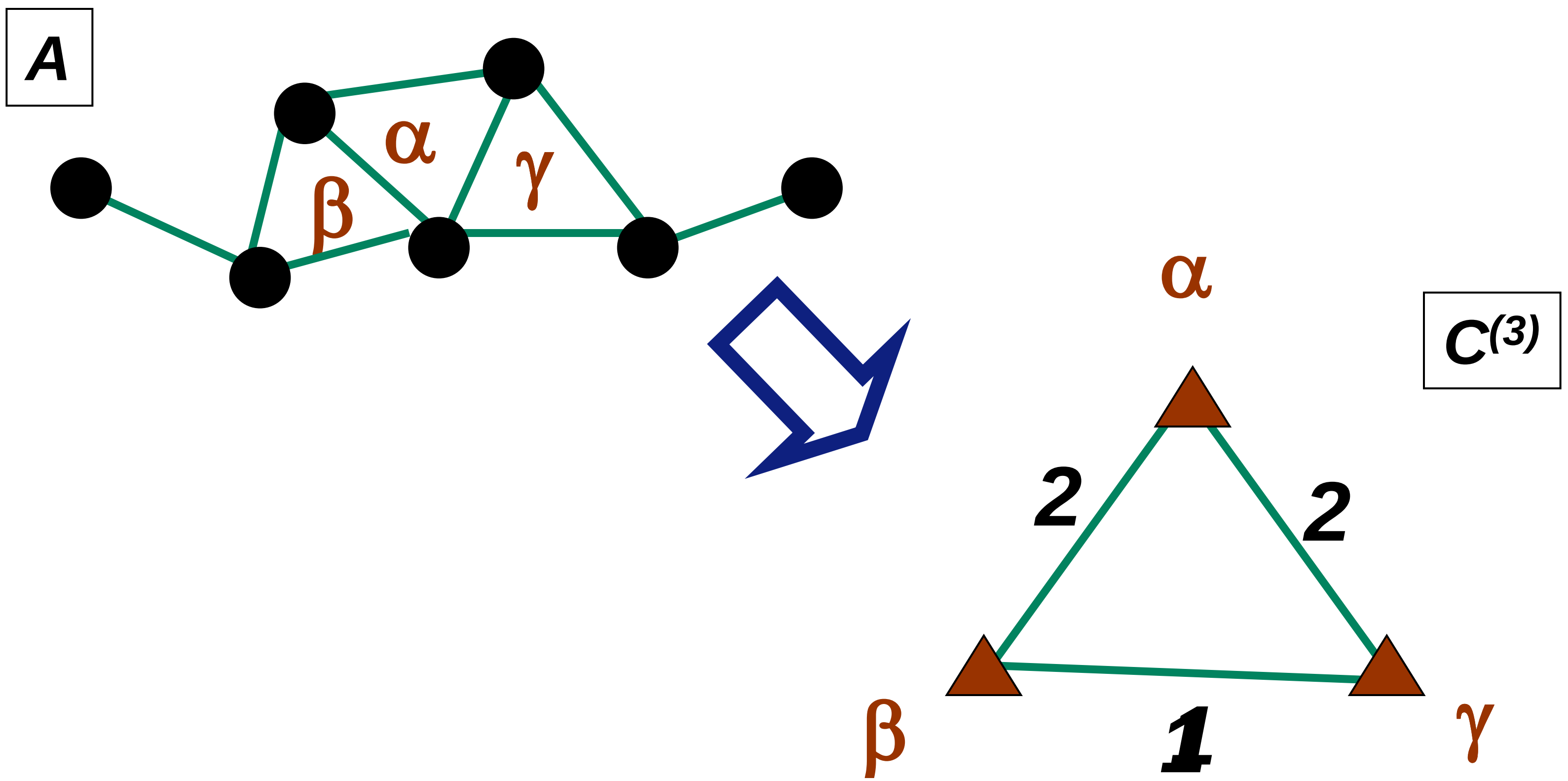
i = vertex ***A***

α,β* = *n-cliques in ***A***

(Rarely used in the Social Science literature)

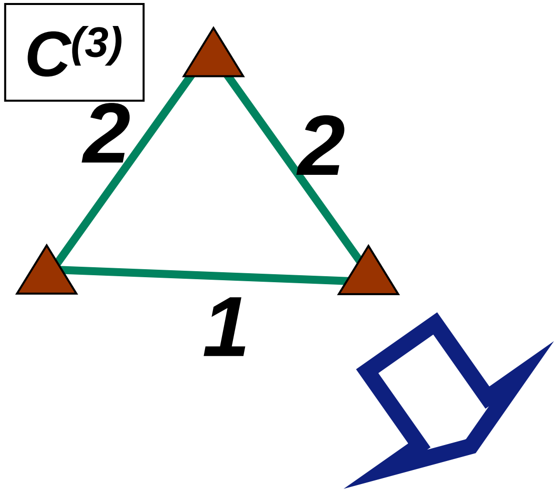
Edge Weights in Basic Clique Graph

Weight of an edge in the basic Clique Graph ***C*** records the number of vertices common to two ***n***-cliques in in the original graph ***A*.**



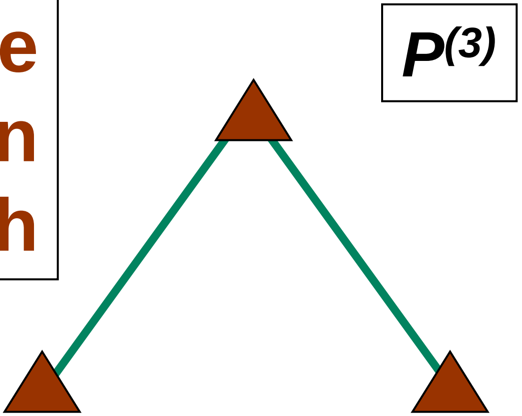
Clique Percolation Graph Construction

Threshold edges of clique Graph with *(n-1)*



i.e. only retain cliques which in original graph ***A*** share all but one vertex with another clique

Clique Percolation Graph



Connected Components = Communities [Palla et al, '05]

$$P_{\alpha\beta} = \Theta\left(C_{\alpha\beta} - n + 1\right)$$

Basic Clique Graph Weights – A problem

- The weight of basic Clique Graph $C^{(3)}$ records the number of vertices common to two n-cliques in in the original graph A.

