VE444: Networks

Yifei Zhu, assistant professor
University of Michigan-Shanghai Jiao Tong University

JOINT INSTITUTE
交大密西根学院

Homophily

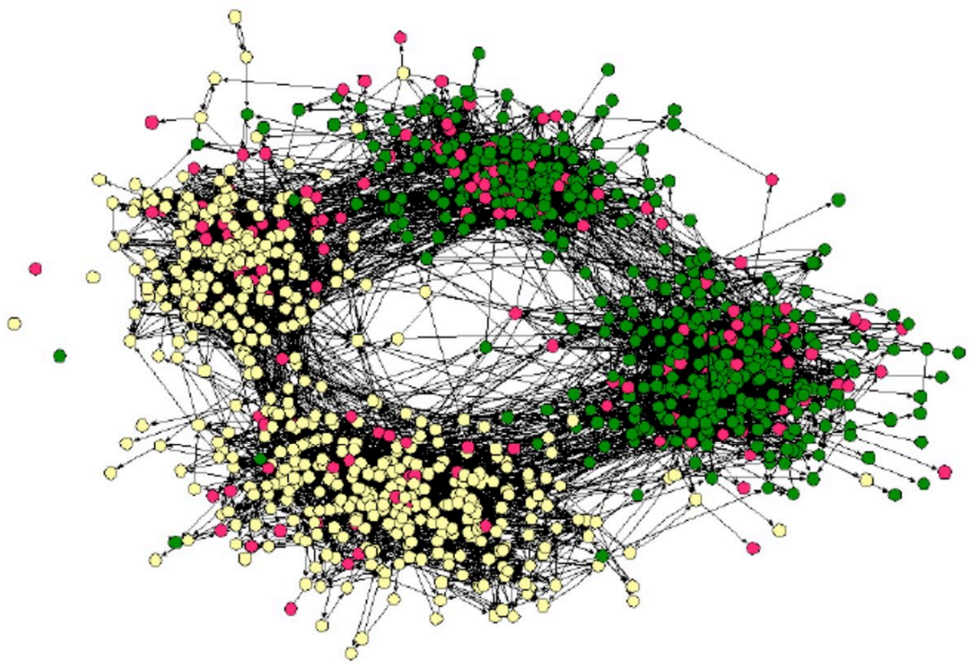
Networks in their surrounding contexts

- **Homophily**: the tendency of individuals to associate and bond with similar others
  - *"Birds of a feather flock together"*
  - "similarity begets friendship" – Plato
  - "love those who are like themselves" -- Aristotle
  - It has been observed in a vast array of network studies, based on a variety of attributes (e.g., age, gender, organizational role, etc.)
  - **Example**: people who like the same music genre are more likely to establish a social connection (meeting at concerts, interacting in music forums, etc.)

Correlations Exists in Networks

Example:

- Real social network
  - Nodes = people
  - Edges = friendship
  - Node color = race
- People are segregated by race due to homophily



(Easley and Kleinberg, 2010)

Homophily and Friendship

- For an individual, he has two types of characteristics
  - Intrinsic: gender, race, mother tongue, etc
  - Changeable: where he lives, expertise, what he likes, etc
- Homophily is the external reason for the creation of social networks
  - Common in race, locations, expertise, interests
- One key question in social sciences
  - Commonality → friendship ? (selection)
  - Friendship → commonality ? (social influence)
  - **Example**: I recommend my “peculiar” musical preferences to my friends, until one of them grows to like my same favorite genres ☺

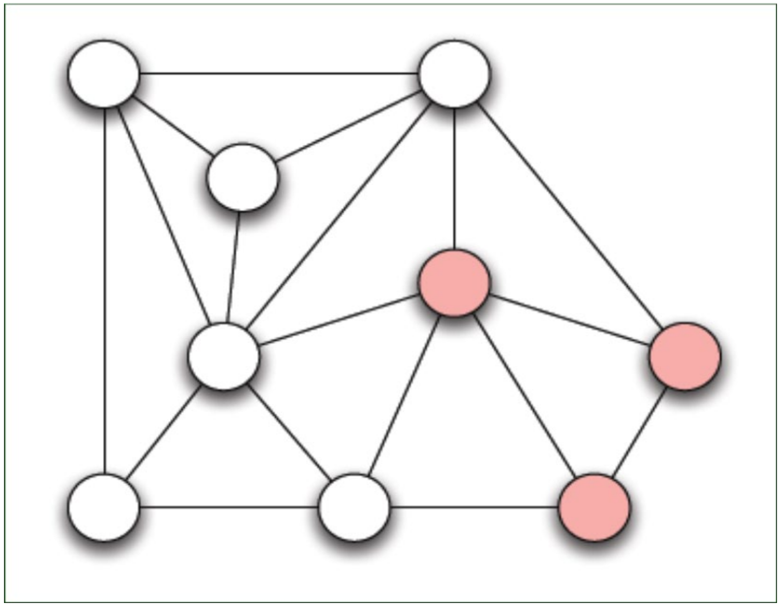
Measuring homophily

- Given a social network where the nodes have only two properties: red and white

- The information we can have:
  - The number of nodes (n), the number of links (e)
  - The ratio of different colors: p, q = 1 – p
  - The number of links (s) where the two end nodes have the same color
- If not homophily?

