Online Social Networks and Media

Homophilly

Networks with Positive and Negative ties

Chapter 4, from D. Easley and J. Kleinberg book

HOMOPHILLY

Introduction

Surrounding context: factors other than node and edges that affect how the network structure evolves

Homophily: people tend to be similar to their friends

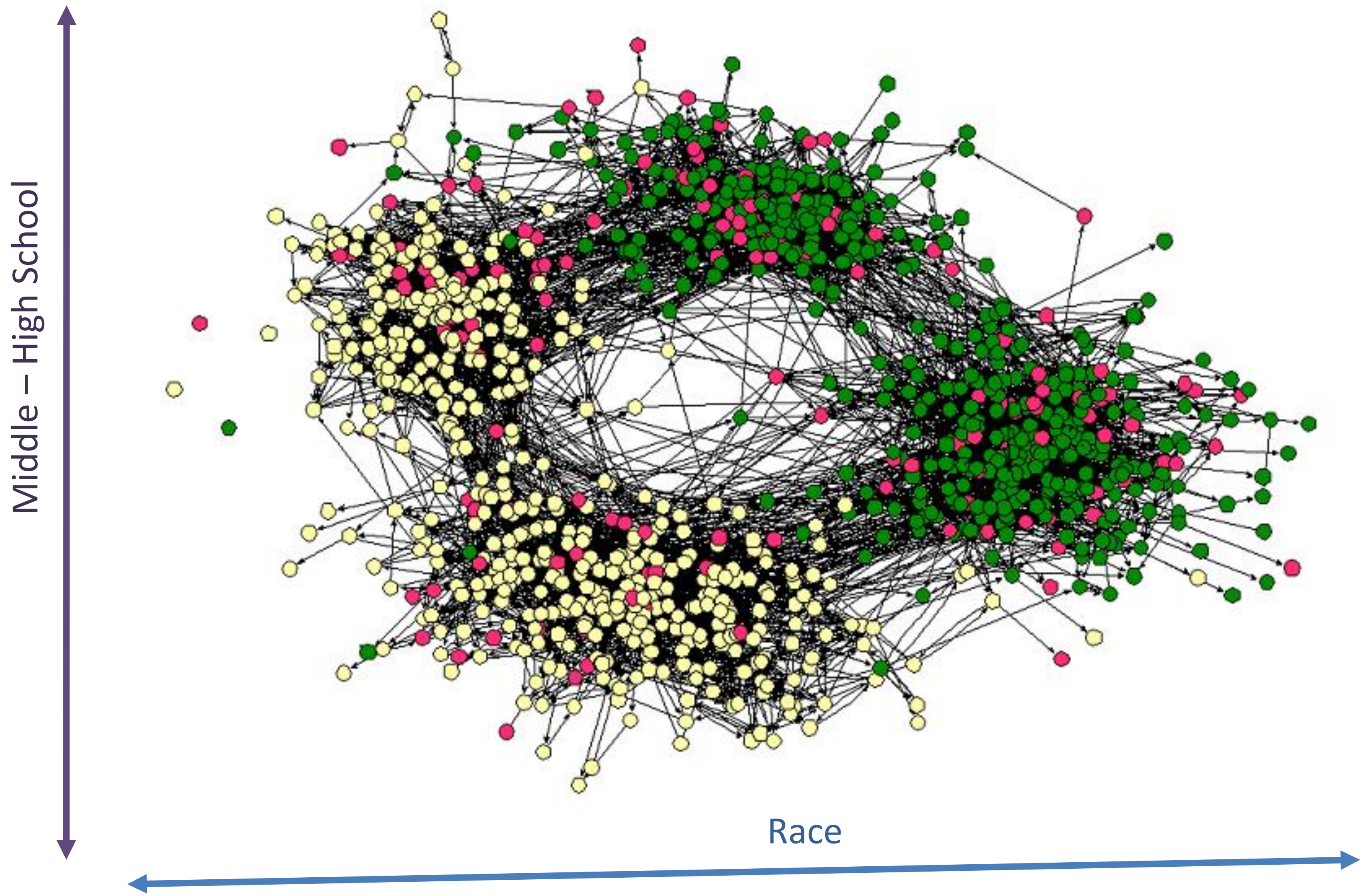
Αριστοτέλης love those who are like themselves

Πλάτωνα Όμοιος ομοίω αεί πελάζει (similarity begets friendship)

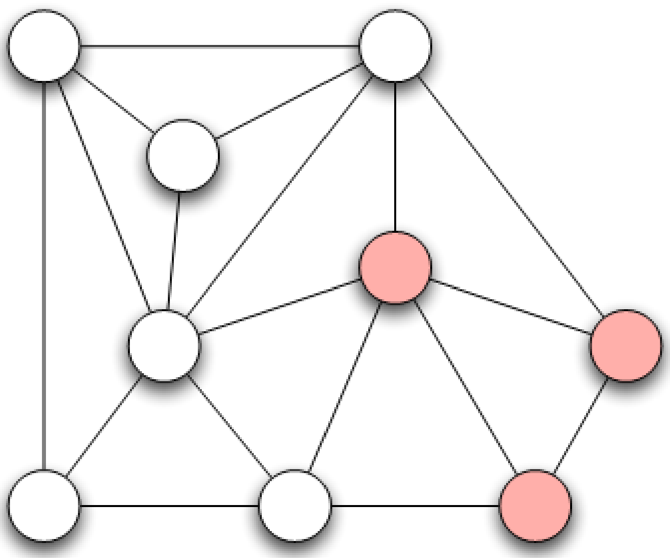
Birds of a feather flock together

Factors intrinsic to the network (introduced by a common friend) and contextual factors (eg attend the same school)

Homophily



Measuring Homophily



If the fraction of cross-gender edges is significantly less than expected, then there is evidence for homophily

gender male with probability p
gender female with probability q

Probability of cross-gender edge?

$$\frac{\#cross\_gender\_edges}{\#edges} << 2pq$$

Measuring Homophily

- "significantly" less than
- Inverse homophily
- Characteristics with more than two values:
  - Number of heterogeneous edges (edge between two nodes that are different)

Mechanisms Underlying Homophily: Selection and Social Influence

Selection: tendency of people to form friendships with others who are like then

Socialization or Social Influence: the existing social connections in a network are influencing the individual characteristics of the individuals

Social Influence <u>as the inverse</u> of Selection

Mutable & immutable characteristics

The Interplay of Selection and Social Influence

Longitudinal studies in which the social connections and the behaviors within a group are tracked over a period of time

Why?

- Study teenagers, scholastic achievements/drug use (peer pressure and selection)
- Relative impact?
- Effect of possible interventions (example, drug use)

The Interplay of Selection and Social Influence

Christakis and Fowler on obesity, 12,000 people over a period of 32-years

People more similar on obesity status to the network neighbors than if assigned randomly

Why?
(i) Because of selection effects, choose friends of similar obesity status,
(ii) Because of confounding effects of homophily according to other characteristics that correlate with obesity
(iii) Because changes in the obesity status of person's friends was exerting an influence that affected her

(iii) As well -> "contagion" in a social sense

Tracking Link Formation in Online Data: interplay between selection and social influence

- Underlying social network
- Measure for behavioral similarity

Wikipedia
Node: Wikipedia editor who maintains a user account and user talk page
Link: if they have communicated with one writing on the user talk page of the other

Editor's behavior: set of articles she has edited

$$\frac{|N_A \cap N_B|}{|N_A \cup N_B|}$$

Neighborhood overlap in the bipartite affiliation network of editors and articles consisting only of edges between editors and the articles they have edited

FACT: Wikipedia editors who have communicated are significantly more similar in their behavior than pairs of Wikipedia editors who have not (homomphily), **why?**
Selection (editors form connections with those have edited the same articles) vs Social Influence (editors are led to the articles of people they talk to)

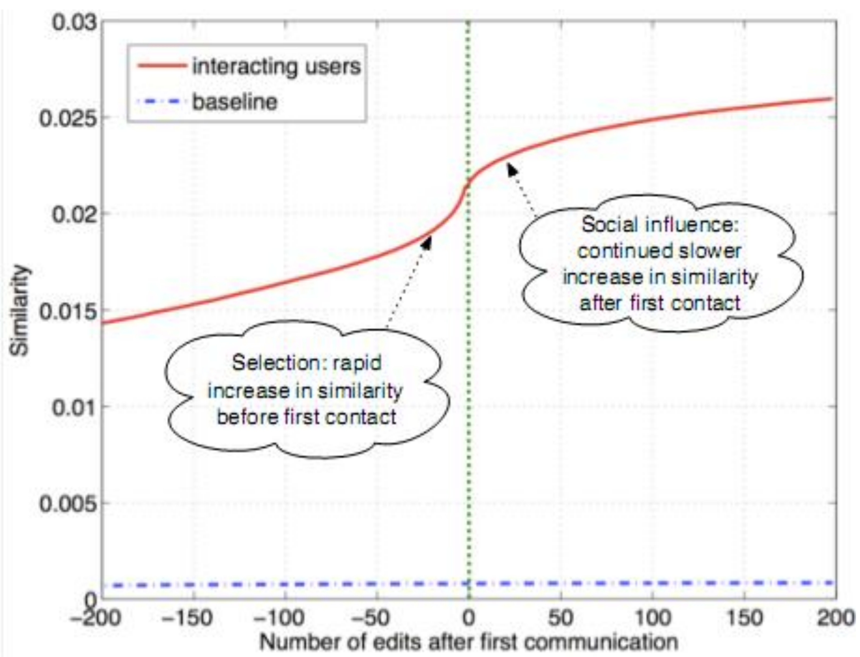
Tracking Link Formation in Online Data: interplay between selection and social influence

Actions in Wikipedia are time-stamped

For each pair of editors A and B who have ever communicated,

- ○ Record their similarity over time
- ○ Time 0 when they first communicated -- Time moves in discrete units, advancing by one “tick” whenever either A or B performs an action on Wikipedia
- ○ Plot one curve for each pair of editors

Average, single plot: average level of similarity relative to the time of first interaction



Similarity is clearly increasing both before and after the moment of first interaction (both selection and social influence)

Not symmetric around time 0 (particular role on similarity): Significant increase before they meet

Blue line shows similarity of a random pair (non-interacting)

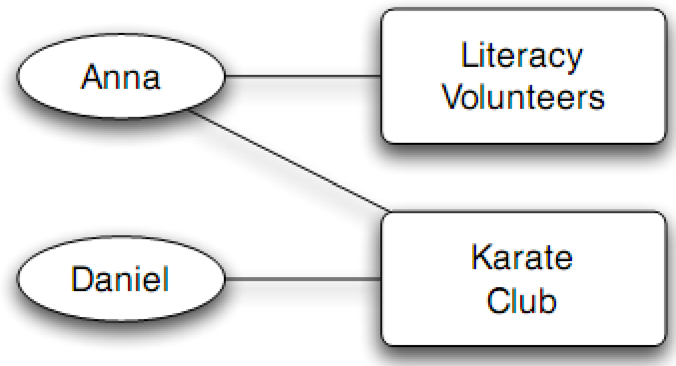
Affiliation

A larger network that contains both people and context as nodes

foci

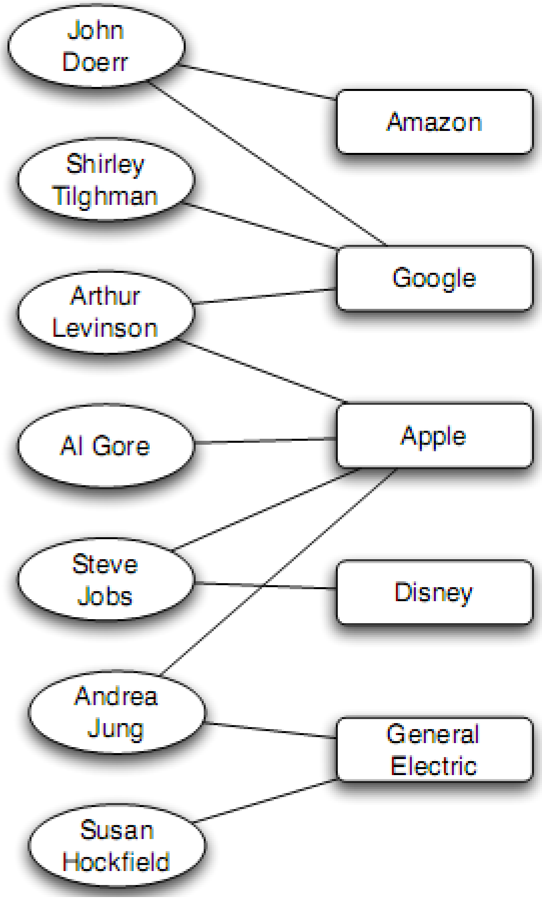
Affiliation network

Bipartite graph
A node for each person and a node for each focus
An edge between a person A and focus X, if A participates in X