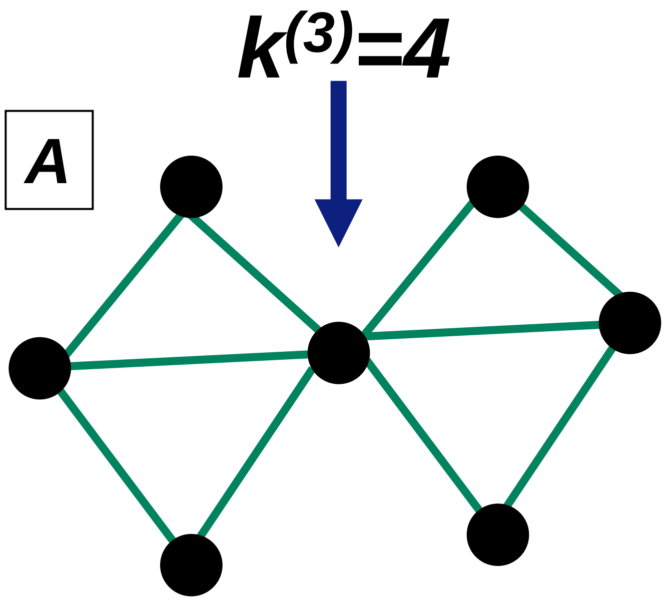
- **Thus a vertex which is in $k^{(n)}$ cliques will contribute a total weight of**

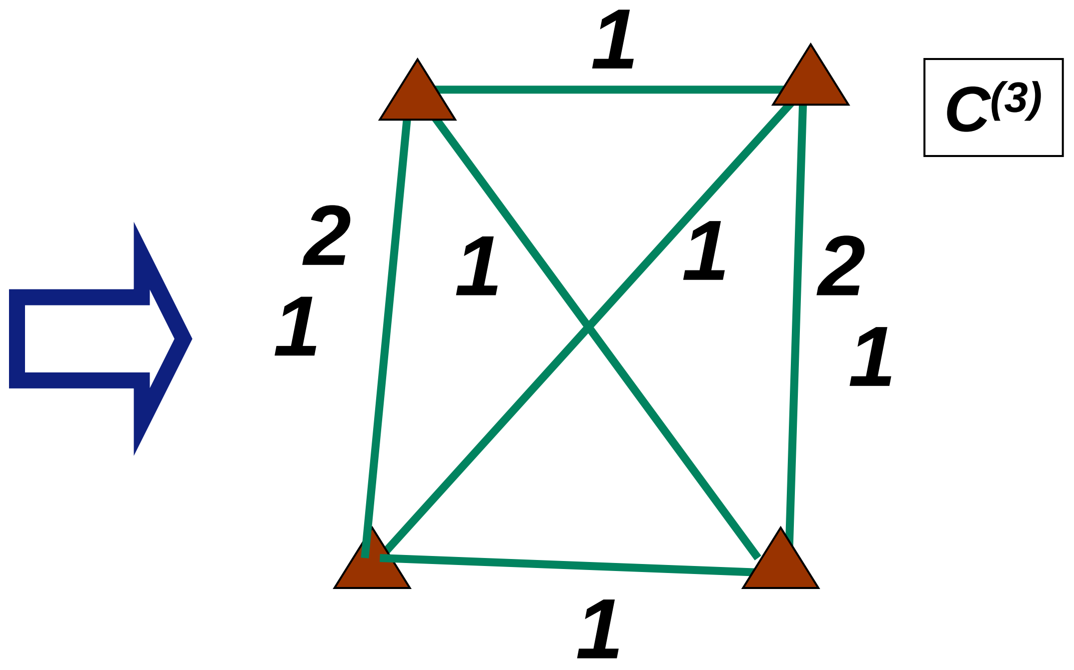
$$k^{(n)} (k^{(n)} -1)/2$$

to the basic clique graph $C^{(3)}$.

Basic Clique Graph Weights – A problem

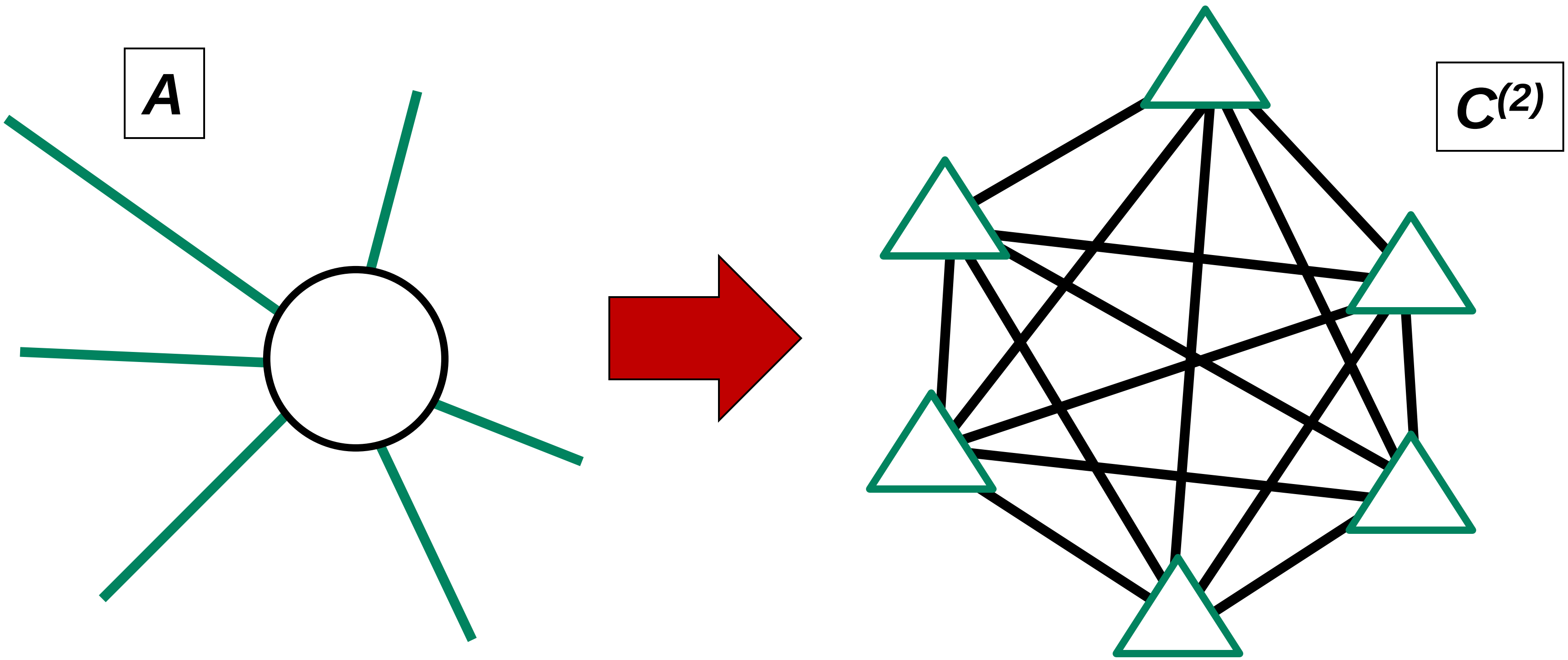
- **A vertex in a large number of cliques contributes too much weight to the clique graph *C*.**





The Problem for ***n*=2** Edges/2-cliques

High degree vertices in original graph ***G*** over represented by edges in Line Graph $\boldsymbol{C^{(2)}(A)}$.