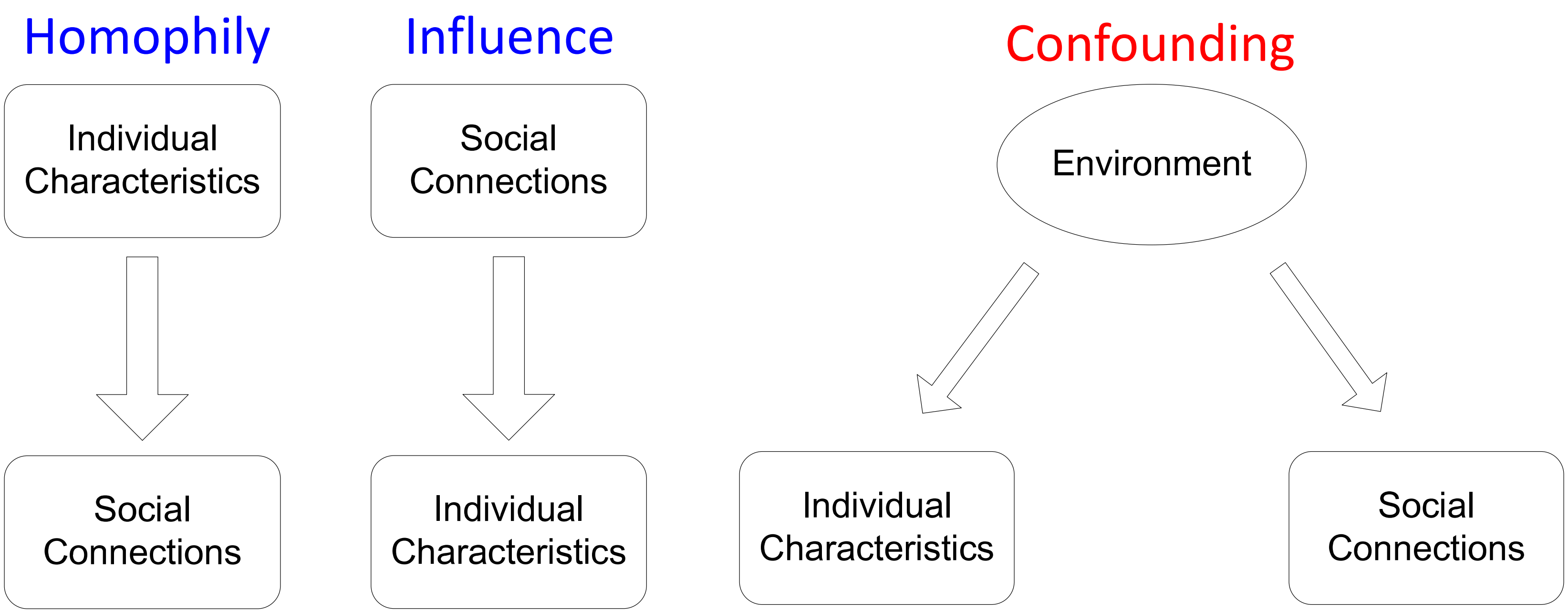
Measuring homophily

- Homophily test: If the fraction of cross-attributes edges is significantly less than 2pq, then there is evidenced for homophily.
- Example:
  - The number of nodes n = 9
  - The number of links e = 18
  - The ratio of red nodes p = 1/3
  - The ratio of white nodes q = 2/3
- Statistical significance test required
- Inverse homophily



Correlations Exist in Networks

- Individual behaviors are correlated in a network environment
- Three main types of dependencies that lead to correlation:



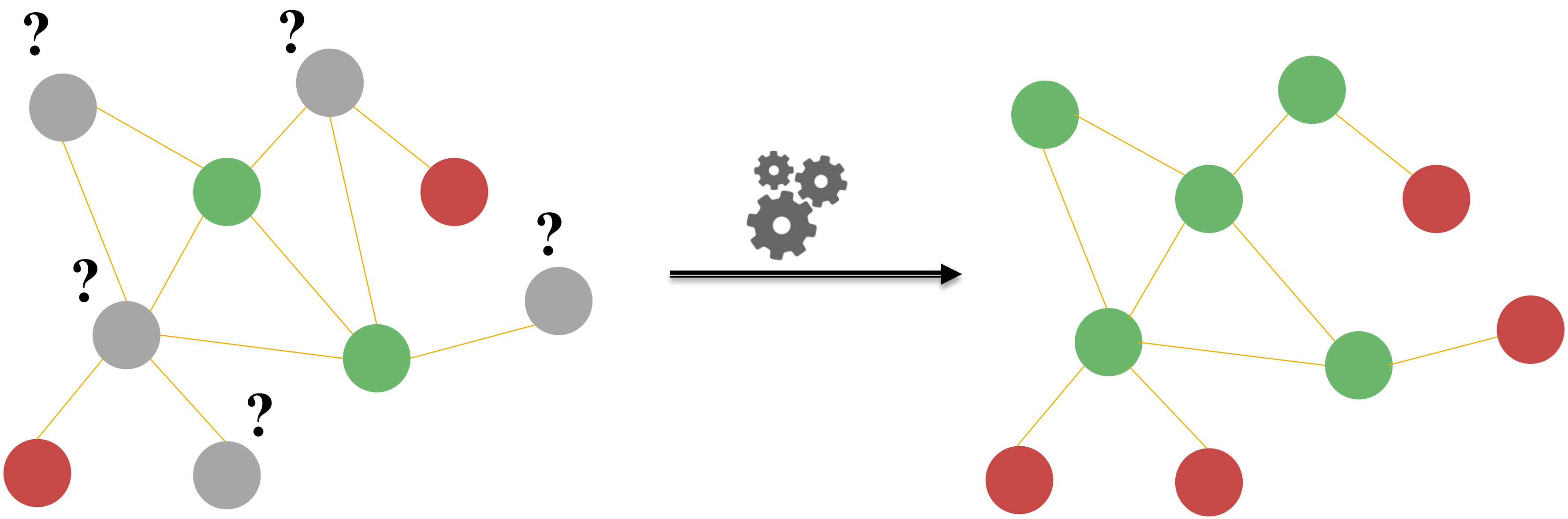
Application of homophily

- How do we leverage this correlation observed in networks to help predict node labels?
- Similar nodes are typically close together or directly connected:
  - "**Guilt-by-association**": If I am connected to a node with label X, then I am likely to have label X as well.
  - Example: Malicious/benign web page: Malicious web pages link to one another to increase visibility, look credible, and rank higher in search engines

Fake Review Spam Detection

- Behavioral analysis
  - individual features, geographic locations, login times, session history, etc.
- Language analysis
  - use of superlatives, lots of self-referencing, rate of misspellings, many agreement words, …
- Easy to fake: **individual behaviors, content of review**
- Hard to fake: **graph structure**
  - Graphs capture relationships between reviewers, reviews, stores

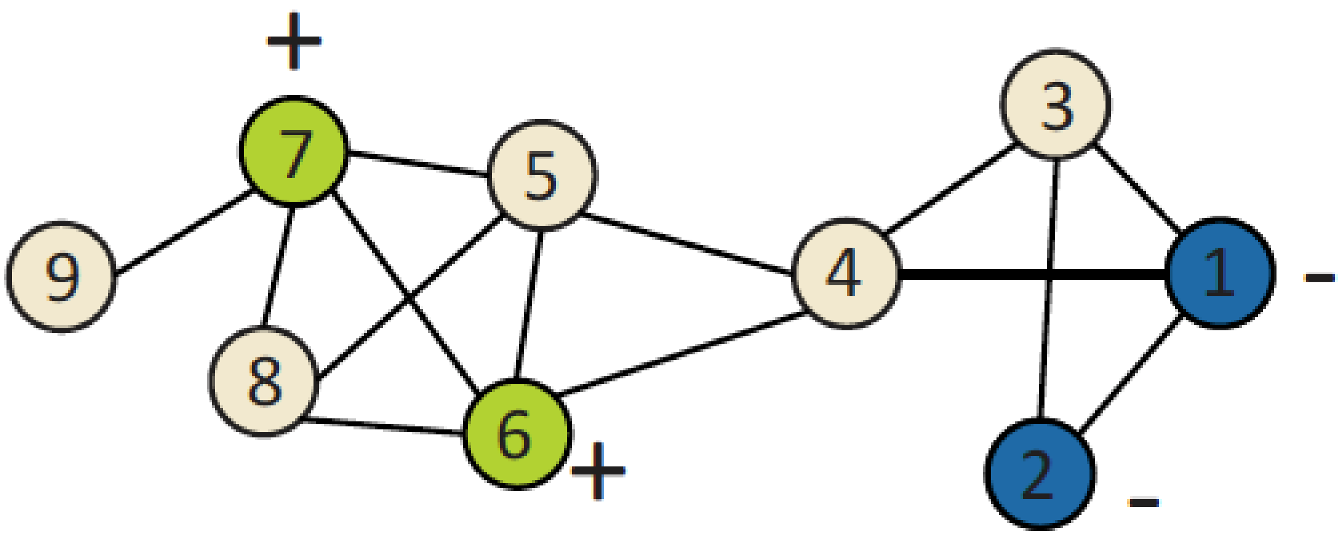
Node Classification



- Given labels of some nodes
- Let's predict labels of unlabeled nodes
- This is called semi-supervised node classification

Classification with Network Data

- How do we leverage this correlation observed in networks to help predict node labels?