Affiliation

Example:
Board of directors



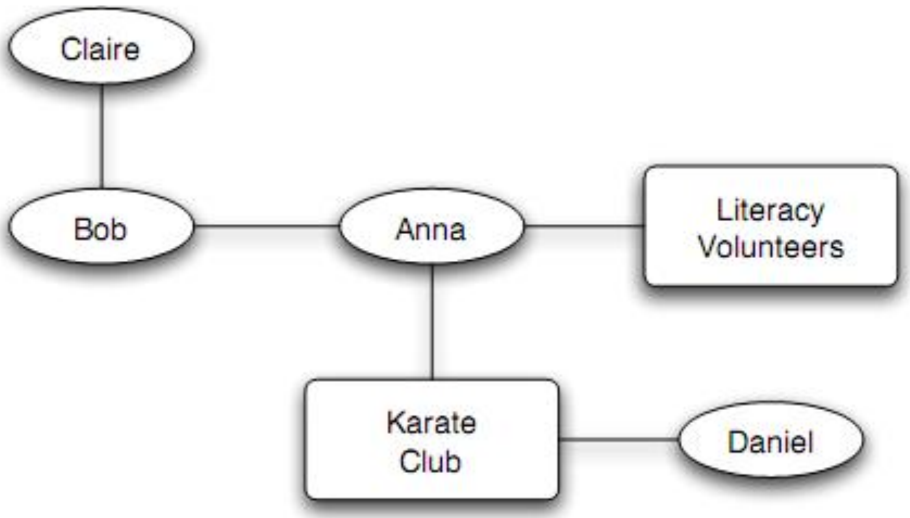
- Companies implicitly links by having the same person sit on both their boards
- People implicitly linked by serving together on a aboard
- Other contexts, president of two major universities and a former Vice-President

Co-evolution of Social and Affiliation Networks

Social Affiliation Network

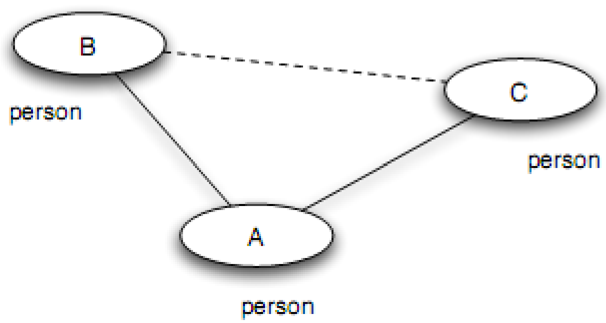
Two type of edges:

1. *Friendship:* between two people
2. *Participation:* between a person and a focus

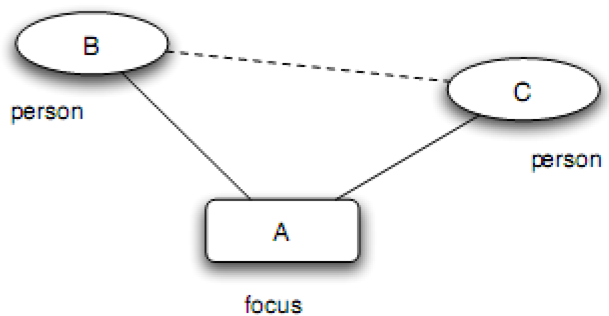


- Co-evolution reflect the interplay of selection and social influence: if two people in a shared focus opportunity to become friends, if friends, influence each other foci.

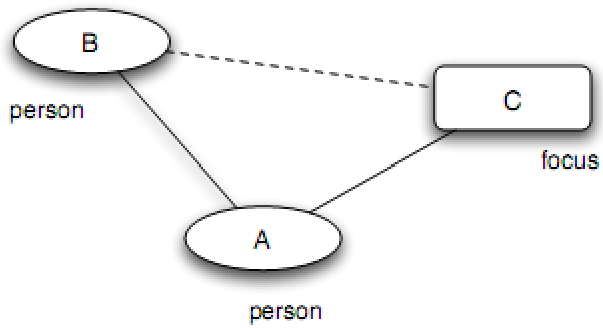
Co-evaluation of Social and Affiliation Networks: Closure process



Triadic closure: (two people with a friend in common - A introduces B to C)

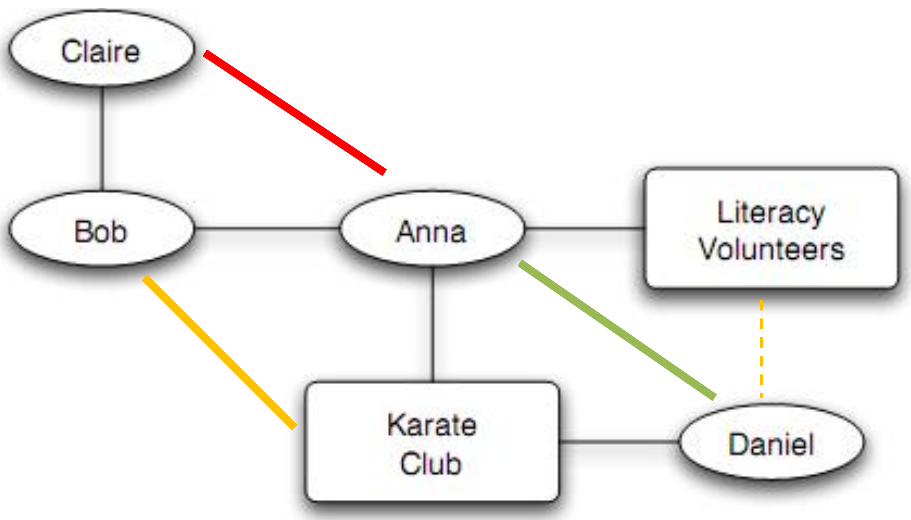


Focal closure: (two people with a focus in common - focus A introduces B to C) (selection)



Membership closure: (a person joining a focus that a friend is already involved in - A introduces focus C to B) (social influence)

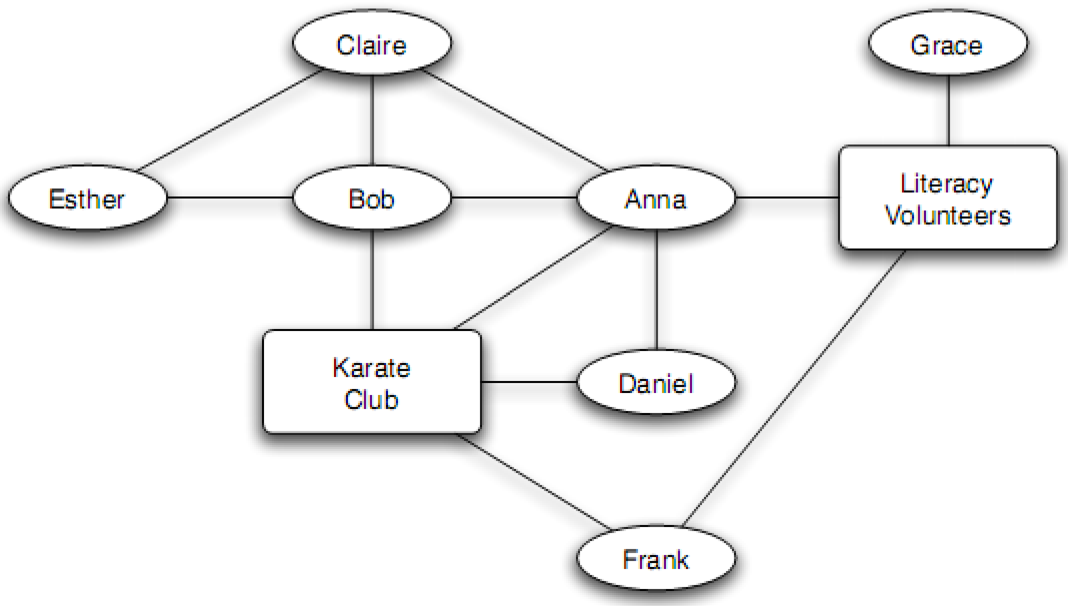
Co-evaluation of Social and Affiliation Networks



Tracking Link Formation in Online Data: triadic closure

Triadic closure:

- How much more likely is a link to form between two people if they have *a friend* in common
- How much more likely is a link to form between two people if they have *multiple* friends in common?



Tracking Link Formation in Online Data: triadic closure

Take two snapshots of the network at different times

I. For each *k*, identify all pairs of nodes that have exactly *k* friends in common in the first snapshot, but who are not directly connected

II. Define *T(k)* to be the fraction of these pairs that have formed an edge by the time of the second snapshot

III. Plot *T(k)* as a function of *k*

T(0): rate at which link formation happens when it does not close any triangle
T(k): the rate at which link formation happens when it does close a triangle (*k* common neighbors, triangles)

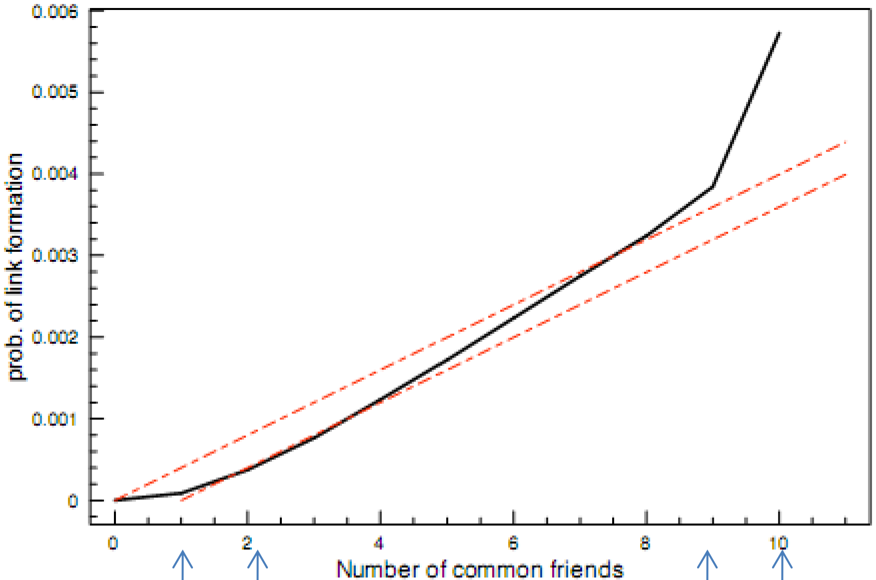
Tracking Link Formation in Online Data: triadic closure

E-mail ("who-talks-to-whom" dataset type)
Among 22,000 undergrad and grad students (large US university)
For 1-year

Network evolving over time

- At each instance (snapshot), two people join, if they have exchanged e-mail in each direction at some point in the past 60 days
- Multiple pairs of snapshots ->
- Built a curve for T(k) on each pair, then average all the curves

Snapshots – one day apart (average probability that two people form a link per day)



From 0 to 1 to 2 friends

From 8 to 9 to 10 friend (but on a much smaller population)

✓ Almost linear

Having two common friends produces significantly more than twice the effect compared to a single common friend

Tracking Link Formation in Online Data: triadic closure

Baseline model:

Assume triadic closure:

Each common friend two people have gives them an independent probability *p* of forming a link each day