Degree *k* vertex

k(k-1)/2 edges

Better Clique Graph Weight Construction

A vertex common to a pair of cliques contributes a weight equal to

$$1/(k^{(n)}-1)$$

to an edge in the new weighted Clique Graph $D^{(n)}$.

$$D_{\alpha\beta}^{(n)} = \sum_i \frac{B_{i\beta}^{(n)} B_{i\beta}^{(n)}}{k^{(n)}-1}\left(1-\delta_{\alpha\beta}\right)$$

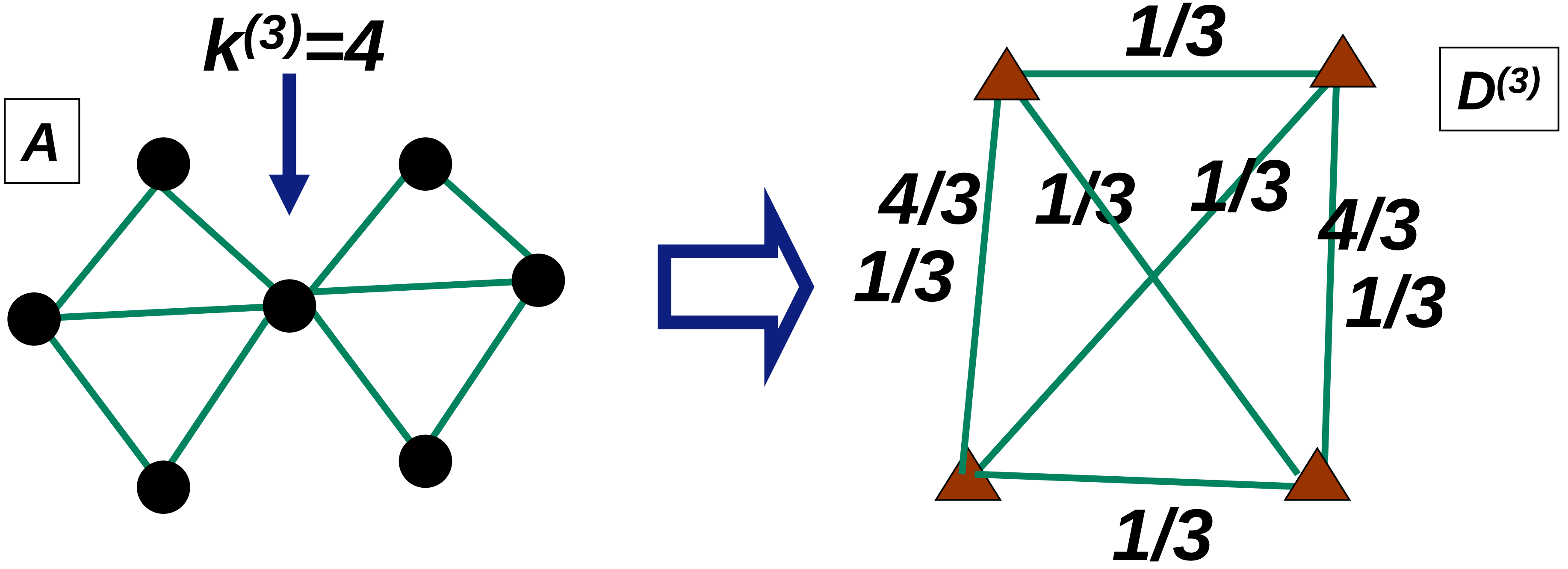
(vertex is in $k^{(n)}$ n-cliques)

Better Clique Graph Weight Construction

A vertex common to a pair of cliques contributes a weight equal to

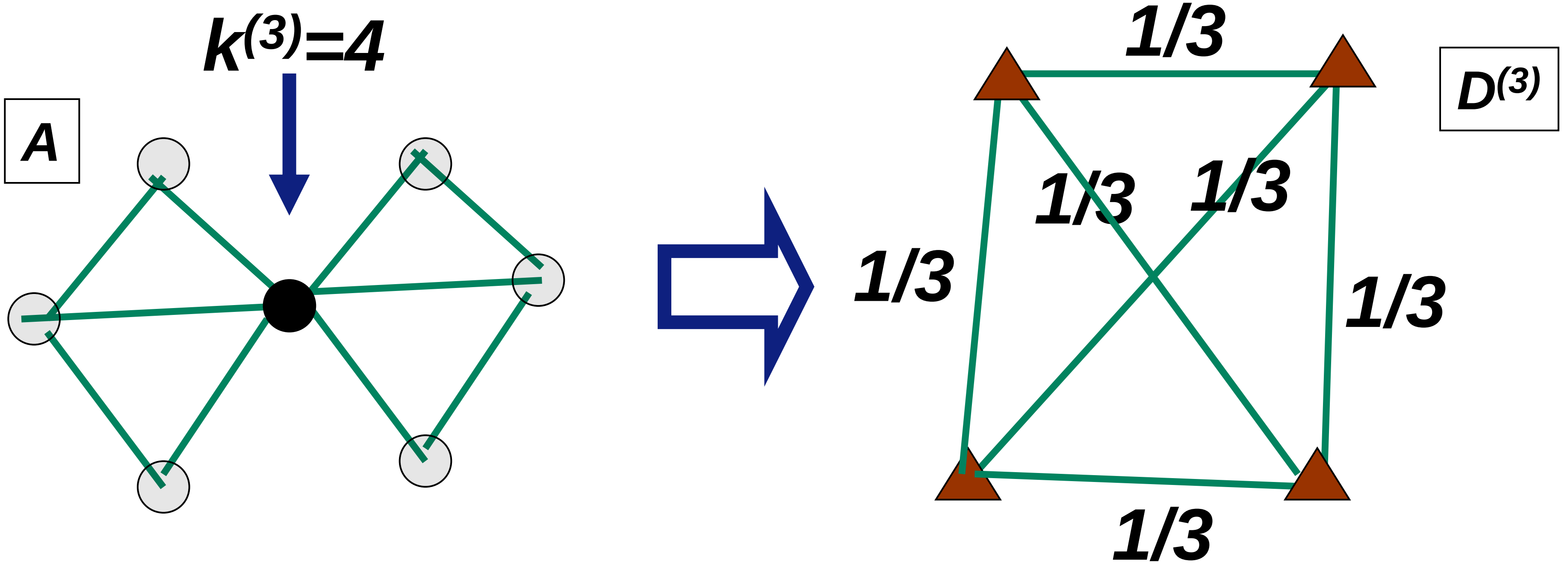
$$1/(k^{(n)} - 1)$$

to an edge in the new weighted Clique Graph $D^{(3)}$. (vertex is in $k^{(n)}$ n-cliques)



Better Clique Graph Weight Construction

Now each vertex *i* in the original graph A contributes a weight to the new weighted Clique Graph $D^{(n)}$ proportional to the number of cliques containing that vertex, $k^{(n)}$.