How do we predict the labels for the nodes in beige?

Available information

- **Classification label** of an object O in network may depend on:
  - Features of O
  - Labels of the objects in O's neighborhood
  - Features of objects in O's neighborhood

Collective classification overview

- Markov Assumption: *the label Y_i of one node i depends on the labels of its neighbors N_i*

$$P(Y_i|i) = P(Y_i|N_i)$$

- Collective classification involves 3 steps:

Local Classifier	Relational Classifier	Collective Inference
• Assign initial labels	• Capture correlations between nodes	• Propagate correlations through network

Collective Classification: Overview

Local Classifier

- Assign initial labels

Relational Classifier

- Capture correlations between nodes

Collective Inference

- Propagate correlations through network

Local Classifier: Used for initial label assignment

- Predicts label based on node attributes/features
- Standard classification task
- Does not use network information

Relational Classifier: Capture correlations based on the network

- Learns a classier to label one node based on the labels and/or attributes of its neighbors
- This is where network information is used

Collective Inference: Propagate the correlation

- Apply relational classifier to each node iteratively
- Iterate until the inconsistency between neighboring labels is minimized
- Network structure substantially affects the final prediction

Probabilistic Relational Classifier

- **Basic idea:** Class probability of Y_i is a weighted average of class probabilities of its neighbors
- For labeled nodes, initialize with ground-truth Y labels
- For unlabeled nodes, initialize Y uniformly
- Update all nodes in a random order until convergence or until maximum number of iterations is reached

Probabilistic relational classifier

- **Repeat** for each node i and label c

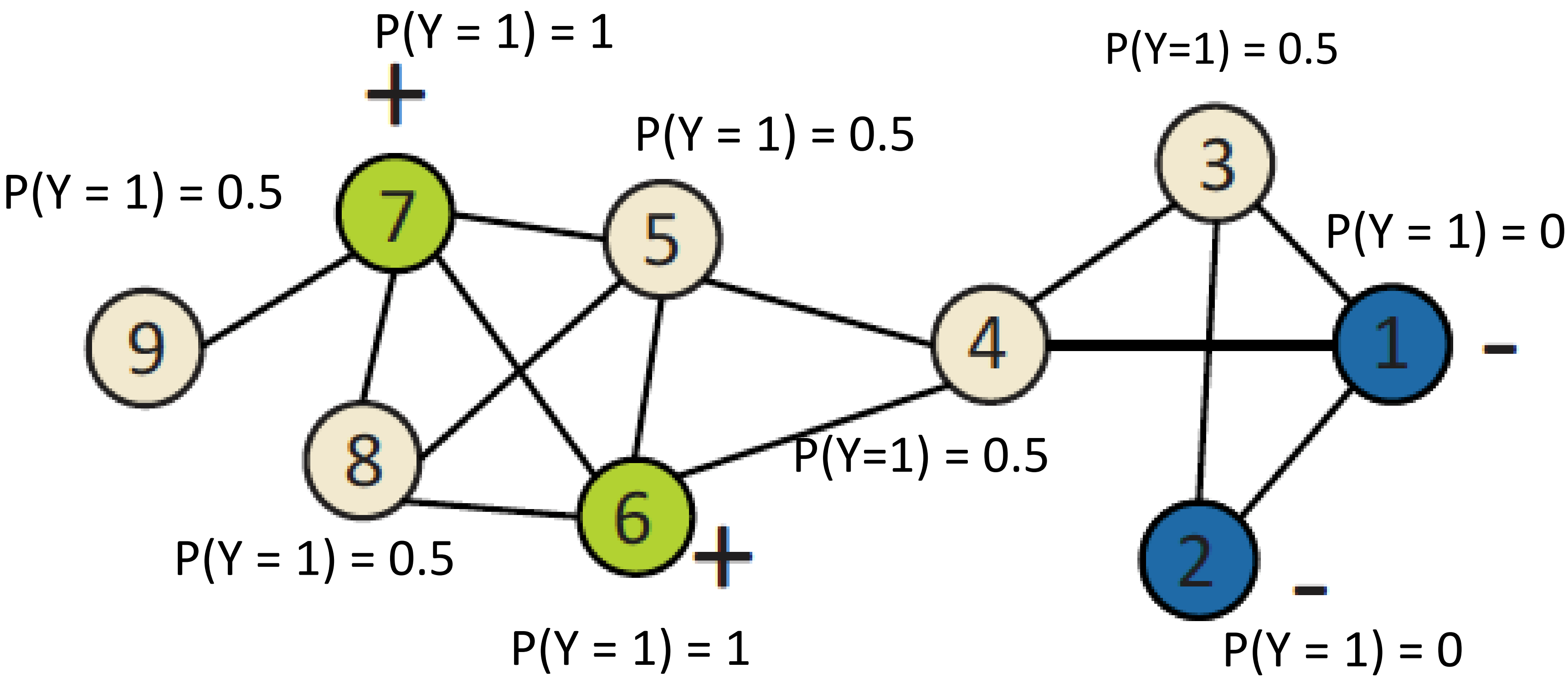
$$P(Y_i = c) = \frac{1}{|N\_i|} \sum_{(i,j)\in E} W(i,j) P(Y_j = c)$$

  - $W(i,j)$ is the edge strength from i to j
  - *N_i is the number of neighbors of i*

- Challenges:
  - Convergence is not guaranteed
  - Model cannot use node feature information

