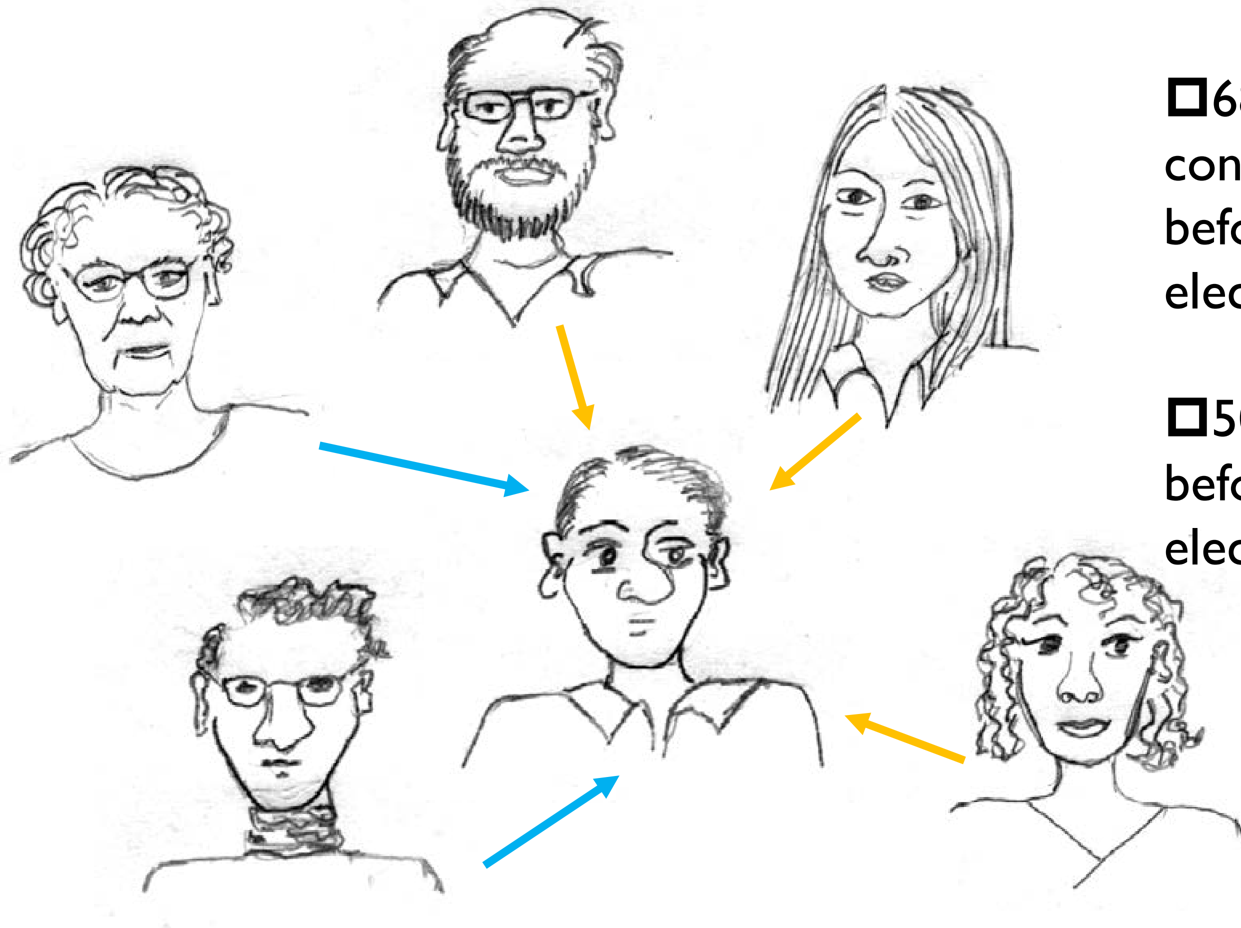
Web/Networks Structure Mining: Influence

Viral Marketing?