Applications

Now apply ***ANY*** standard vertex based algorithm to vertices of clique graph $\boldsymbol{C^{(n)}}$ or $\boldsymbol{D^{(n)}}$ or other weighted version

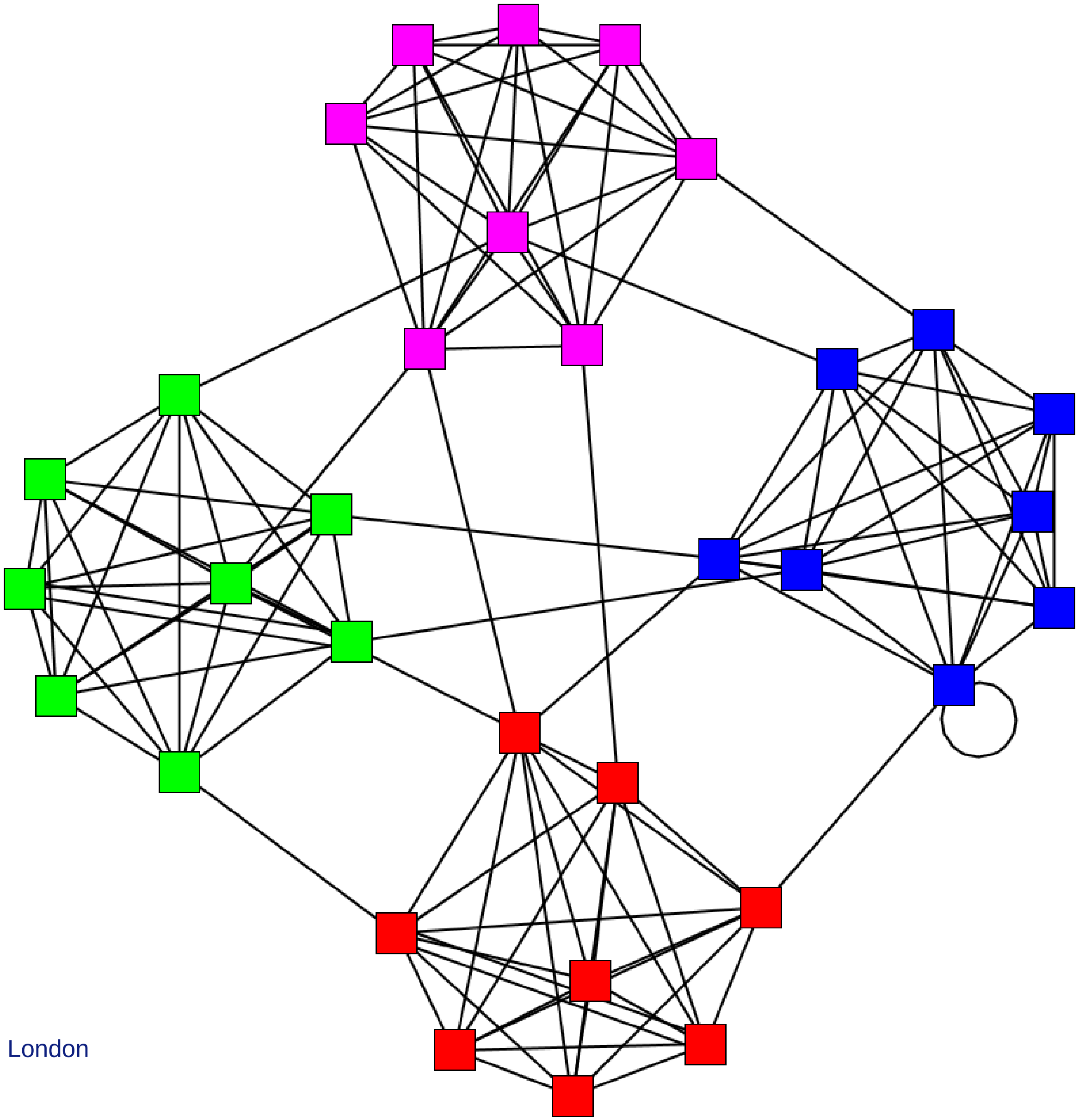
- Degree distribution ⇒ clique overlap distribution
- Cluster coefficient ⇒ overlap of order ***n*** cliques
- Community detection ⇒ clique communities
- etc, etc

- Graphs, Data and our Vertex Centric Viewpoint
- Clique Graphs
- **Overlapping Communities**
- Line Graphs and Overlapping Communities
- Conclusions

Communities in Networks

Rough definition:-

A *community* is a subgraph which is more tightly connected than average



Community
= cluster
= cohesive subgroup

Advantages of Vertex Partition Communities

- Simplest way of assigning communities across whole graph
- Appropriate for some problems
  e.g. assigning pixels in image analysis
- Vast amount of development of theory and methods
  e.g. free code works on graph of 10^8 vertices
  in 20min
  (Louvain method, Blondel et al. 2008)

Limitations of Vertex Partition Communities

In many applications it is too simplistic to assign one vertex to one community

Examples:

- Friendship networks
- Academic papers

Who am I? Communities in Friendship Networks

Friendship networks have:-

- people as vertices
- edges if friends

We all have different types of friends:-

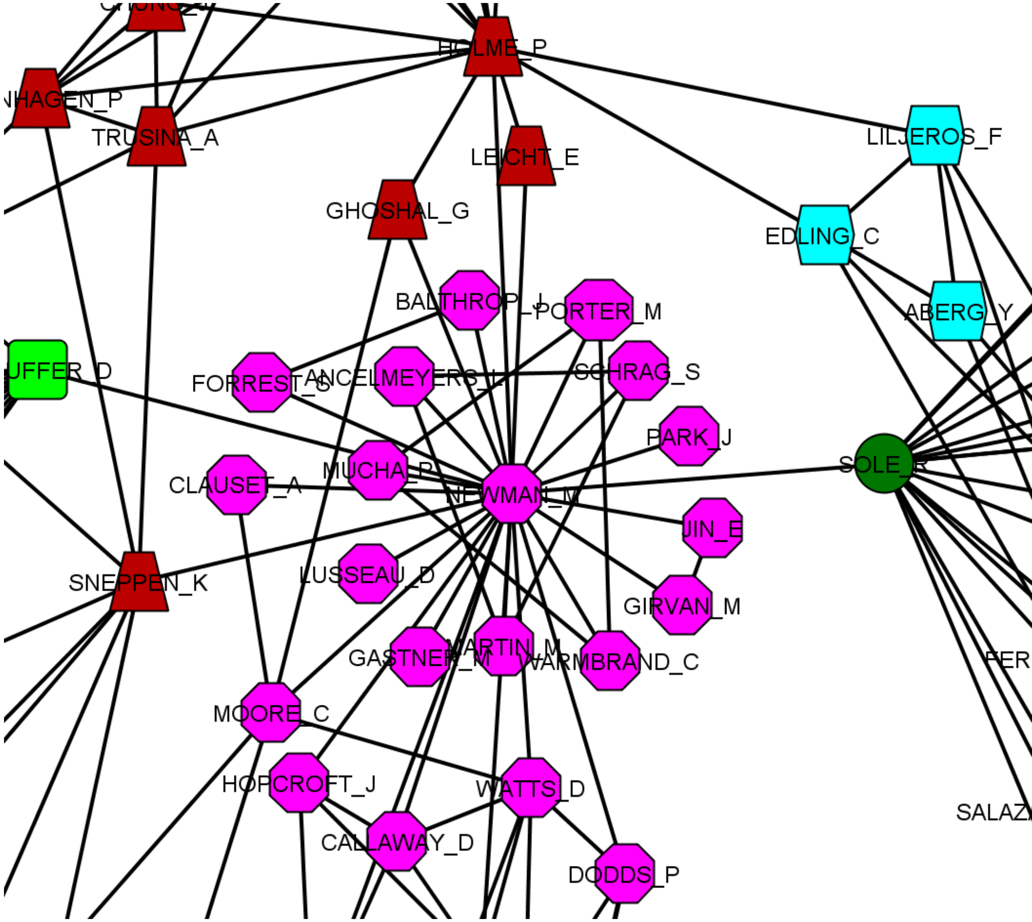
- Family
- Neighbours
- Work Colleagues

I am not just an Evans or an Imperial employee

What am I? Communities from Academic Papers

Coauthorship Networks:-

- authors as vertices
- edges between coauthors of a paper



Academics who work across boundaries always assigned to one community

Clique Graphs for Community Detection

- Produce Clique graph
- Use favourite vertex partition method in clique graph
- Hence assign each clique to a single community
- Deduce community membership of vertices and edges of original graph from clique membership.

⇒Thus vertices and edges maybe part of more than one community.

⇒ Overlapping Communities.

Karate Club [Zachary 1977]

- Community of 34 members of a Karate Club
- Split into two parts during study
  - Group centred on Club Officers
  - Group centred on instructor
- Zachary partitioned vertices into two subsets using Ford-Fulkerson algorithm (source-sink) which matched actual split except for one individual

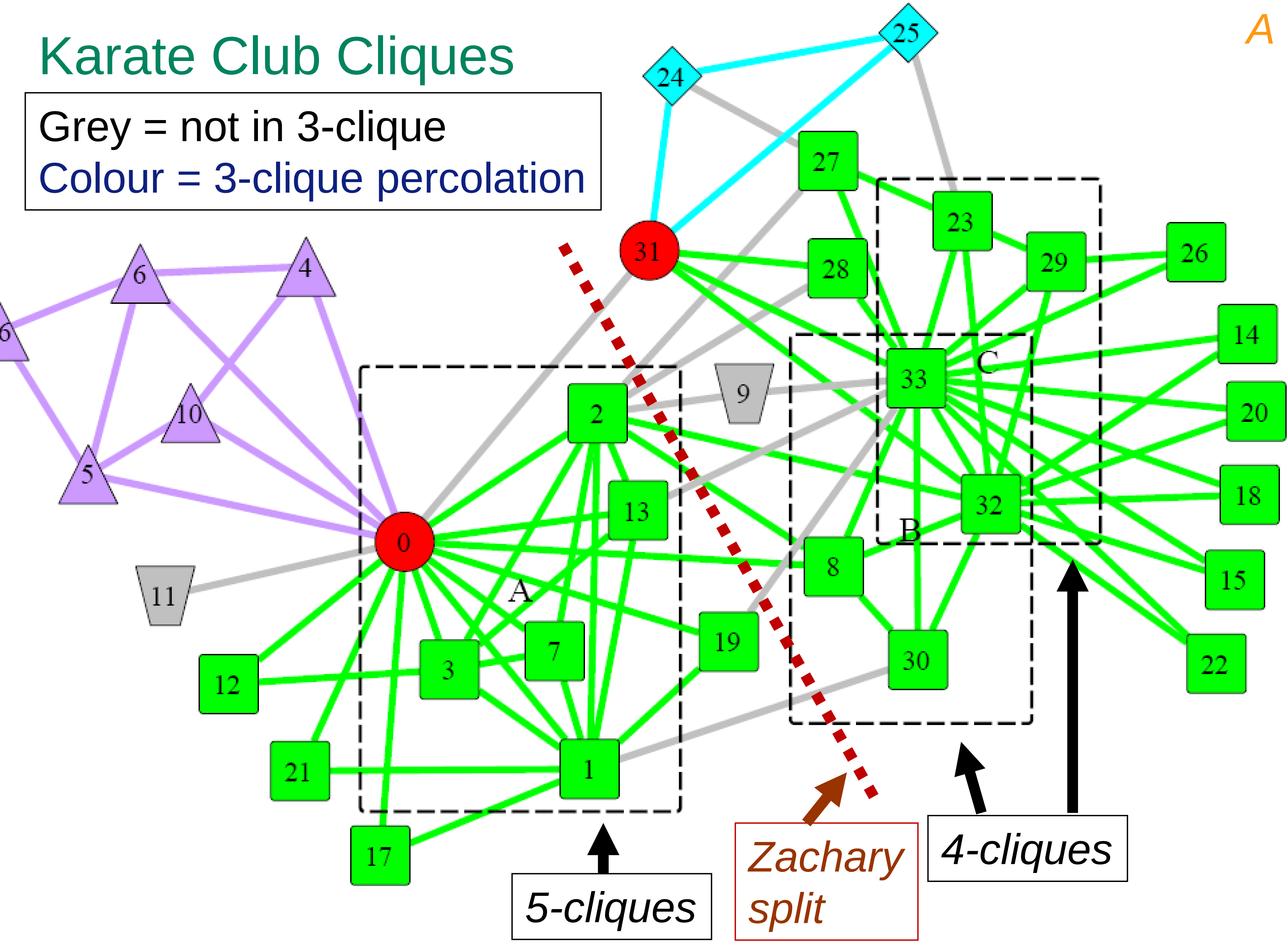
Cliques in Karate Club

- Almost everything is in a 3-clique
  - just 2 vertices and 9 edges not in 3-cliques

- One group of 6 vertices centred on Instructor have two 5-cliques with 4 common vertices (percolating)