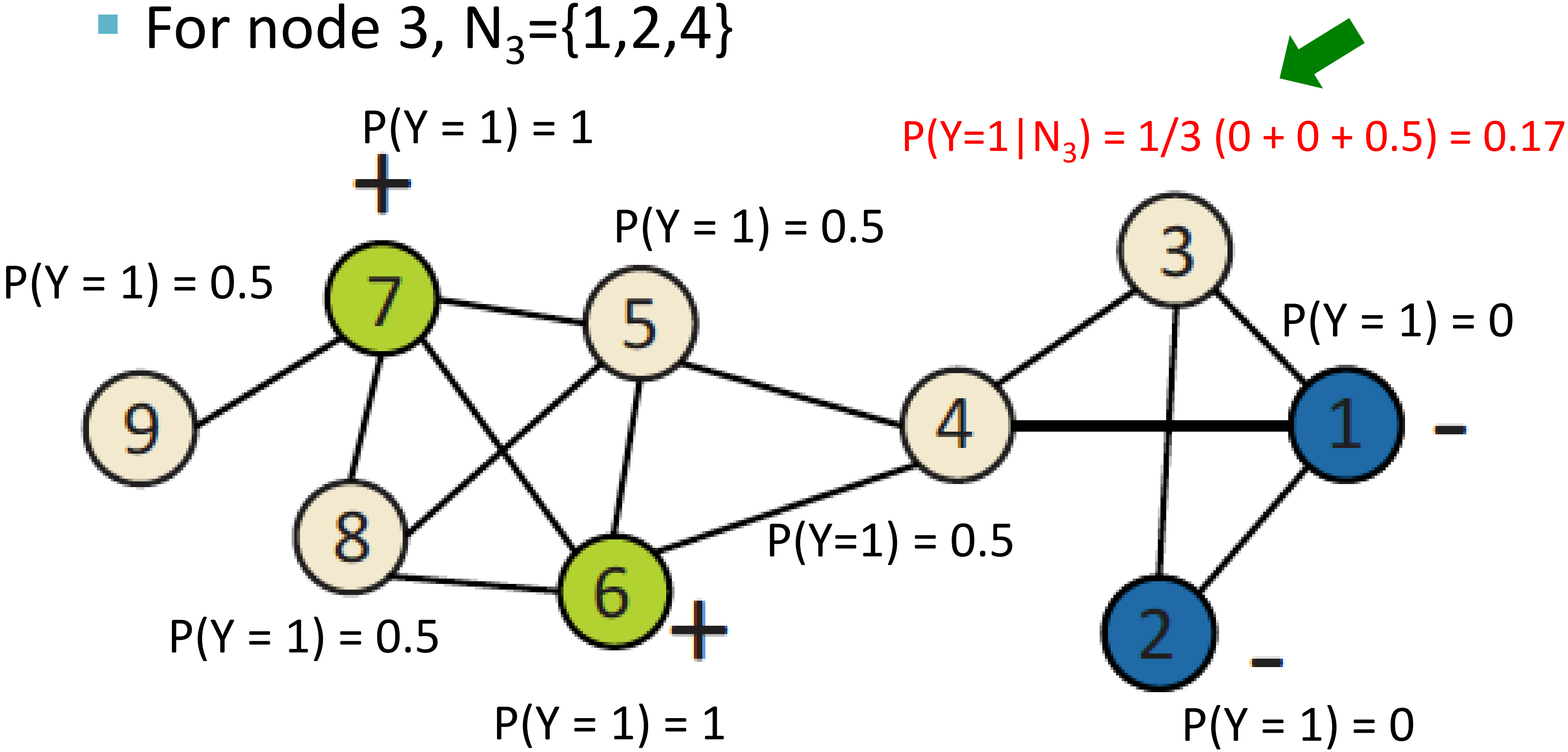
Probabilistic relational classifier example

Initialization: All labeled nodes to their labels, and all unlabeled nodes uniformly



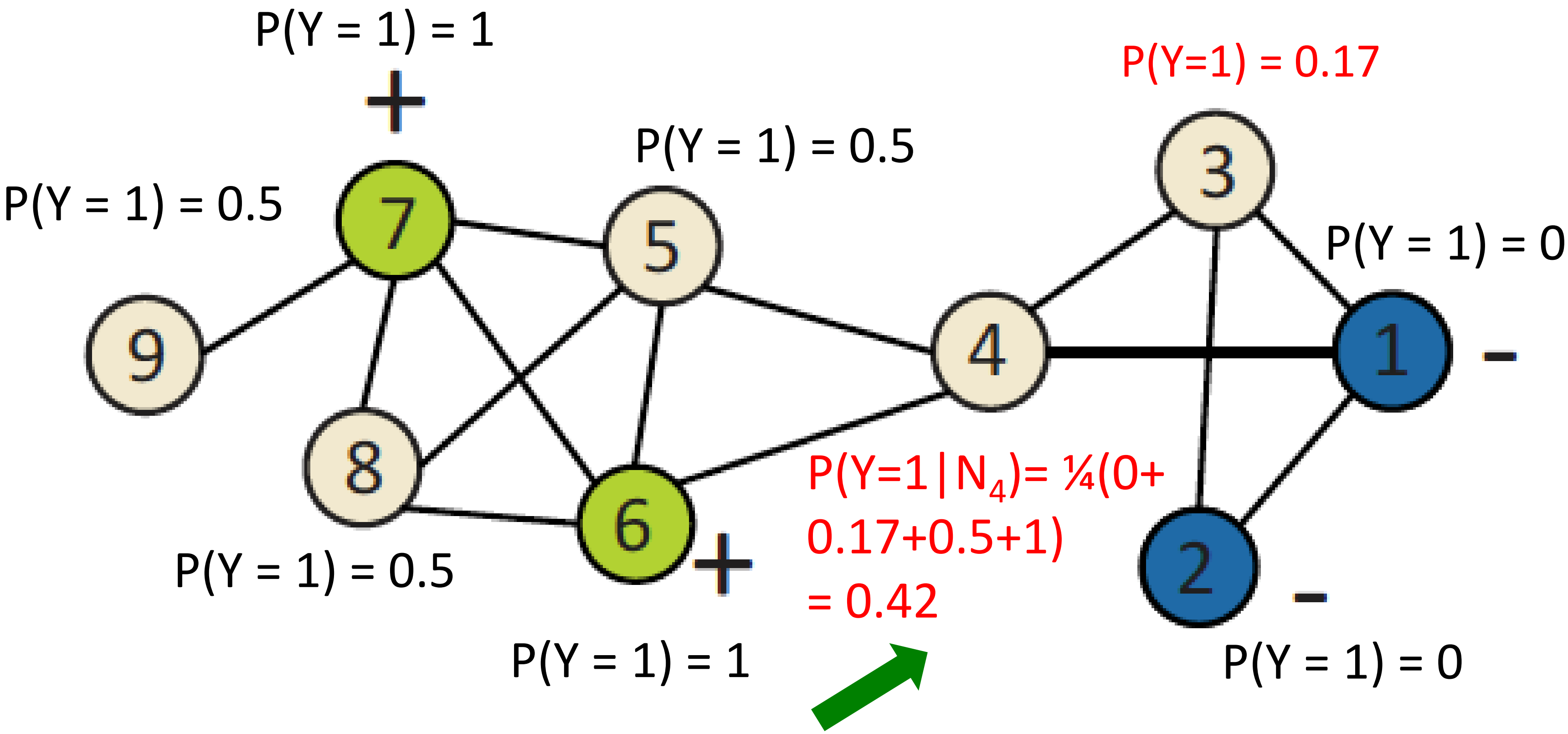
Probabilistic relational classifier example