For two people with *k* friend in common,

Probability not forming a link on any given day

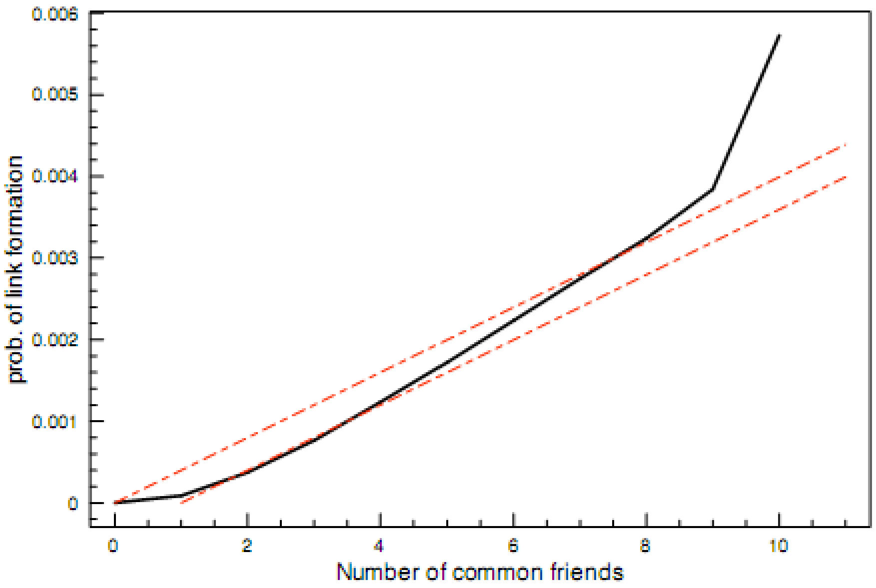
$$(1-p)^k$$

Probability forming a link on any given day

$$T_{baseline}(k) = 1 - (1-p)^k$$

Given the small absolute effect of the first common friend in the data

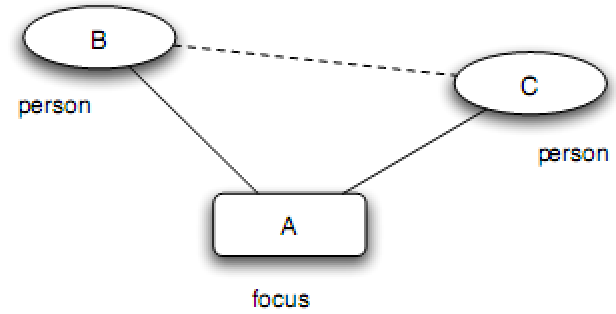
$$T_{baseline}(k) = 1 - (1-p)^{k-1}$$

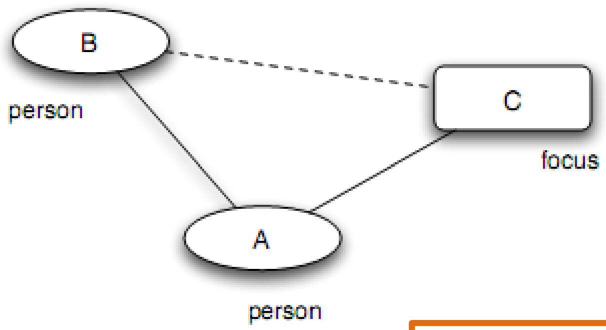


Qualitative similar (linear), but independent assumption too simple

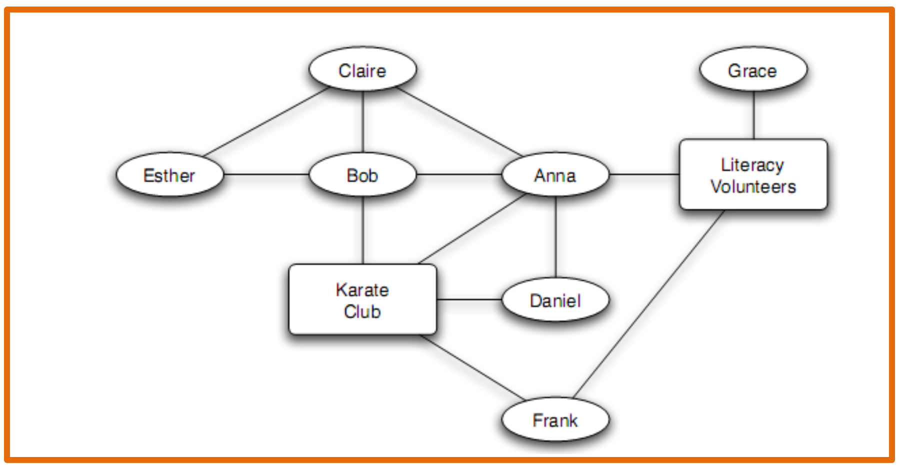
Tracking Link Formation in Online Data: focal and membership closure

Focal closure: what is the probability that two people form a link as a function of the *number of foci* that are jointly affiliated with



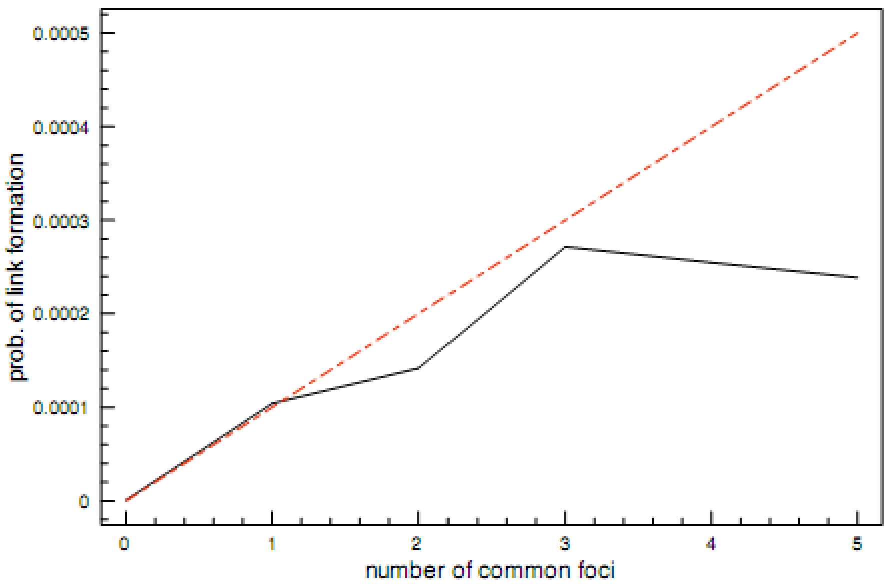


Membership closure: what is the probability that a person becomes involved with a particular focus as a function of the *number of friends* who are already involved in it?



Tracking Link Formation in Online Data: focal closure

E-mail ("who-talks-to-whom" dataset type)
Use the *class schedule* of each student
Focus: class (common focus – a class together)

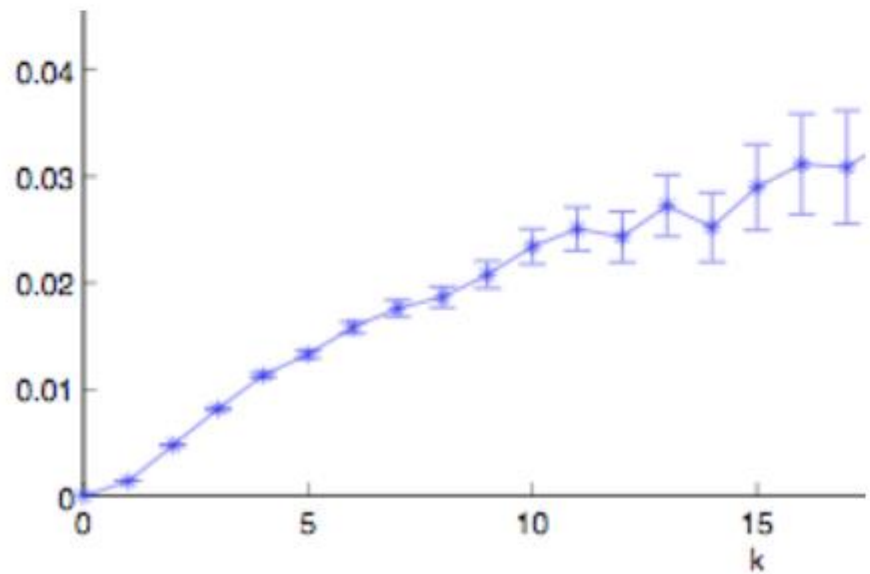


A single shared class same effect as a single shared friend, then different
Subsequent shared classes after the first produce a diminishing returns effect

Tracking Link Formation in Online Data: membership closure

Node: Wikipedia editor who maintains a user account and user talk page
Link: if they have communicated by one user writing on the user talk page of the other
Focus: Wikipedia article
Association to focus: edited the article

Again, an initial increasing effect: the probability of editing a Wikipedia article is more than twice as large when you have two connections into the focus than one



✓ Also, multiple effects can operate simultaneously

A Spatial Model of Segregation

- Formation of ethnically and racially homogeneous neighbors in cities

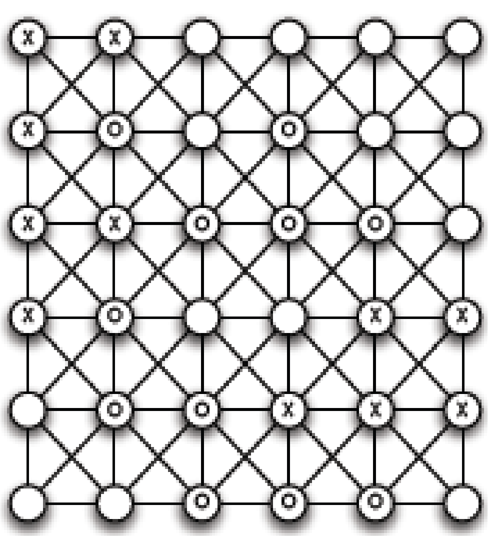
(a) *Chicago, 1940* (b) *Chicago, 1960*

A Spatial Model of Segregation: The Schelling Model

Simple model at a local level

- Population of individuals called agents, of type X or type O
- Agents reside in cells of a grid
- Neighbor cells that touch it (including diagonal)

Possible to show as a graph, but use geometric grid



x	x				
x	o		o		
x	x	o	o	o	
x	o			x	x
	o	o	x	x	x
		o	o	o	

A Spatial Model of Segregation: The Schelling Model

Threshold *t*: Each agent wants to have at least *t* agents of its own type as neighbors

If an agent discovers that fewer than *t* of its neighbors are of the same type of itself, then it has an interest to move to a new cell

Unsatisfied (shown with *)

t = 3

x	x				
x	o		o		
x	x	o	o	o	
x	o			x	x
	o	o	x	x	x
		o	o	o	

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

A Spatial Model of Segregation: The Schelling Model

Agents move in sequence of *rounds*:
In each round, consider unsatisfied agents at some order, move to an unoccupied cell where it will be satisfied

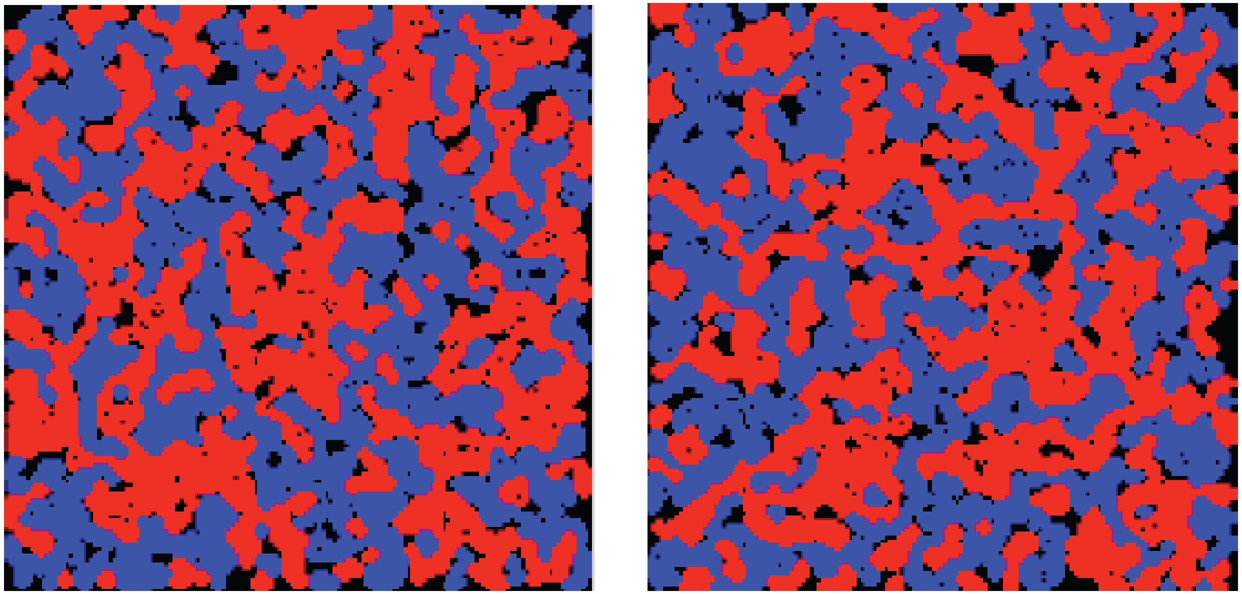
How to move? (in a random order, downwards?) Where to move? what if no satisfying position?