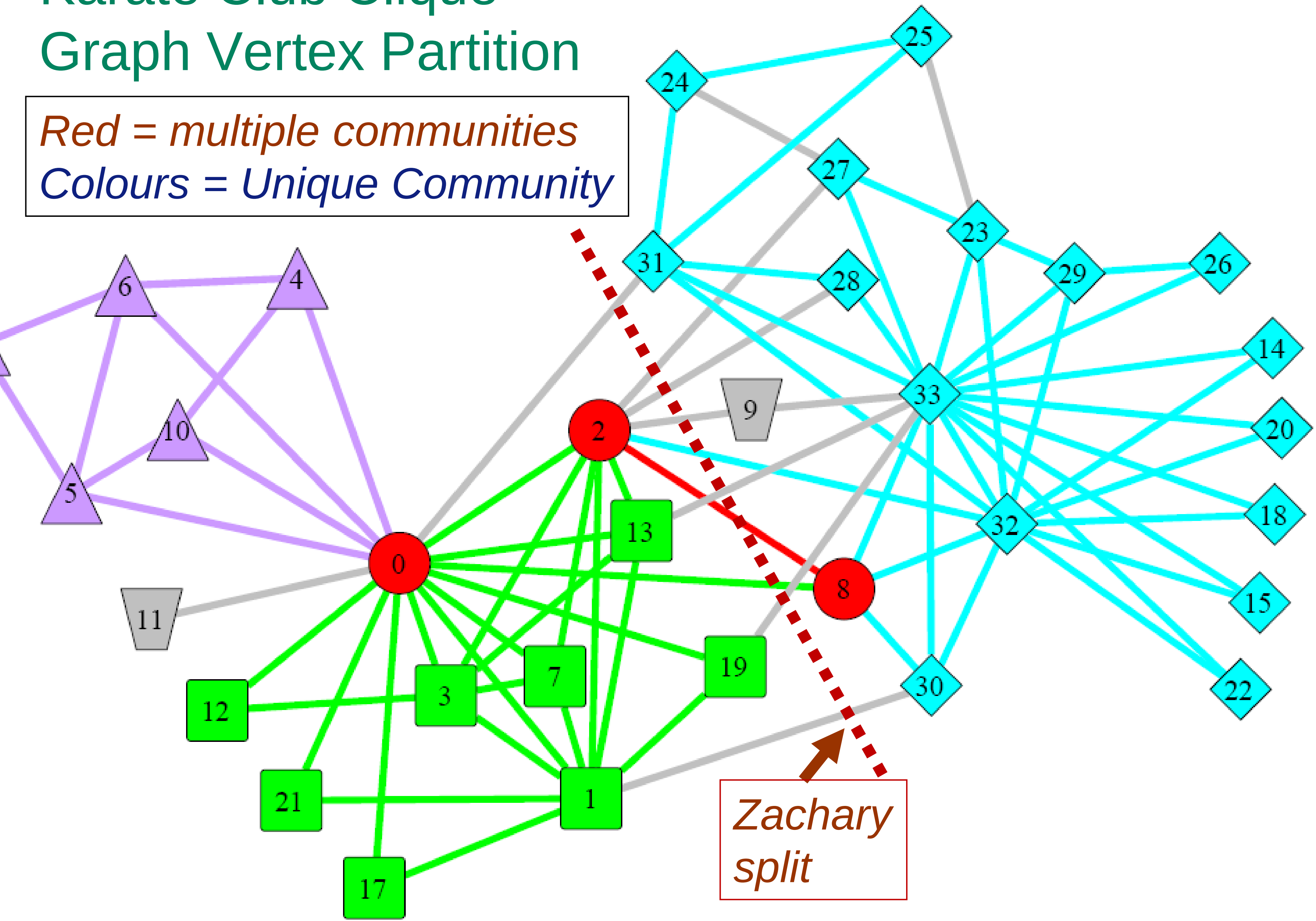
- Two other 4-cliques centred on chief officers, only 2 common vertices (non-percolating)

Karate Club Cliques



Karate Club Clique
Graph Vertex Partition

Red = multiple communities
Colours = Unique Community



Analysis of Karate Club Results for Triangles

- Clique percolation doesn't work here

- Despite the fact that almost everything is in a 3-clique, vertex partitioning of a 3-clique graph $D^{(3)}$ works extremely well.

- Graphs, Data and our Vertex Centric Viewpoint
- Clique Graphs
- Overlapping Communities
- Line Graphs and Overlapping Communities
- Conclusions

2-Clique Graph = Line Graph

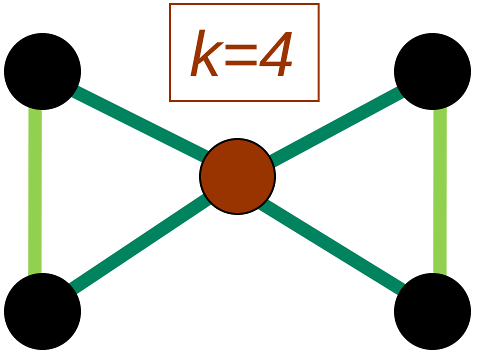
- The 2-clique (=edge) graph $C^{(2)}$ is an unweighted graph called a **Line Graph** ***L(A)***
  – Long history [Whitney, '32; Krausz '43; Harary & Norman '60]
  – Much work in mathematical literature

- The 2-clique (=edge) graph $D^{(2)}$ is a **Weighted Line Graph** ***WL(A)***

[TSE+Lambiotte '09]

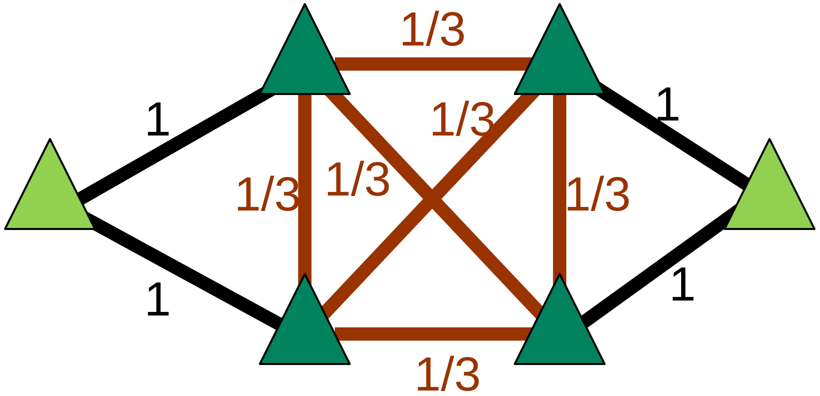
Example – Bow Tie Graph

Graph ***A***



Weighted Line Graph

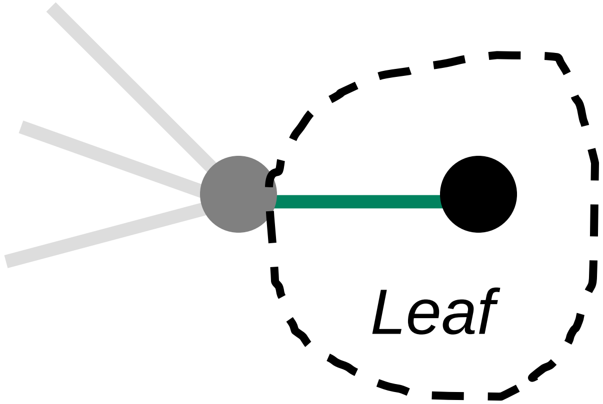
$\boldsymbol{WL(A) = D^{(2)}}$

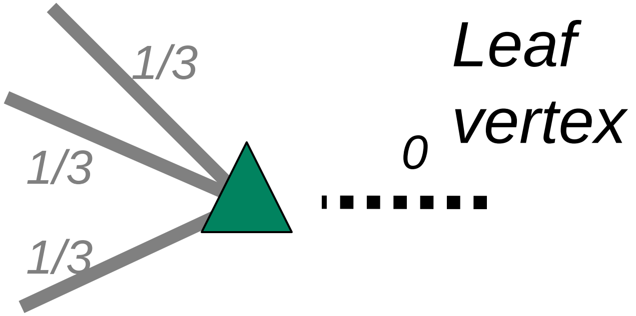


Edges 6
Strength 2

Nice Property of Weighted Line Graph

- Strength of each vertex in ***WL(A)*** is usually ***2*** since every edge in ***A*** has two ends
- Exception for edges of "leaves" in ***A*** which produce strength ***1*** vertices in ***WL(A)***.