- Update for the 1st Iteration:
  - For node 3, N_3={1,2,4}



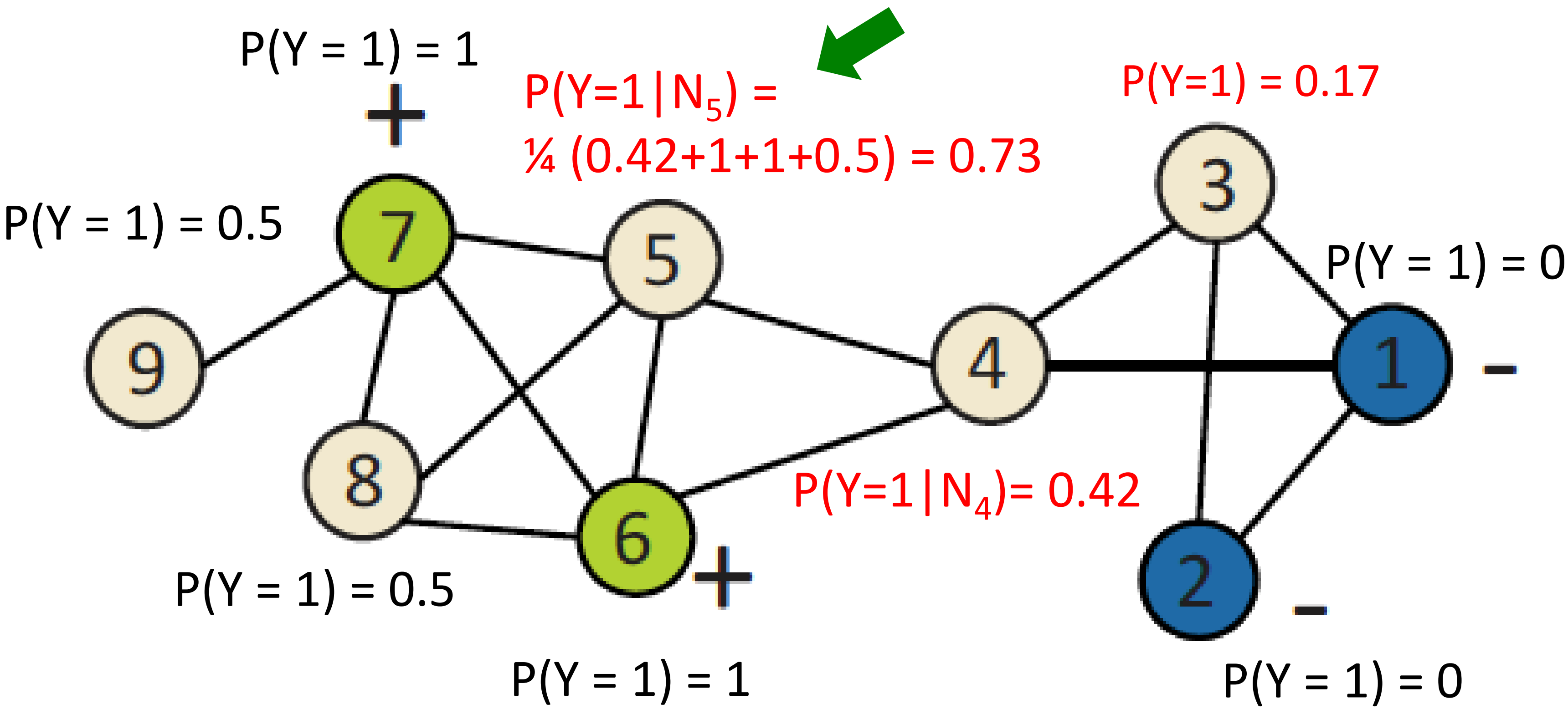
Probabilistic relational classifier example

- Update for the 1st Iteration:
  - For node 4, N_4={1,3, 5, 6}



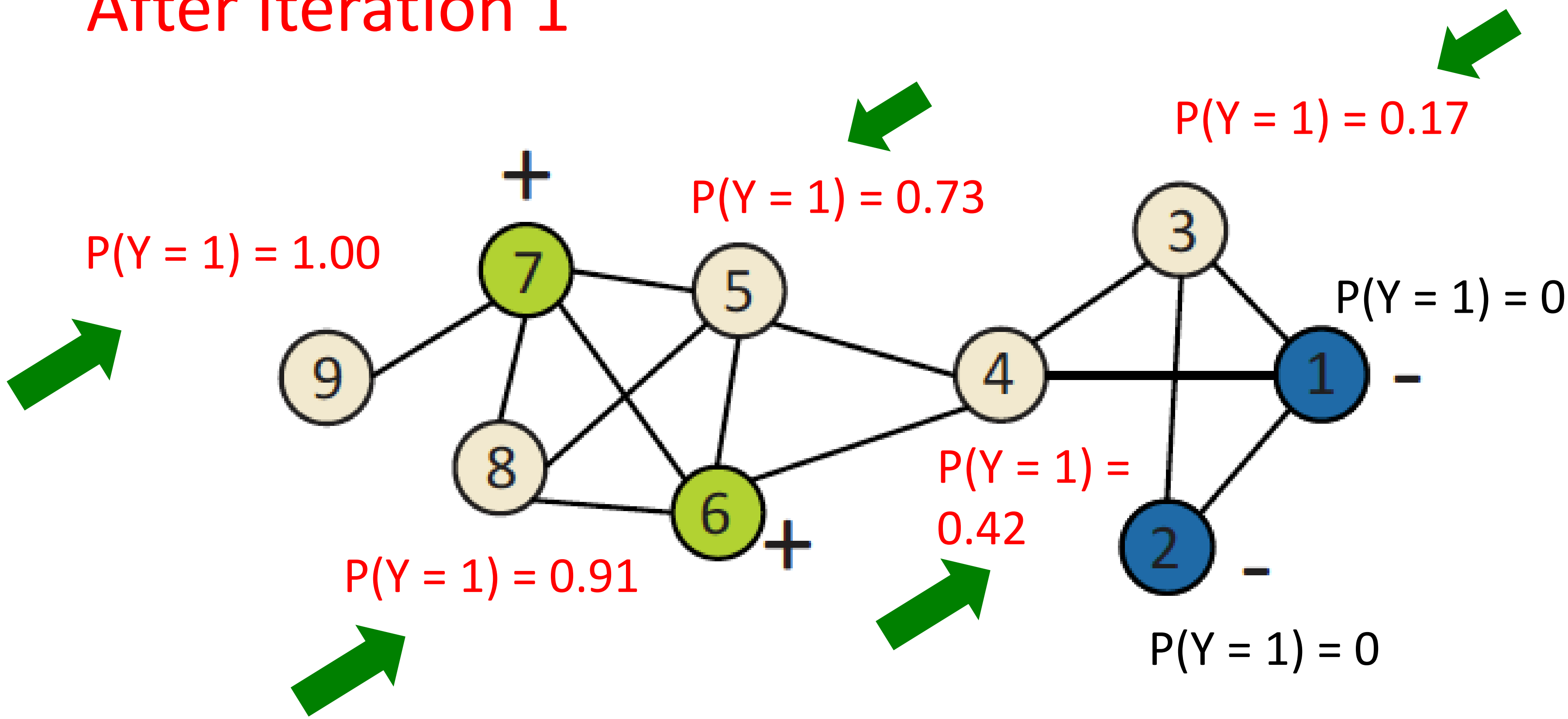
Probabilistic relational classifier example