X1*	X2*				
X3	O1*		O2		
X4	X5	O3	O4	O5*	
X6*	O6			X7	X8
	O7	O8	X9*	X10	X11
		O9	O10	O11*	

X3	X6	O1	O2		
X4	X5	O3	O4		
	O6	X2	X1	X7	X8
O11	O7	O8	X9	X10	X11
	O5	O9	O10*		

t = 3. one row at a time working downwards, agent moves to the nearest cell that will make it satisfied

A Spatial Model of Segregation: The Schelling Model



(a) *A simulation with threshold 3.* (b) *Another simulation with threshold 3.*

Figure 4.17: Two runs of a simulation of the Schelling model with a threshold t of 3, on a 150-by-150 grid with 10,000 agents of each type. Each cell of the grid is colored red if it is occupied by an agent of the first type, blue if it is occupied by an agent of the second type, and black if it is empty (not occupied by any agent).

- Simulation, unsatisfied agents move to a random location
- ~50 rounds, all satisfied
- Different random starts
- Large homogeneous regions

A Spatial Model of Segregation: The Schelling Model

Spatial segregation is taking place even if no individual agent is actively seeking it

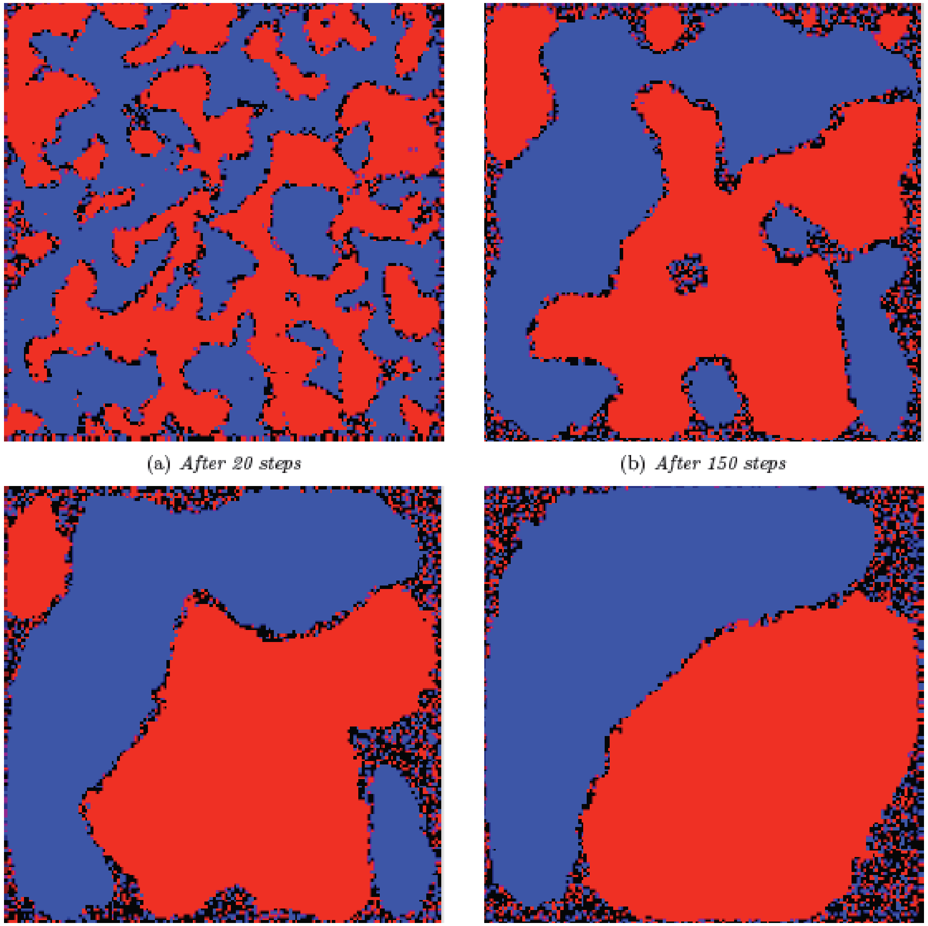
For t = 3, satisfied even in the minority among its neighbors

Requirements not globally incompatible

X	X	O	O	X	X
X	X	O	O	X	X
O	O	X	X	O	O
O	O	X	X	O	O
X	X	O	O	X	X
X	X	O	O	X	X

If we start from a random configuration, attach to clusters, grow, some fall below, move, "unraveling" of more integrated regions

A Spatial Model of Segregation: The Schelling Model



(a) *After 20 steps*
(b) *After 150 steps*
(c) *After 350 steps*
(d) *After 800 steps*

Figure 4.19: Four intermediate points in a simulation of the Schelling model with a threshold t of 4, on a 150-by-150 grid with 10,000 agents of each type. As the rounds of movement progress, large homogeneous regions on the grid grow at the expense of smaller, narrower regions.

End of Chapter 4

Homophily (selection vs social influence)

Graphs with more than one type of nodes (bipartite)

- Affiliation networks

Spatial model of segregation

Chapter 5, from D. Easley and J. Kleinberg book

NETWORK WITH POSITIVE AND NEGATIVE TIES

Structural Balance

What about negative edges?

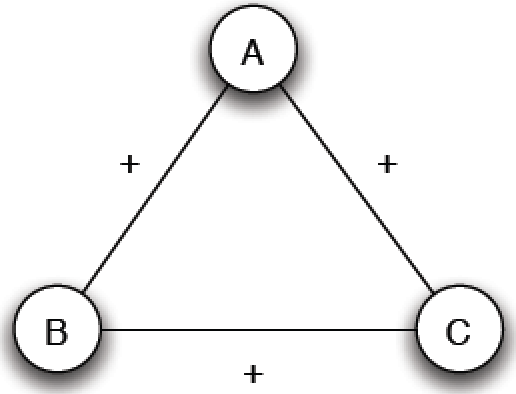
Initially, a complete graph (or clique): every edge either + or -

Let us first look at individual triangles

- Lets look at 3 people => 4 Cases
- See if all are equally possible (local property)

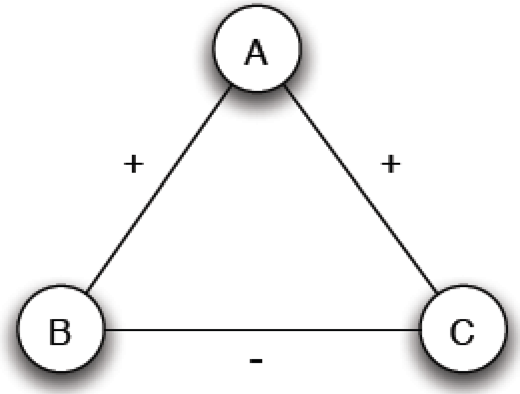
Structural Balance

Case (a): 3 +



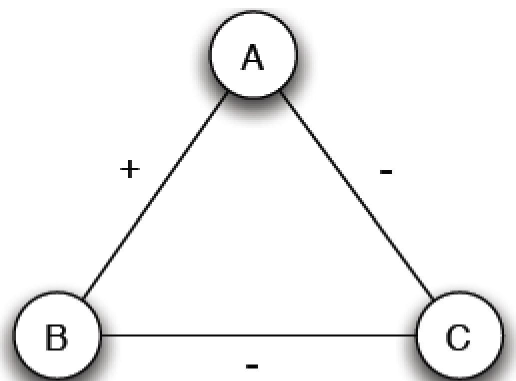
Mutual friends

Case (b): 2 +, 1 -



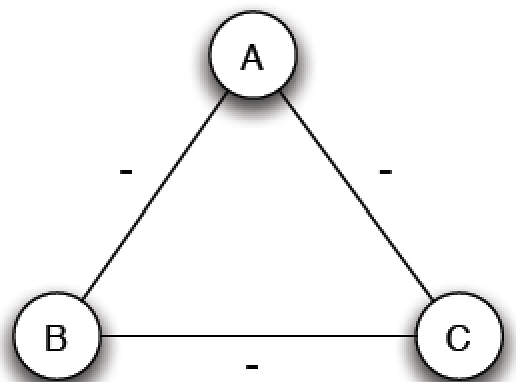
A is friend with B and C, but B and C do not get well together

Case (c): 1 +, 2 -



A and B are friends with a mutual enemy

Case (d): 3 -

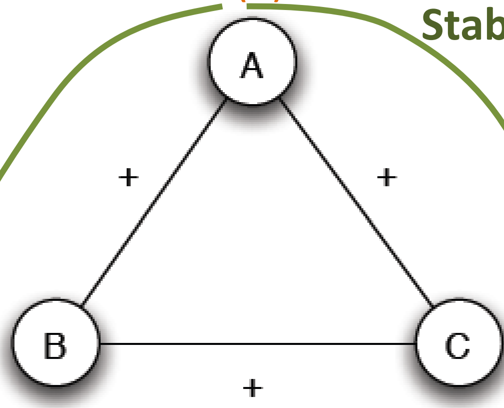


Mutual enemies

Structural Balance

Case (a): 3 +

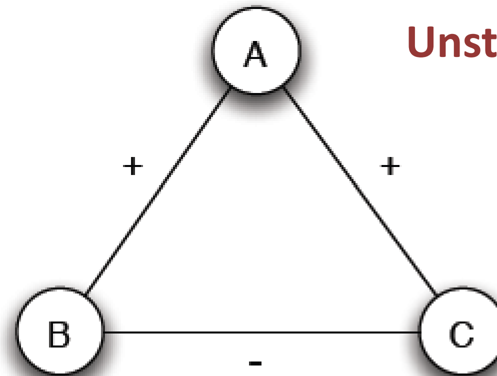
Stable or balanced



Mutual friends

Case (b): 2 +, 1 -

Unstable

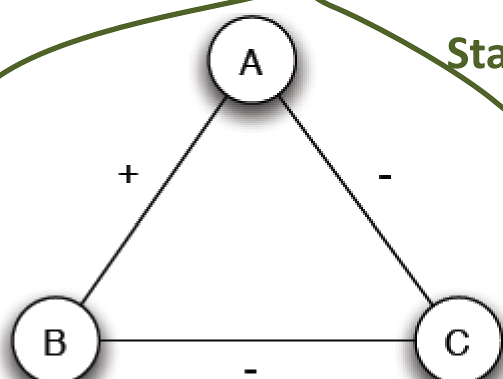


A is friend with B and C, but B and C do not get well together

Implicit force to make B and C friends (- => +) or turn one of the + to -

Case (c): 1 +, 2 -

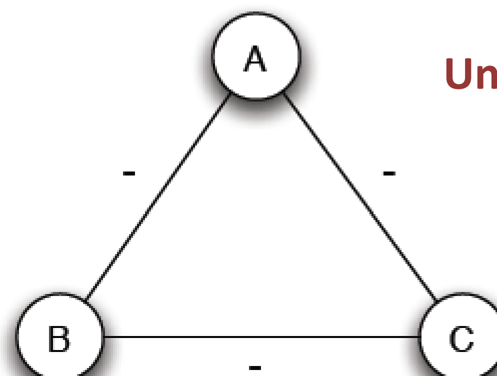
Stable or balanced



A and B are friends with a mutual enemy

Case (d): 3 -

Unstable



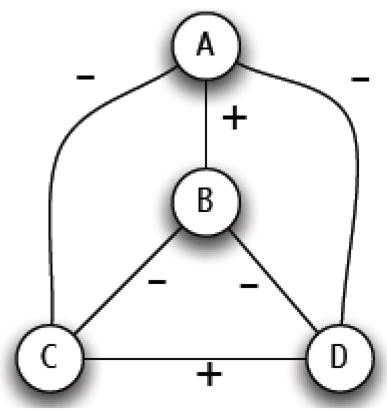
Mutual enemies

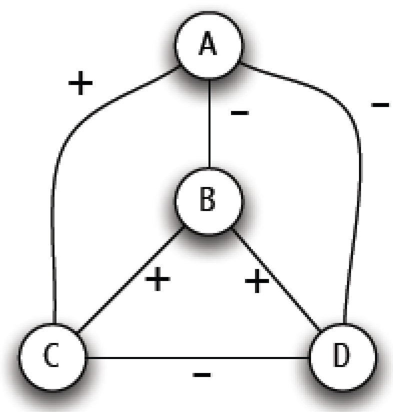
Forces to team up against the third (turn 1 – to +)

Structural Balance

A labeled complete graph is balanced if every one of its triangles is balanced

> ***Structural Balance Property:*** For every set of three nodes, if we consider the three edges connecting them, either all three of these are labeled +, or else exactly one of them is labeled – (odd number of +)



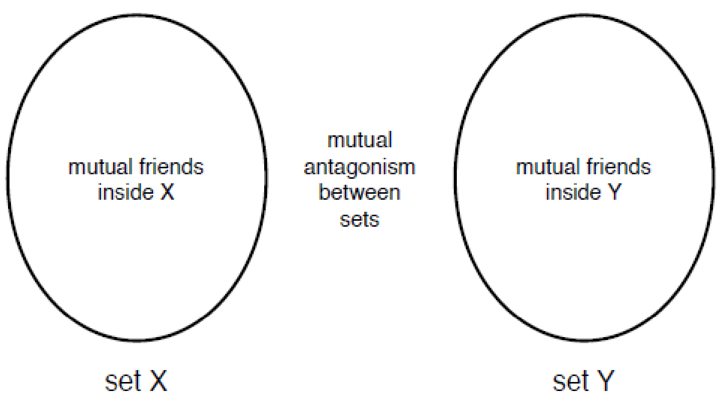


What does a balanced network look like?