From a Vertex to an Edge Centric Viewpoint

- Take your graph ***A*** with ***N*** vertices and ***<k>*** edges

- Make a Weighted Line Graph ***WL(A)*** with ***N<k>*** vertices and ***O(N<k²>)*** edges

- Run any vertex based algorithm on ***WL(A)*** and you are running it on the edges of ***G***.

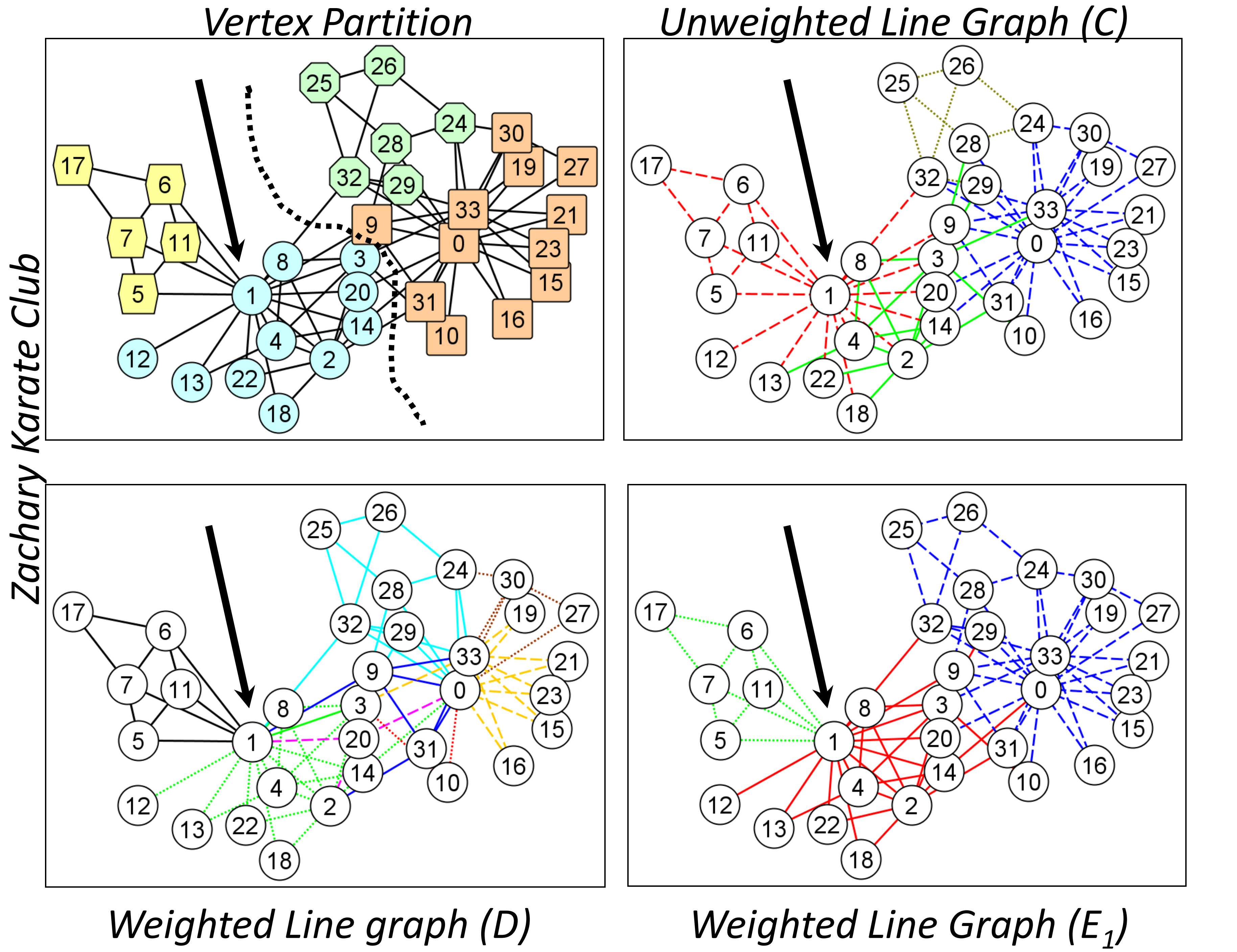
Overlapping Communities

Use weighted line graphs to study edge colourings and hence to deduce overlapping communities

- Karate Club
- South Florida Words Association Data
- Words from Titles of Scientific Papers

Karate Club [Zachary 1977]

- Community of 34 members of a Karate Club
- Split into two parts during study
  - Group centred on Club Officers
  - Group centred on instructor
- Zachary partitioned vertices into two subsets using Ford-Fulkerson algorithm (source-sink) which matched actual split except for one individual

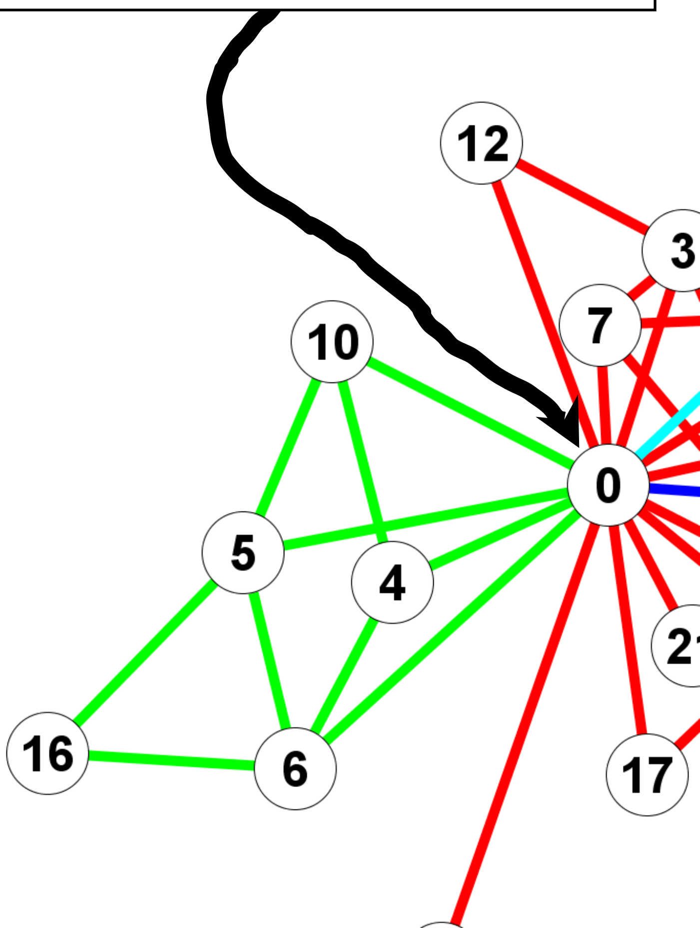

Vertex Partition
Unweighted Line Graph (C)
Zachary Karate Club
Weighted Line graph (D)
Weighted Line Graph (E1)