- Update for the 1st Iteration:
  - For node 5, N_5={4,6,7,8}



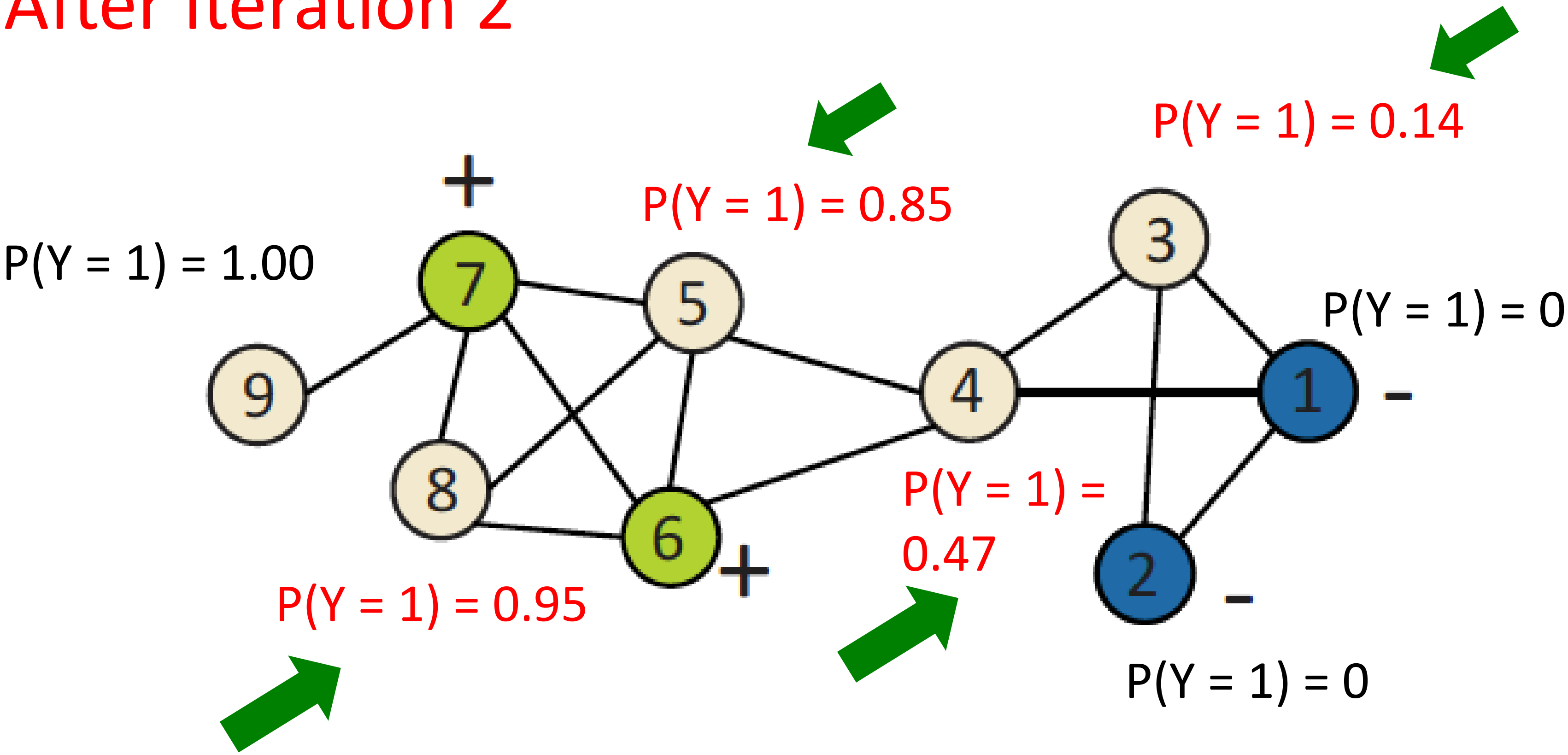
Probabilistic relational classifier example

After Iteration 1



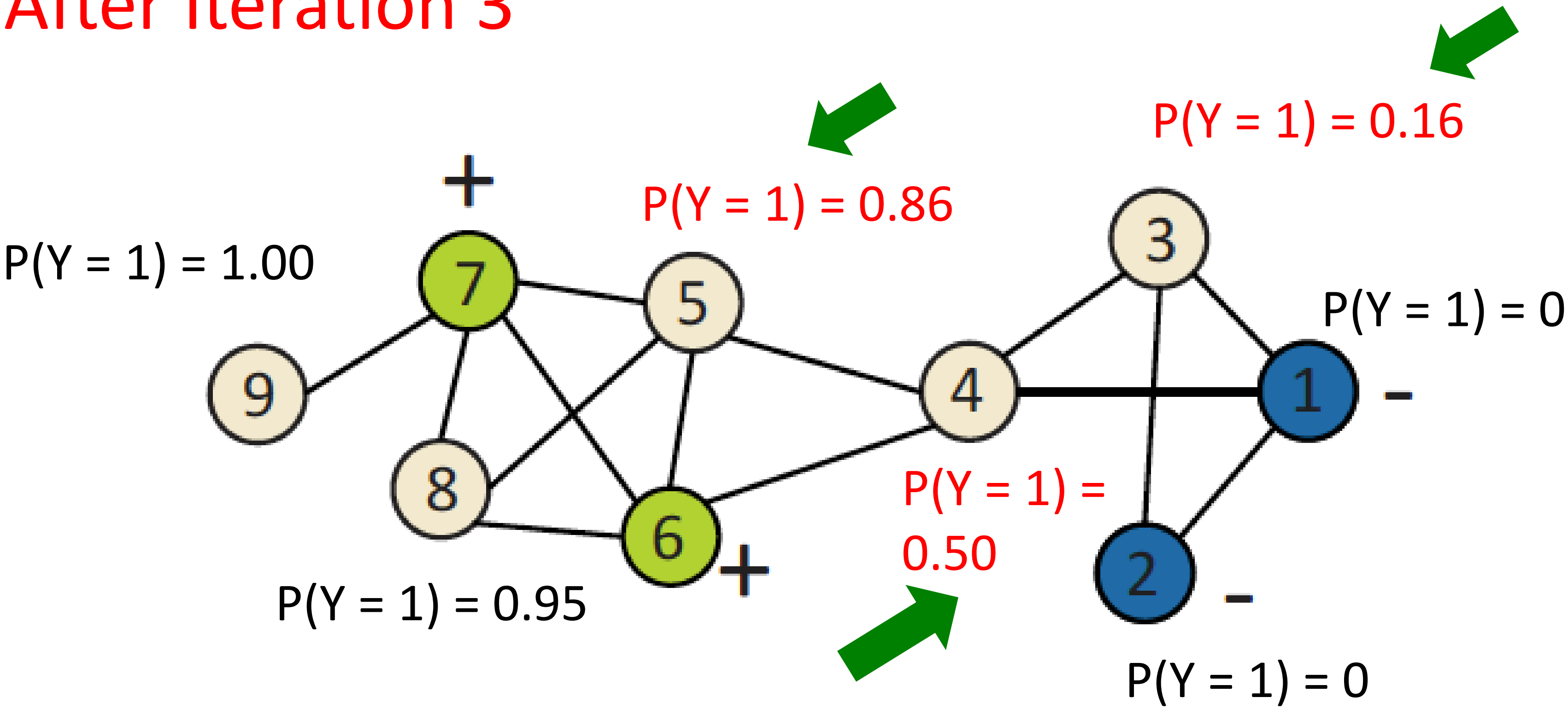
Probabilistic relational classifier example