The Structure of Balanced Networks

Balance Theorem: If a labeled complete graph is balanced,

(a) all pairs of nodes are friends, or
(b) the nodes can be divided into two groups X and Y, such that every pair of nodes in X like each other, every pair of nodes in Y like each other, and every one in X is the enemy of every one in Y.



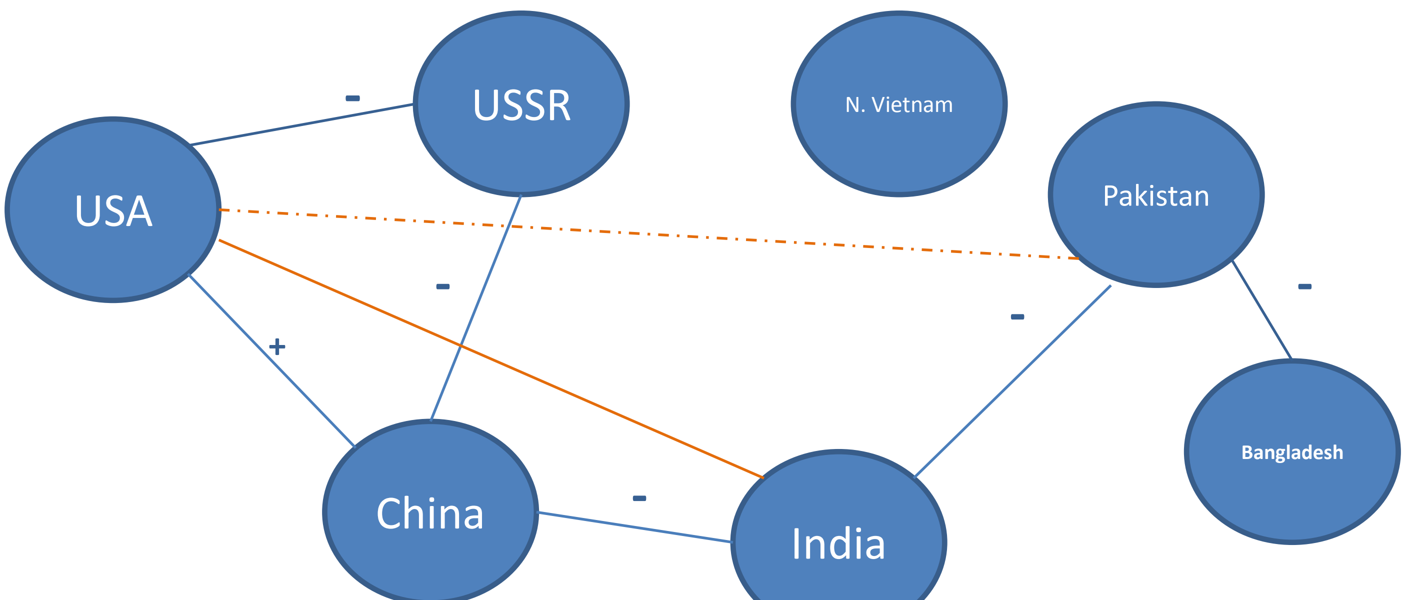
Proof ...

From a local to a global property

Applications of Structural Balance

- ✓ How a network evolves over time
- ✓ Political science: International relationships (I)

The conflict of Bangladesh's separation from Pakistan in 1972 (1)

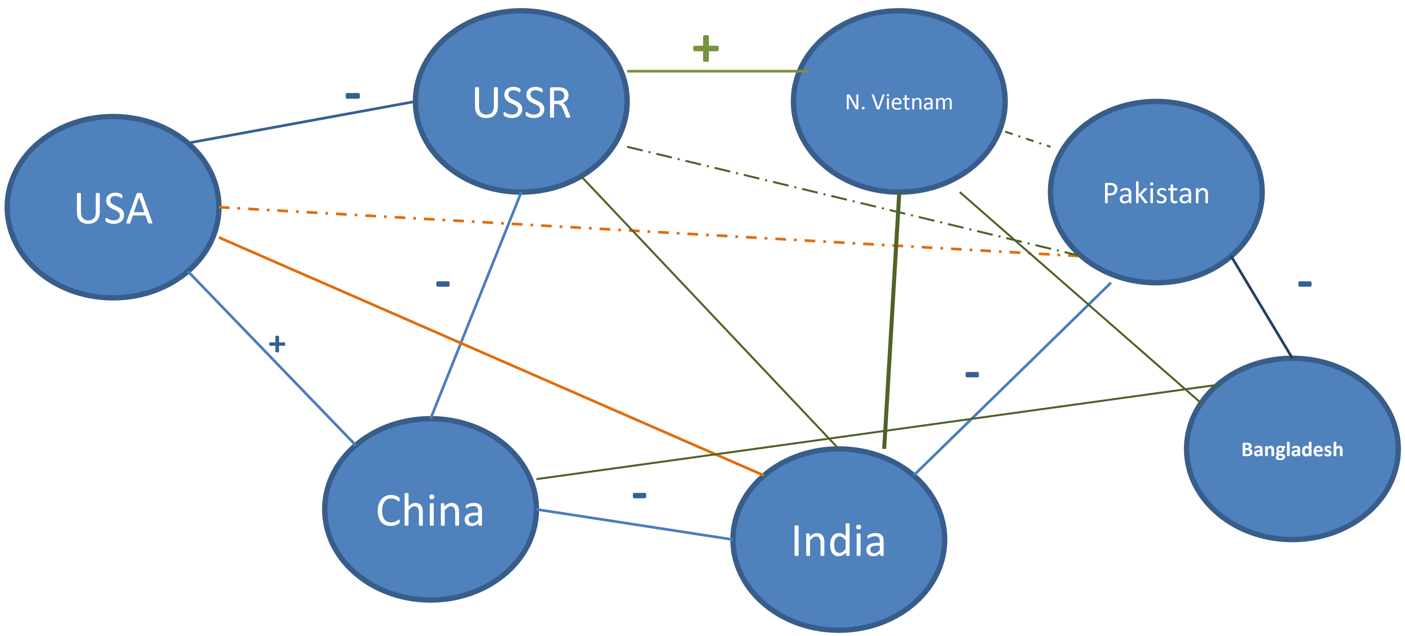


USA support to Pakistan?

Applications of Structural Balance

✓International relationships (I)

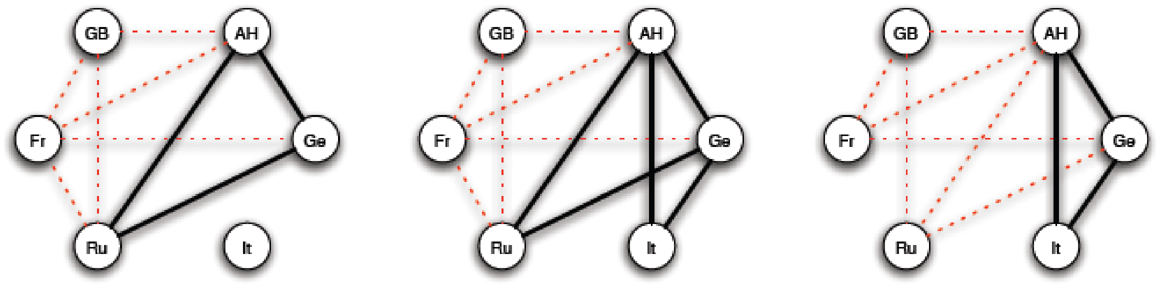
The conflict of Bangladesh’s separation from Pakistan in 1972 (II)



China?

Applications of Structural Balance

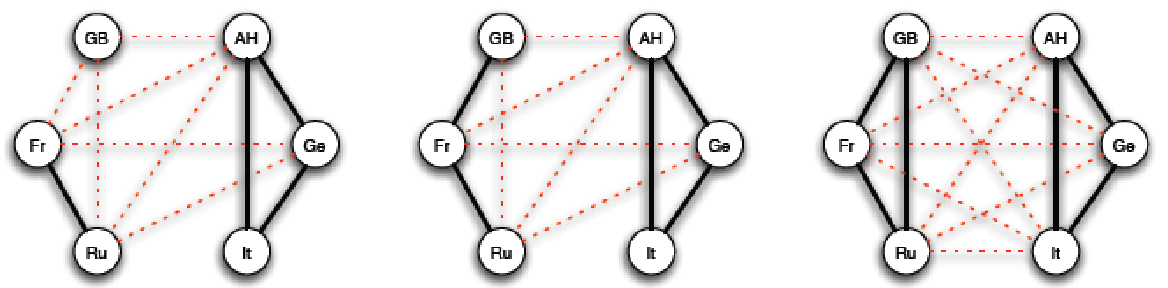
✓ International relationships (II)



(a) *Three Emperors' League 1872–81*

(b) *Triple Alliance 1882*

(c) *German-Russian Lapse 1890*



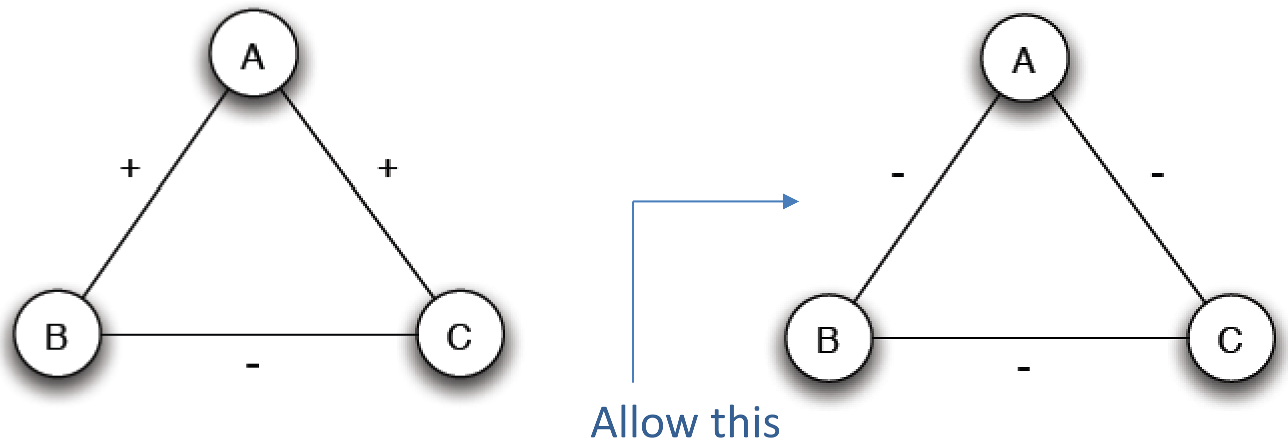
(d) *French-Russian Alliance 1891–94*

(e) *Entente Cordiale 1904*

(f) *British Russian Alliance 1907*

Figure 5.5: The evolution of alliances in Europe, 1872-1907 (the nations GB, Fr, Ru, It, Ge, and AH are Great Britain, France, Russia, Italy, Germany, and Austria-Hungary respectively). Solid dark edges indicate friendship while dotted red edges indicate enmity. Note how the network slides into a balanced labeling — and into World War I. This figure and example are from Antal, Krapivsky, and Redner [20].