Karate Club Analysis

Vertices in One Edge Community

#	k	Fraction k In Green C
5	4	100%
6	4	100%
10	3	100%
4	3	100%
16	2	100%
0 (Mr_Hi)	16	25%

Mr Hi (the Instructor) bridges several groups



Karate Club Edge Partition

Vertices can be members of many communities

An overlapping community structure for vertices

Name	Community	Total k	k in C
0 Mr Hi	0	16	10
	1		4
	2		1
	3		1
33 John A	3	17	12
	0		3
	2		2

South Florida Word Association Data

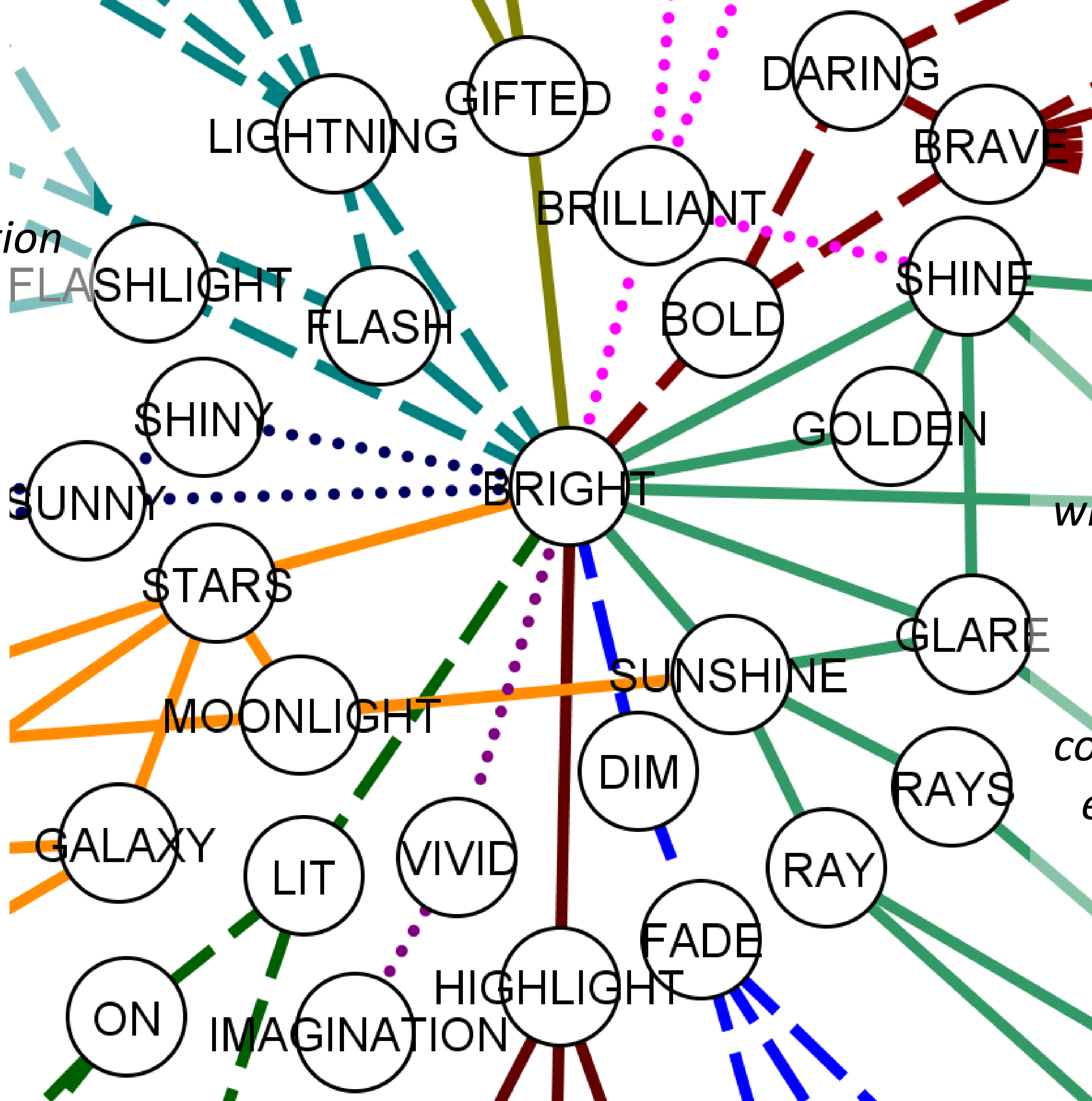
Data from 6,000 participants, nearly three-quarters of a million responses to 5,019 stimulus words.

Original graph A has

- Stimulus words as vertices
- Edges connecting words if paired in data more than a specified threshold.

http://w3.usf.edu/FreeAssociation/

South Florida Word Association Data



Only showing vertices which have 90% of edges in one edge community except for BRIGHT

Title of Papers Data

Data based on collection of science papers from a single institution over several years. Forms a bipartite graph of:-

- 26255 vertices representing the papers
- 17761 vertices representing terms - stemmed words from titles after stop words removed
- 210229 edges

Edge Partition of Terms in Paper Titles

- Some words have all edges in one partition
  – they define these communities e.g. cassini
- Other words have edges in several communities
  – stop words e.g. signature

Stem	Total k	k in C
interplanetari	78	78
cassini	62	62
heliospher	59	59
magnetopaus	53	53
spacecraft	52	52
signatur	91	32
solitari	30	10
radar	21	7
mhd	18	6

- Graphs, Data and our Vertex Centric Viewpoint
- Clique Graphs
- Overlapping Communities
- Line Graphs and Overlapping Communities
- Conclusions

Conclusions

- Clique graphs move focus from vertices to cliques with minimal effort
- Altering weights avoids problem of over representation of high degree vertices
- 2-clique graphs are *Weighted Line Graphs* [Evans and Lambiotte, 09]
- Generalisation to motifs straightforward
- Community detection on clique graph produces overlapping vertex communities for original graph