After Iteration 2



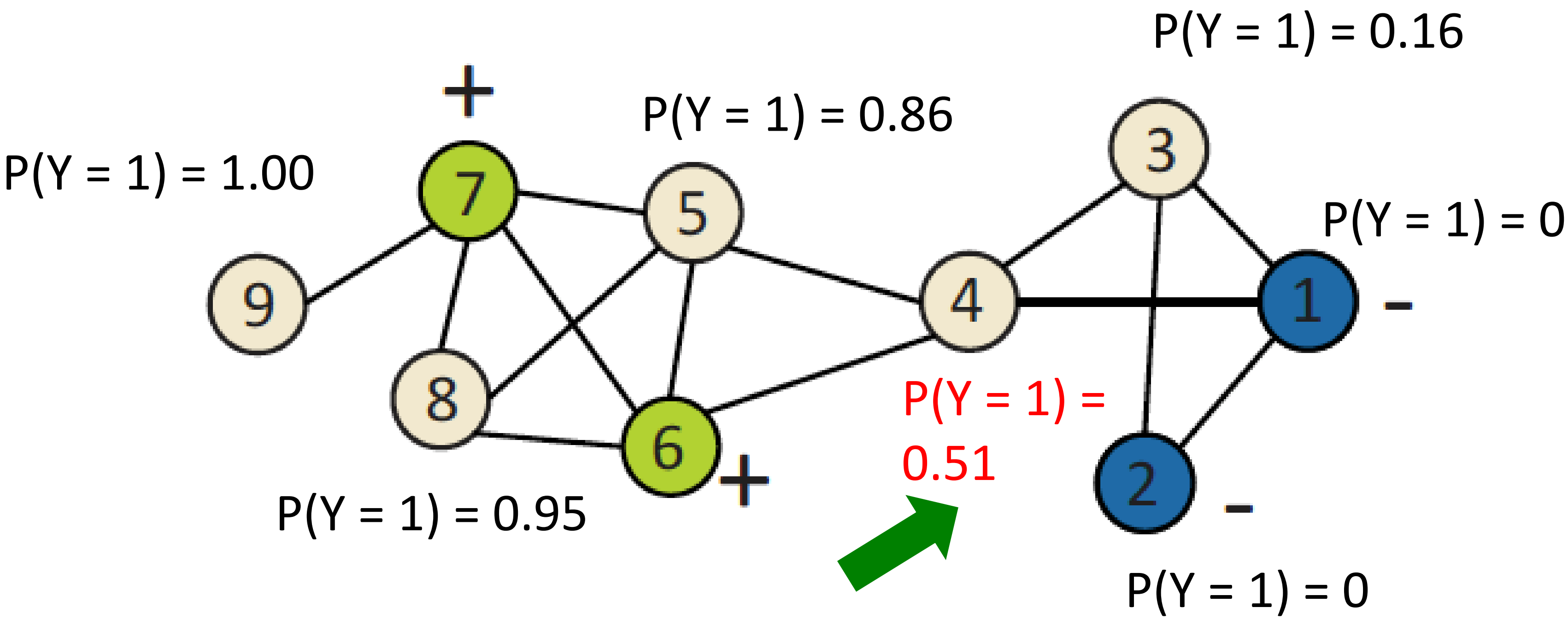
Probabilistic relational classifier example

After Iteration 3



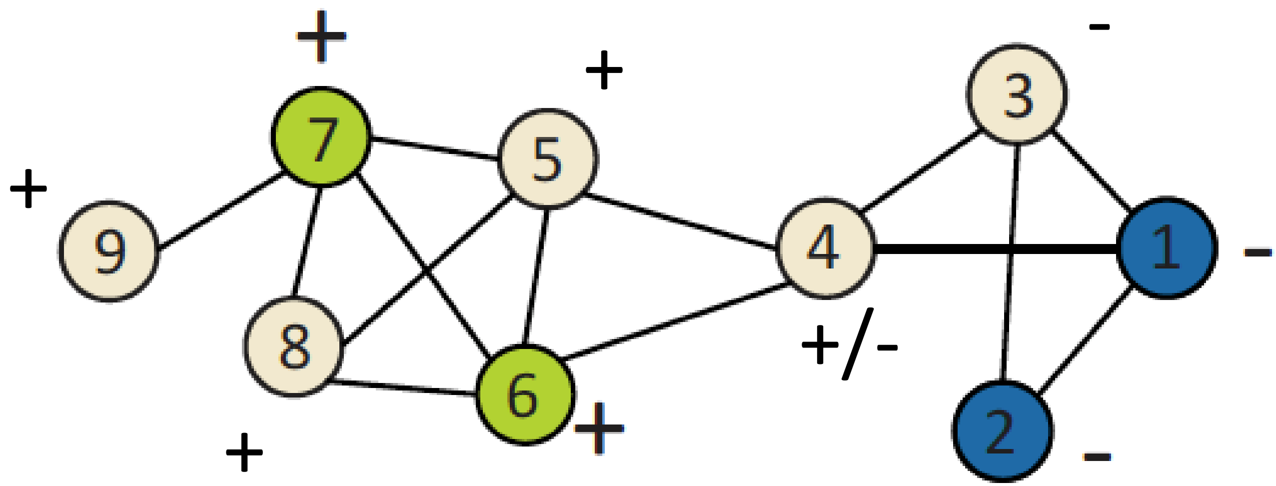
Probabilistic relational classifier example

After Iteration 4



Probabilistic relational classifier example

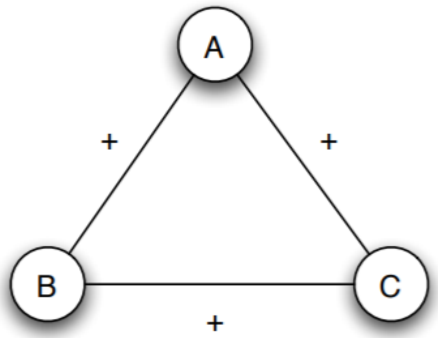
- All scores stabilize after 5 iterations:
  - Nodes 5, 8, 9 are + ($P(Y_i = 1) > 0.5$)
  - Node 3 is – ($P(Y_i = 1) < 0.5$)
  - Node 4 is in between ($P(Y_i = 1) = 0.5$)



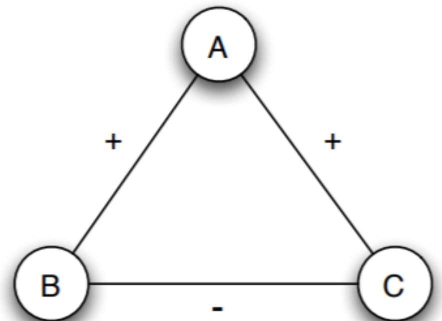
Structural Balance

Local effects can have global consequences that are observable at the level of the network as a whole

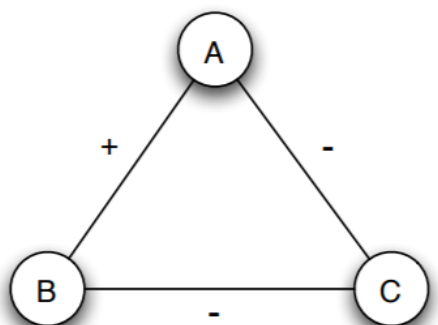
Starting from the local



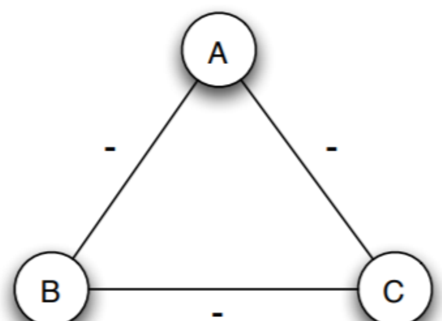
(a) *A, B, and C are mutual friends: balanced.*



(b) *A is friends with B and C, but they don't get along with each other: not balanced.*



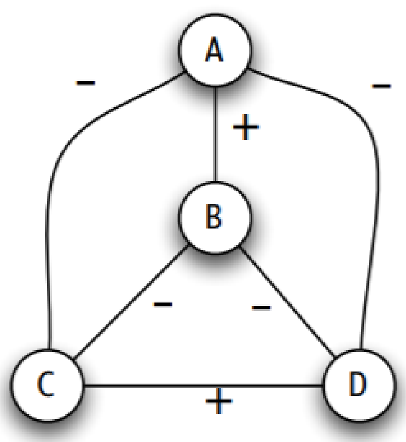
(c) *A and B are friends with C as a mutual enemy: balanced.*