A Weaker Form of Structural Balance



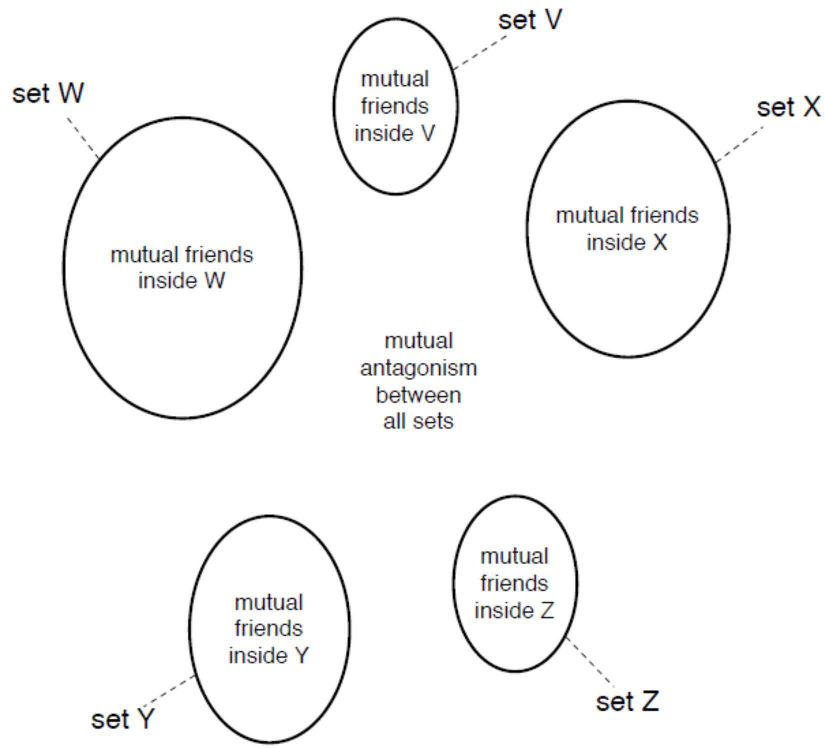
Weak Structural Balance Property: There is no set of three nodes such that the edges among them consist of exactly two positive edges and one negative edge

A Weaker Form of Structural Balance

Weakly Balance Theorem: If a labeled complete graph is weakly balanced, its nodes can be divided into groups in such a way that every two nodes belonging to the same group are friends, and every two nodes belonging to different groups are enemies.

Proof …

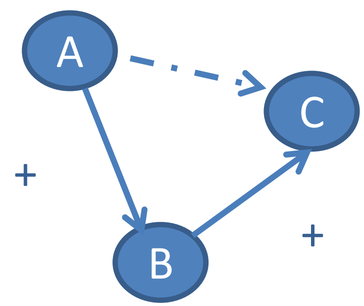
A Weaker Form of Structural Balance



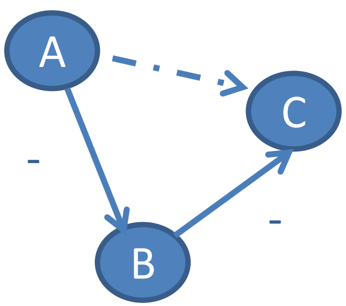
Trust, distrust and online ratings

Evaluation of products and trust/distrust of other users

Directed Graphs



A trusts B, B trusts C, A ? C



A distrusts B, B distrusts C, A ? C
If distrust enemy relation, +
A distrusts means that A is better than B, -

Depends on the application
Rating political books or
Consumer rating electronics products

Generalizing

1. Non-complete graphs

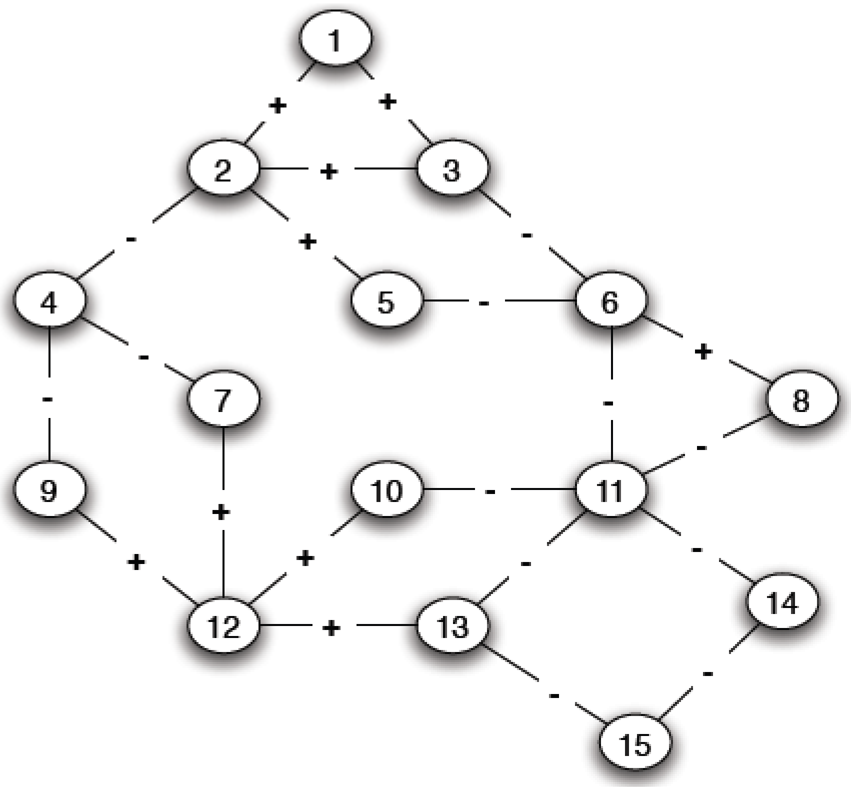
2. Instead of all triangles, “most” triangles, approximately divide the graph

We shall use the original (“non-weak” definition of structural balance)

Structural Balance in Arbitrary Graphs

Thee possible relations

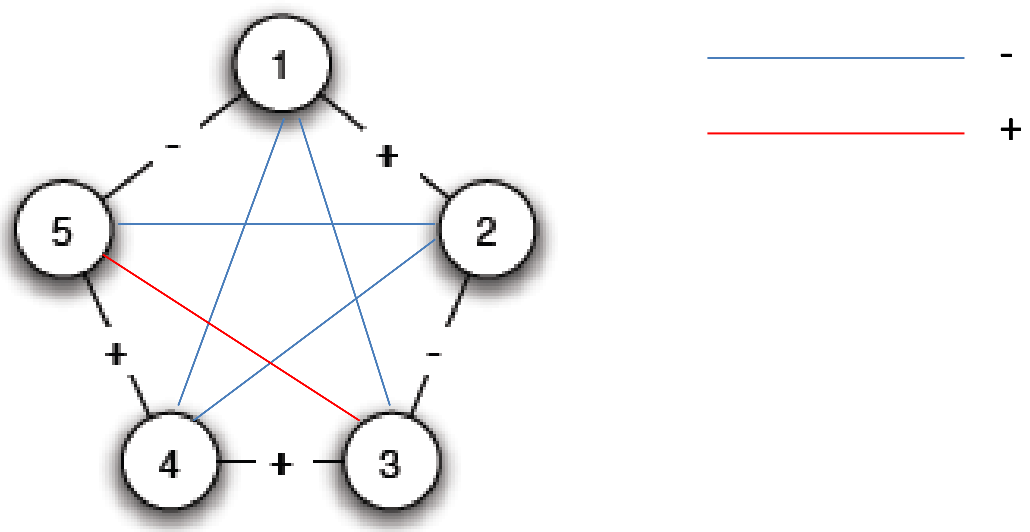
- Positive edge
- Negative edge
- Absence of an edge



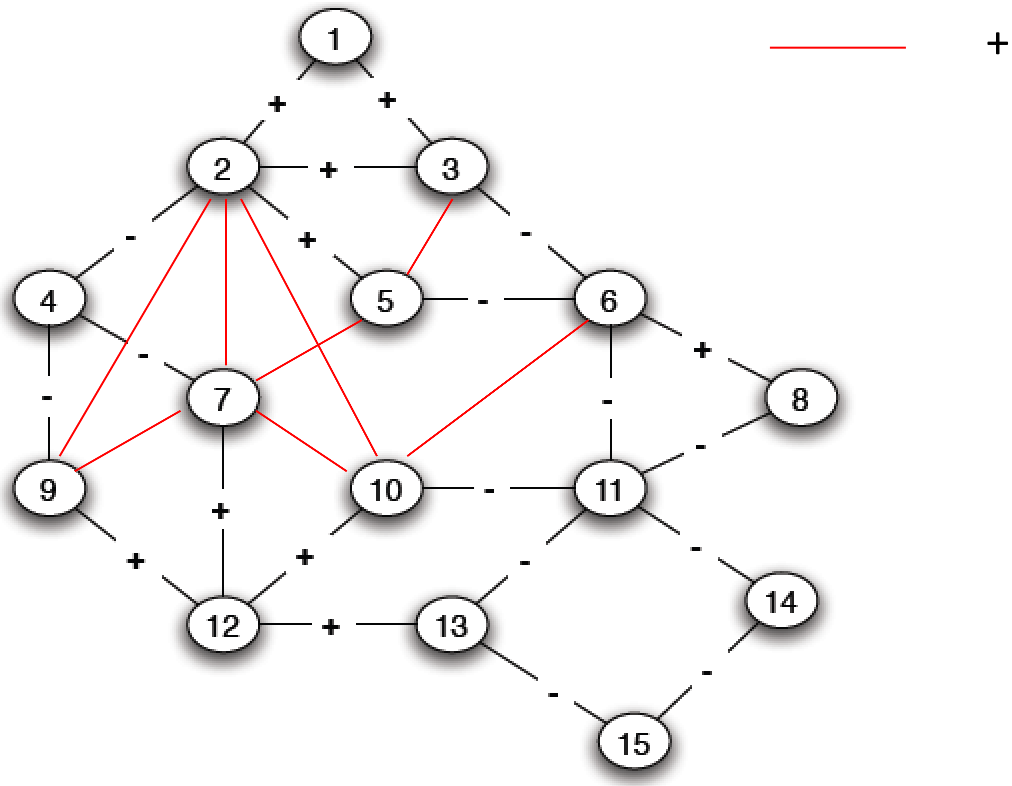
Balance Definition for General Graphs

1. Based on triangles (local view)
2. Division of the network (global view)

A (non-complete) graph is balanced if it can be completed by adding edges to form a signed complete graph that is balanced



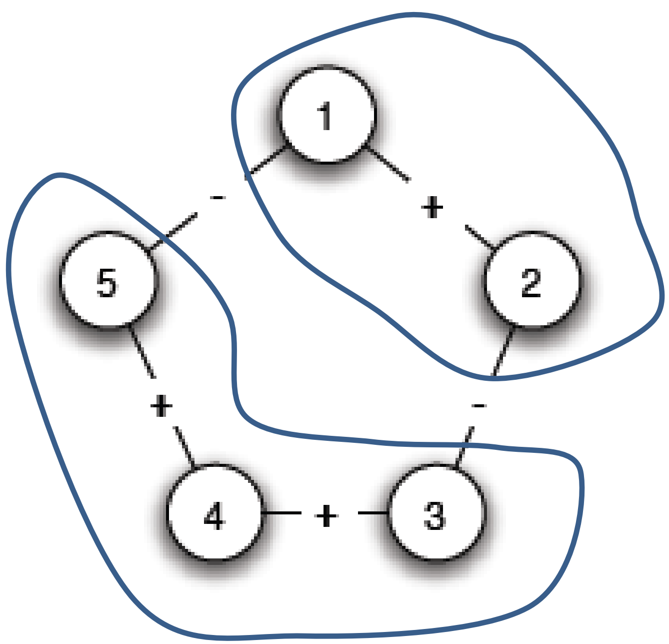
Balance Definition for General Graphs



Balance Definition for General Graphs

1. Based on triangles (local view)
2. Division of the network (global view)

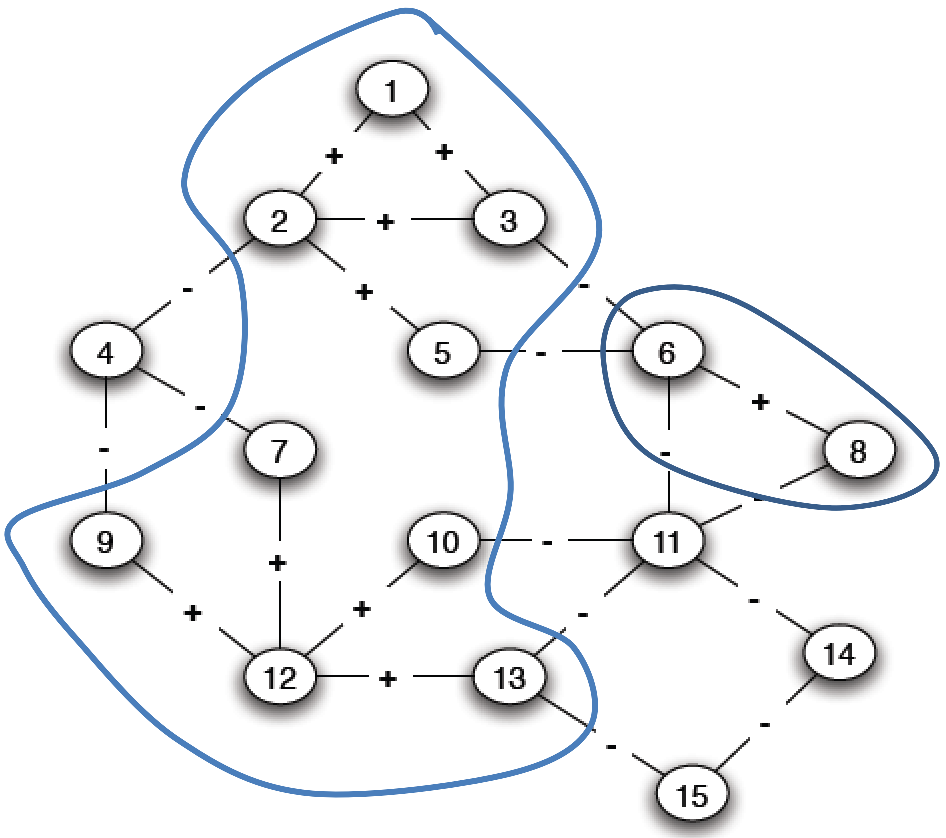
A (non-complete) graph is balanced if it possible to divide the nodes into two sets X and Y, such that any edge with both ends inside X or both ends inside Y is positive and any edge with one end in X and one end in Y is negative