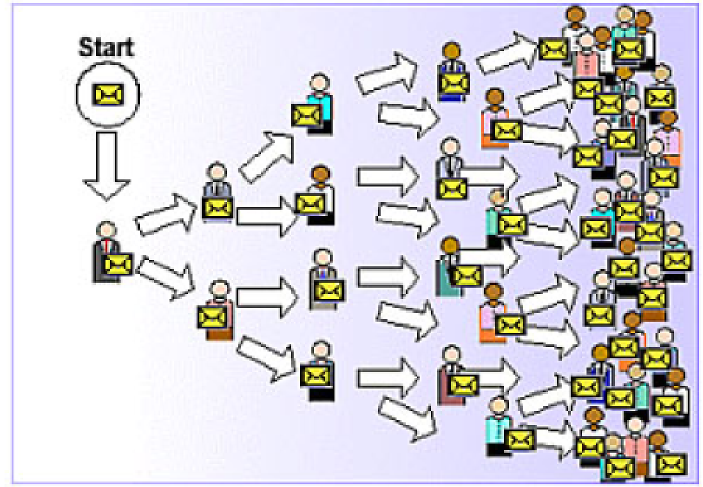
- We are more influenced by our friends than strangers



☐68% of consumers consult friends and family before purchasing home electronics (Burke 2003)

☐50% do research online before purchasing electronics

Viral Marketing



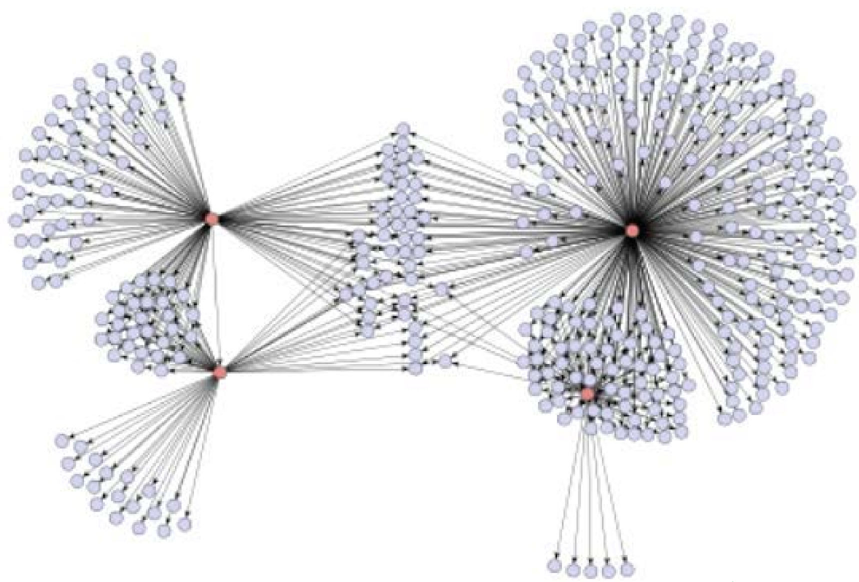
Identify influential customers

Convince them to adopt the product – Offer discount/free samples

These customers endorse the product among their friends

How to Create Big Cascades?

- Blogs – information epidemics （传播）
  - Which are the influential blogs?
  - Which blogs create big cascades?
  - Where should we advertise?



Which node shall we target?

Outline

Nodes, ties and influence

- Importance of nodes
- Strength of ties
- Influence modeling
- Influence maximization

Importance of Nodes

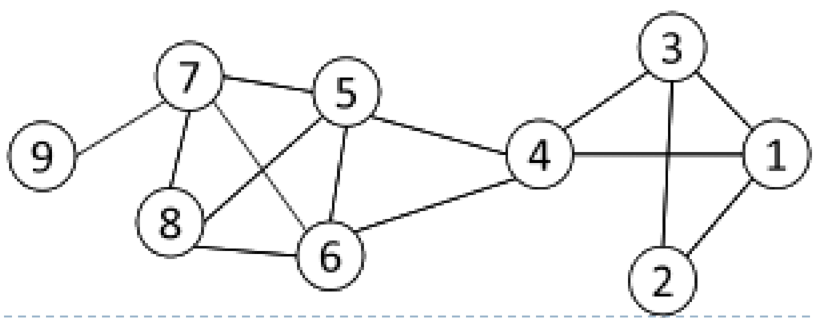
- Not all nodes are equally important
- Centrality（核心性，中心效应）Analysis:
  - Find out the most important nodes in one network

- Commonly-used Measures
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality

Degree Centrality

- The importance of a node is determined by the number of nodes adjacent to it
  - The larger the degree, the more important the node is
  - Only a small number of nodes have high degrees in many real-life networks

- Degree centrality: $C_D(v_i) = d_i = \sum_j A_{ij}$
- Normalized degree centrality: $C'_D(v_i) = d_i/(n-1)$



For node 1, degree centrality is 3;
Normalized degree centrality is
3/(9-1)=3/8.