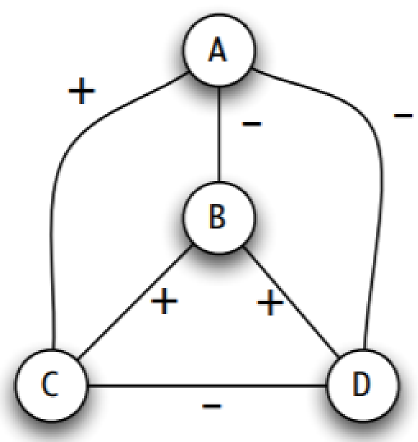


(d) *A, B, and C are mutual enemies: not balanced.*

Structural balance

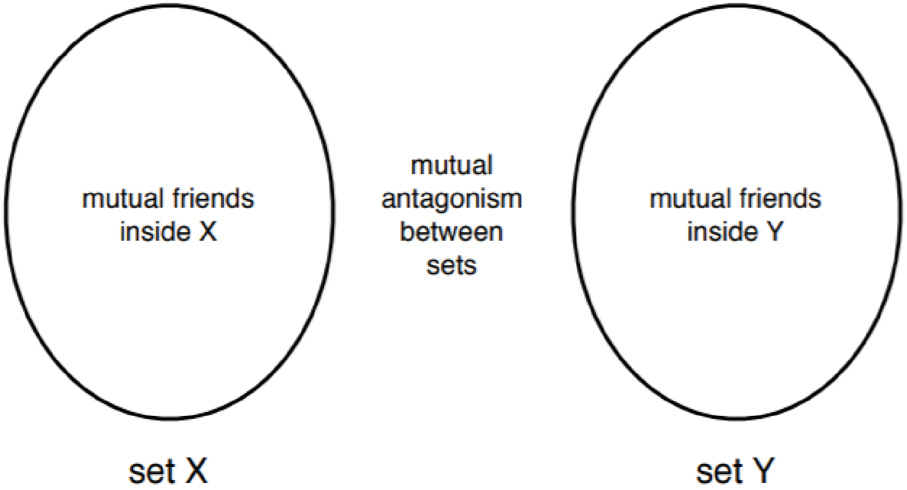
- Structural balance property: For every set of three nodes, if we consider the three edges connecting them, either all three of these edges are labeled +, or else exactly one of them is labeled +.





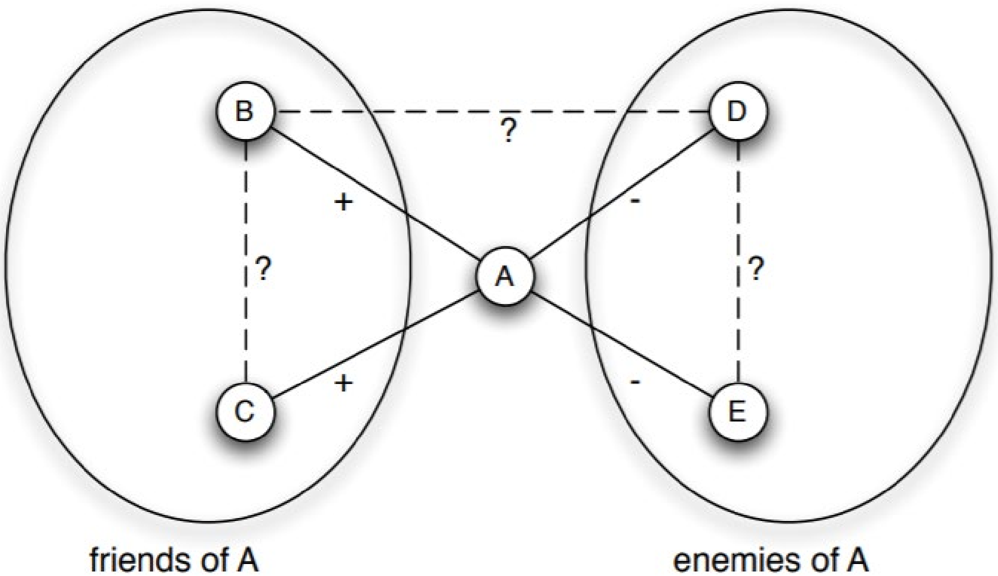
Structure of a balanced networks

- Balance theorem: If a labeled complete graph is balanced, then either all pairs of nodes are friends, or else 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 everyone in X is the enemy of everyone in Y .



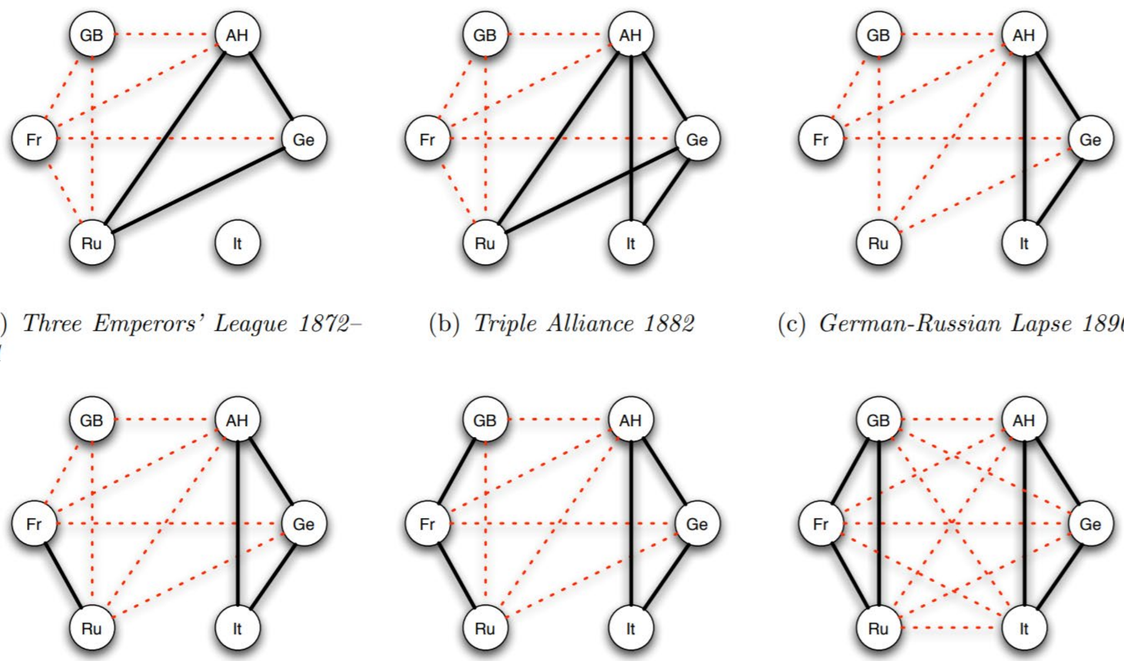
Proving the Balance Theorem

- To satisfy balance theorem, we have to
  - (1) every nodes in X are friends
  - (2) every nodes in Y are friends
  - (3) every node in X is an enemy of every node in Y



Balance: good or bad?

- Search for balance can lead to two implacably opposed alliances



(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*