The **two definition** are **equivalent**:
An arbitrary signed graph is balanced under the first definition, if and only if, it is balanced under the second definitions

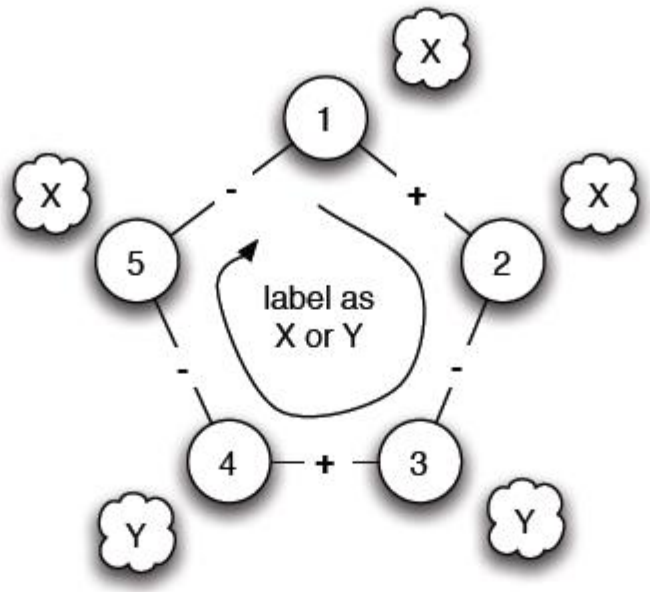
Balance Definition for General Graphs

Algorithm for dividing the nodes?



Balance Characterization

What prevents a network from being balanced?



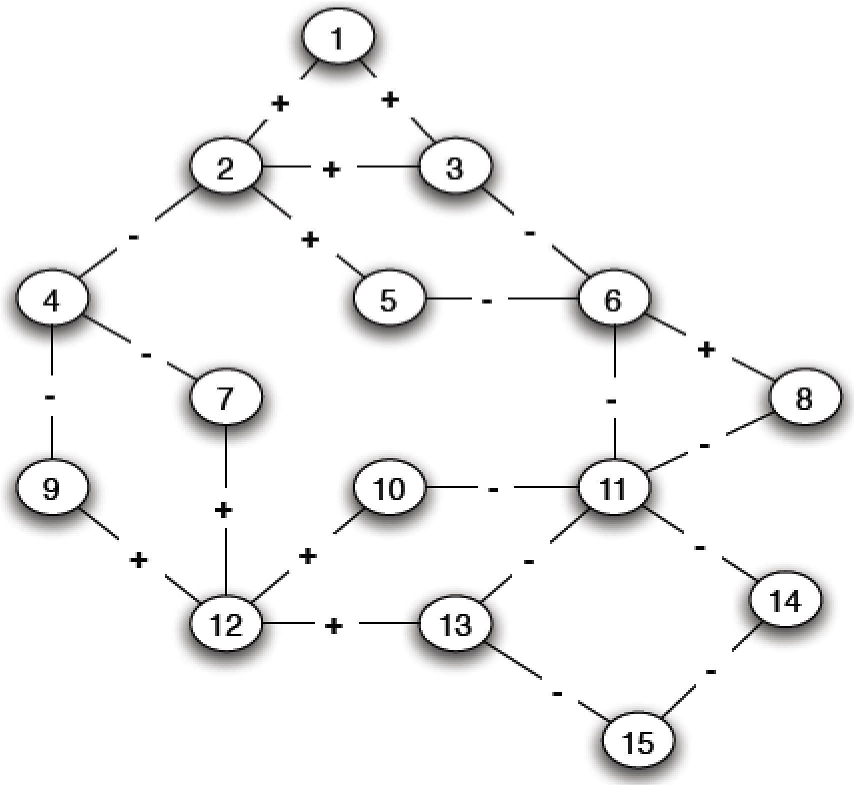
- Start from a node and place nodes in X or Y
- Every time we cross a negative edge, change the set

Cycle with odd number of negative edges

Balance Definition for General Graphs

Cycle with odd number of - => unbalanced

Is there such a cycle with an odd number of -?



Balance Characterization

Claim: A signed graph is balanced, if and only if, it contains no cycles with an odd number of negative edges

(proof by construction)

Find *a balanced division:* partition into sets X and Y, all edges inside X and Y positive, crossing edges negative

Either succeeds or Stops with a cycle containing an odd number of -

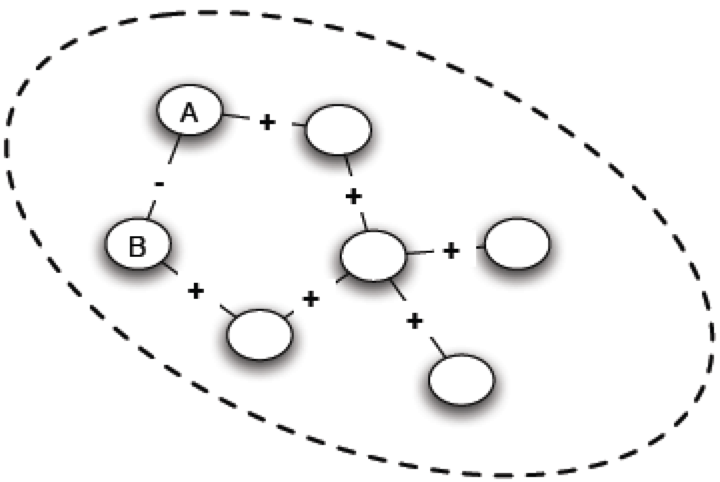
Two steps:

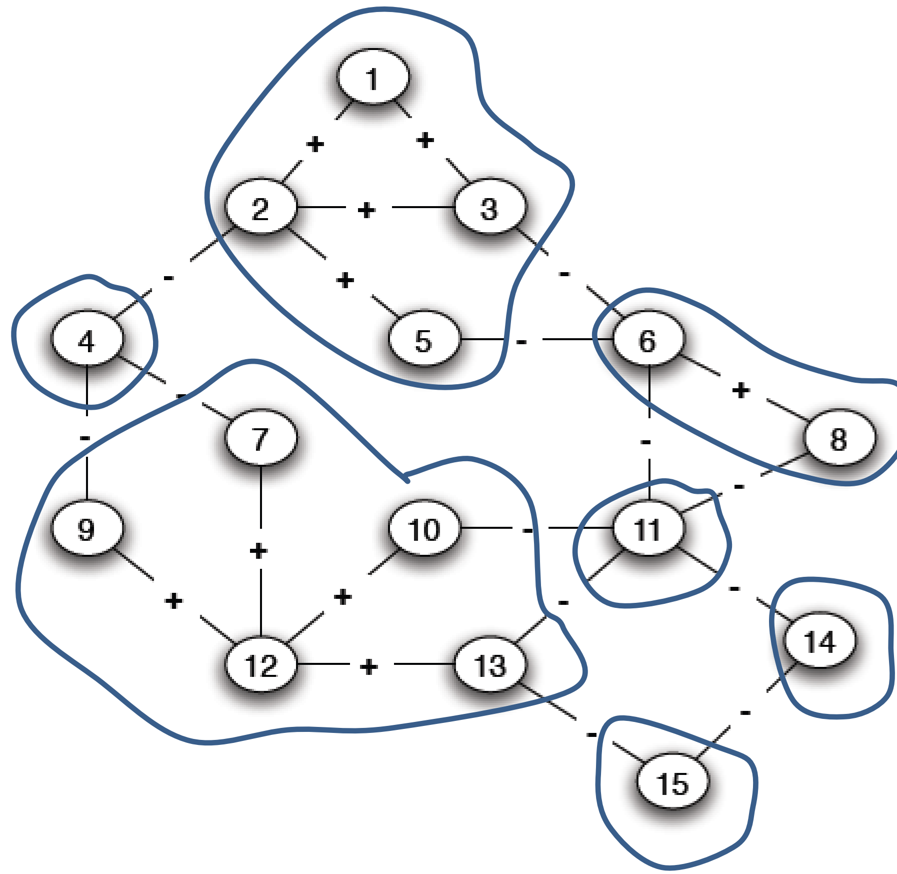
1. Convert the graph into a reduced one with only negative edges
2. Solve the problem in the reduced graph

Balance Characterization: Step 1

1. Find *connected components* (supernodes) by considering only positive edges

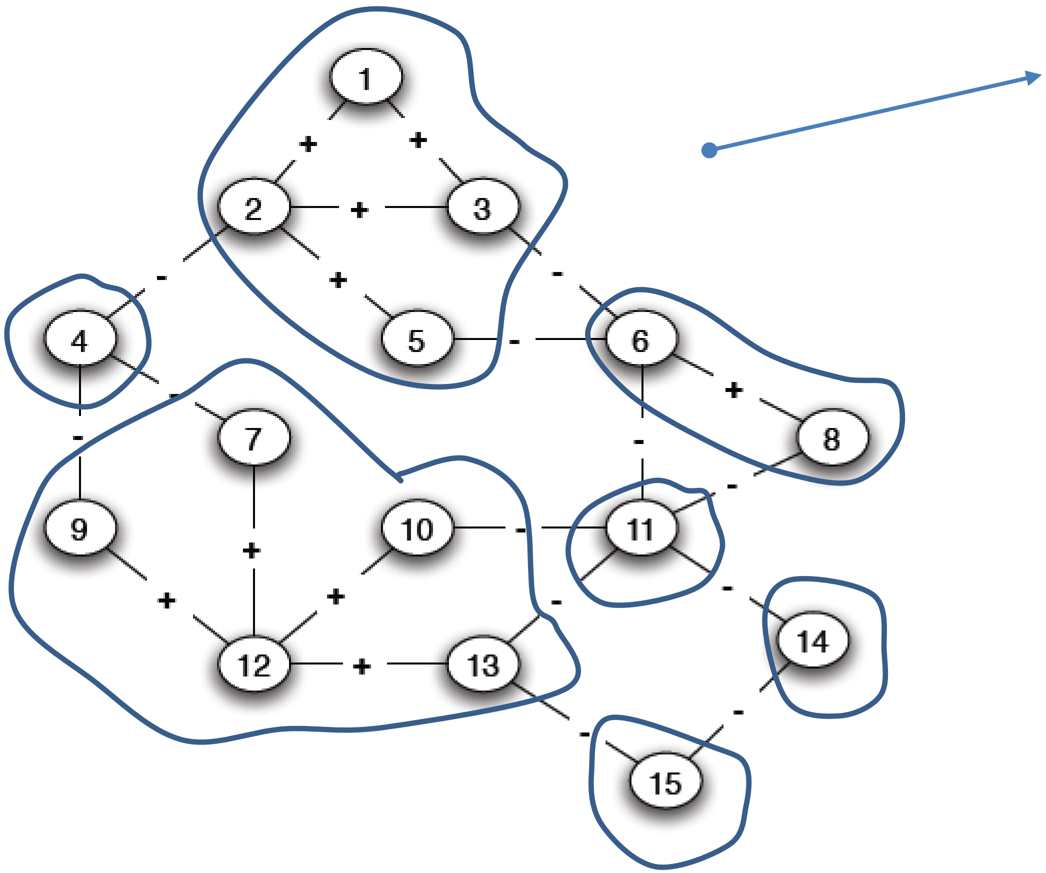
2. Check: Do supernodes contain a negative edge between any pair of their nodes
(a) Yes -> odd cycle (1)
(b) No -> each supernode either X or Y

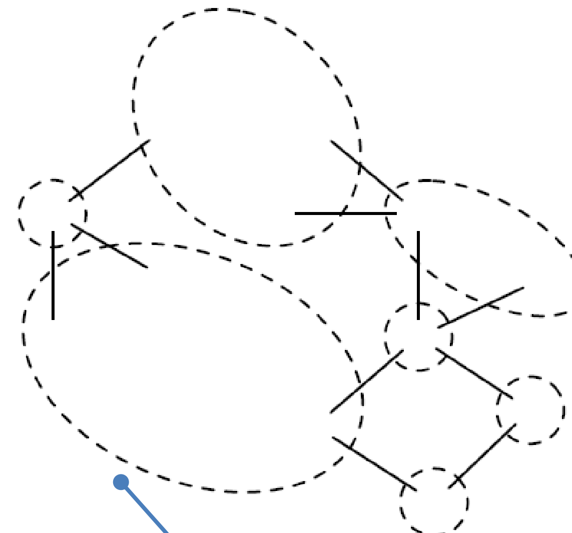


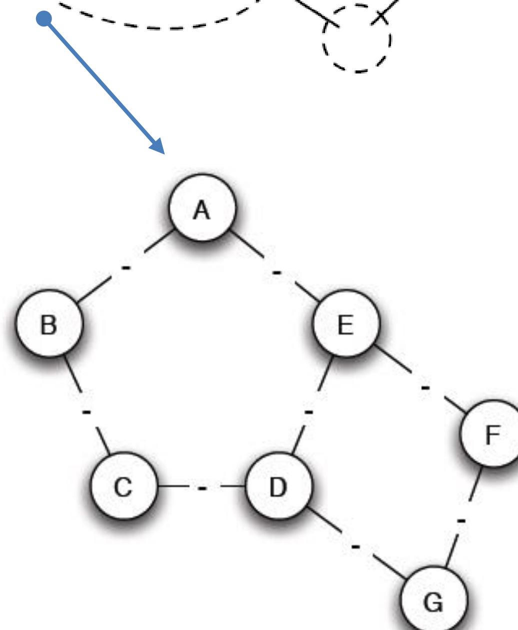


Balance Characterization: Step 1

3. Reduced problem: a node for each supernode, an edge between two supernodes if an edge in the original







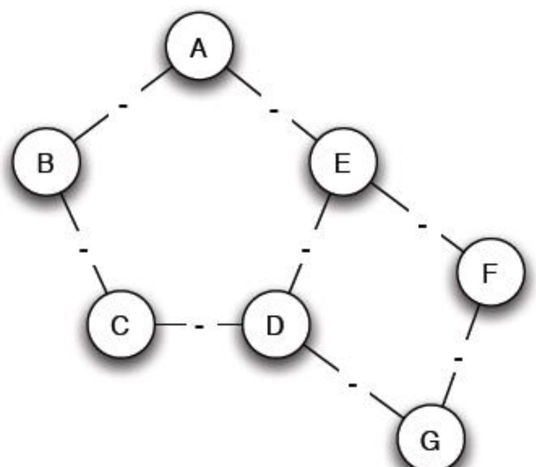
Balance Characterization: Step 2

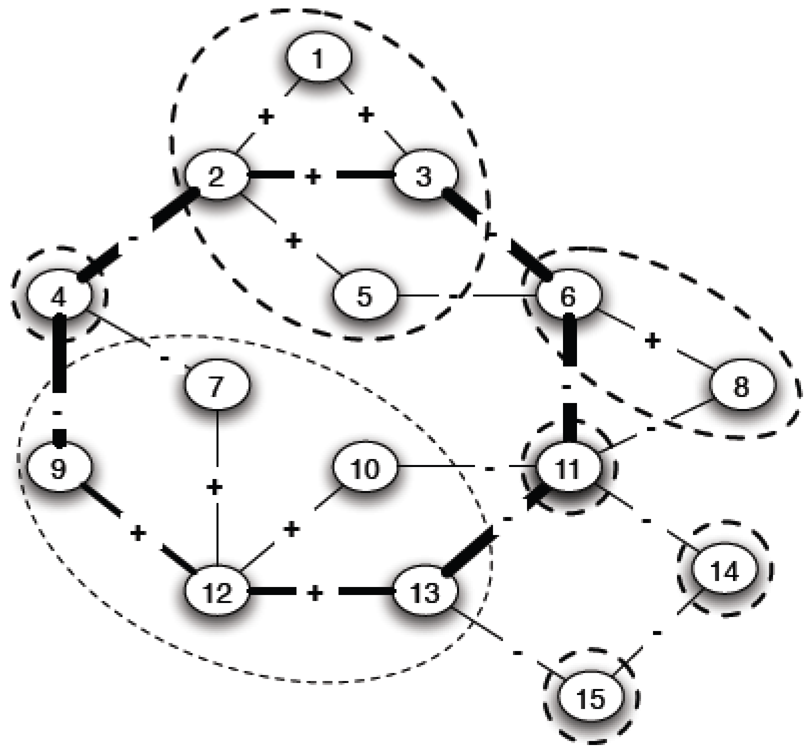
Note: Only negative edges among supernodes

Start labeling by either X and Y

If successful, then label the nodes of the supernode correspondingly

✓ A cycle with an odd number, corresponds to a (possibly larger) odd cycle in the original





Balance Characterization: Step 2

Determining whether the graph is bipartite (there is no edge between nodes in X or Y, *the only edges are from nodes in X to nodes in Y*)

Use Breadth-First-Search (BFS)

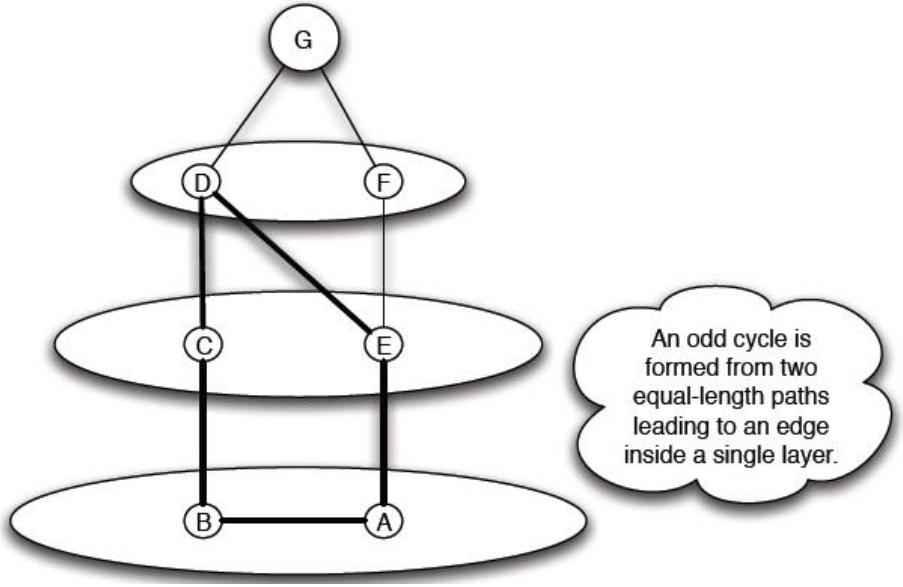
- Start the search at any node and give alternating labels to the vertices visited during the search. *That is, give label X to the starting node, Y to all its neighbors, X to those neighbors' neighbors, and so on.*

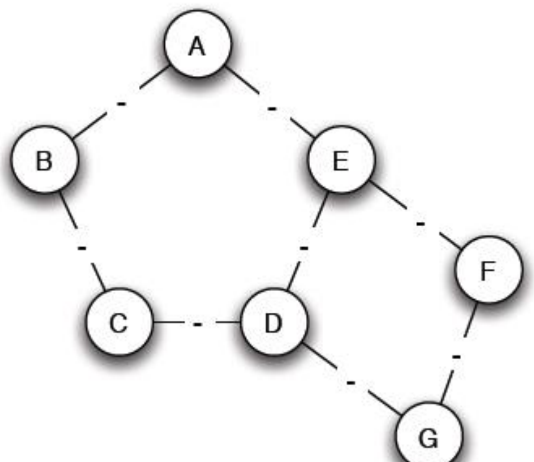
- If at any step a node has (visited) neighbors with the same label as itself, then the graph is not bipartite *(cross-level edge)*

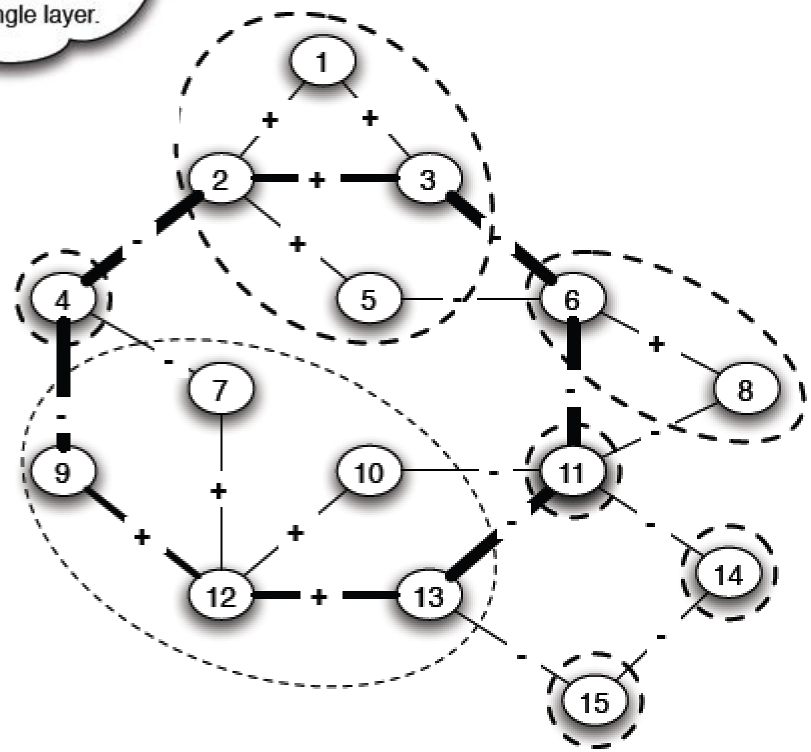
- If the search ends without such a situation occurring, then the graph is bipartite.

Why is this an "odd" cycle?

Balance Characterization







Generalizing

1. Non-complete graphs

2. Instead of all triangles, “most” triangles, approximately divide the graph

Approximately Balance Networks

a complete graph (or clique): every edge either + or -

Claim: If all triangles in a labeled complete graph are balanced, than either
(a) all pairs of nodes are friends or,
(b) the nodes can be divided into two groups X and Y, such that
  (i) every pair of nodes in X like each other,
  (ii) every pair of nodes in Y like each other, and
  (iii) every one in X is the enemy of every one in Y.

Not all, but most, triangles are balanced

Claim: If *at least 99.9%* of all triangles in a labeled compete graph are balanced, then either,
(a) There is a set consisting of *at least 90%* of the nodes in which *at least 90%* of all pairs are friends, or,
(b) the nodes can be divided into two groups X and Y, such that
  (i) *at least 90%* of the pairs in X like each other,
  (ii) *at least 90%* of the pairs in Y like each other, and
  (iii) *at least 90%* of the pairs with one end in X and one in Y are enemies

Approximately Balance Networks

Claim: If *at least 99.9%* of all triangles in a labeled complete graph are balanced, then either,

(a) There is a set consisting of *at least 90%* of the nodes in which *at least 90%* of all pairs are friends, or,

(b) the nodes can be divided into two groups X and Y, such that

  (i) *at least 90%* of the pairs in X like each other,

  (ii) *at least 90%* of the pairs in Y like each other, and

  (iii) *at least 90%* of the pairs with one end in X and one in Y are enemies

Claim: Let ε be any number, such that $0 \leq \varepsilon < 1/8$. If *at least* $1 - \varepsilon$ of all triangles in a labeled complete graph are balanced, then either

(a) There is a set consisting of *at least* $1-\delta$ of the nodes in which *at least* $1-\delta$ of all pairs are friends, or,

(b) the nodes can be divided into two groups X and Y, such that

  (i) *at least* $1-\delta$ of the pairs in X like each other,

  (ii) *at least* $1-\delta$ of the pairs in Y like each other, and

  (iii) *at least* $1-\delta$ of the pairs with one end in X and one in Y are enemies

$$\delta = \sqrt[3]{\varepsilon}$$

Approximately Balance Networks

Basic idea – find a "good" node A (s.t., it does not belong to too many unbalanced triangles) to partition into X and Y

Pigeonhole principle: if *n* items are put into *m pigeonholes* with *n* > *m*, then at least one pigeonhole must contain more than one item

Counting argument based on pigeonhole: compute the average value of a set of objects and then argue that there must be at least one node that is equal to the average or below (or equal and above)

End of Chapter 5

Balanced networks in the case of both positive and negative edges