Closeness Centrality

- "Central" nodes are important, as they can reach the whole network more quickly than non-central nodes
- Importance measured by how close a node is to other nodes
- Average Distance: $D_{avg}(v_i) = \frac{1}{n-1}\sum_{j \neq i}^{n} g(v_i, v_j)$

$g(v_i, v_j)$ denotes the geodesic distance（测地距离） between nodes v_i, v_j

- Closeness Centrality:

$$C_C(v_i) = \left[\frac{1}{n-1}\sum_{j \neq i}^{n} g(v_i, v_j)\right]^{-1} = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)}$$

Closeness Centrality Example

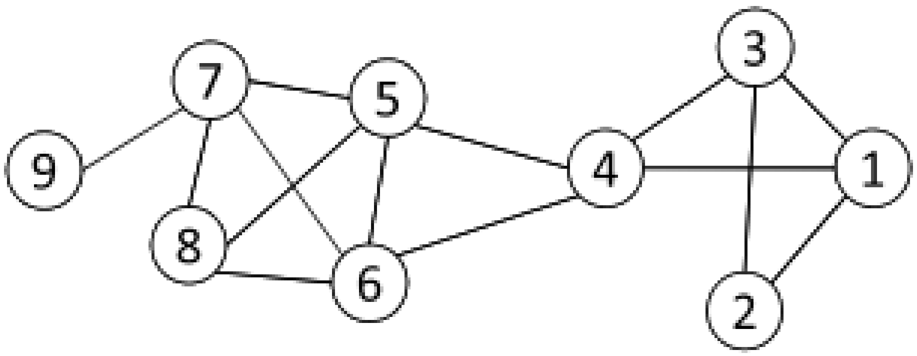


Table 2.1: Pairwise geodesic distance

Node	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

Node 4 is more central than node 3

Betweenness Centrality

- Nodes betweenness counts the number of shortest paths that pass one node
- Nodes with high betweeness are important in communication and information diffusion

- Betweenness centrality: $C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$
  - σ_{st} :The number of shortest paths between *s* and *t*
  - $\sigma_{st}(v_i)$:The number of shortest paths between *s* and *t* that passes v_i
- Normalized betweenness centrality: $C'_B(i) = \frac{C_B(i)}{(n-1)(n-2)/2}$

Betweenness Centrality Example

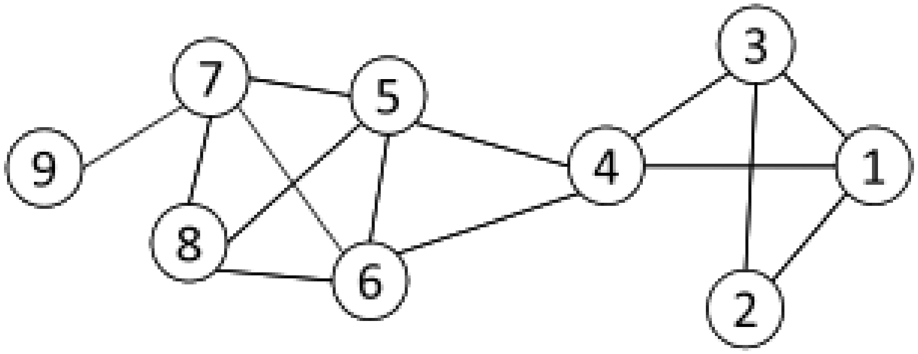


Table 2.2: $\sigma_{st}(4)/\sigma_{st}$

	$s = 1$	$s = 2$	$s = 3$
$t = 5$	1/1	2/2	1/1
$t = 6$	1/1	2/2	1/1
$t = 7$	2/2	4/4	2/2
$t = 8$	2/2	4/4	2/2
$t = 9$	2/2	4/4	2/2

$$C_B(4) = 15$$
$$C_B(5) = 12 \times 0.5 = 6$$

Eigenvector Centrality

- One's importance is determined by his friends' importance
- If one has many important friends, he should be important as well.

$$C_E(v_i) \propto \sum_j A_{ij} C_E(v_i)$$

  - Let x denote the eigenvector centrality of node from v_1 to v_n

$$\mathbf{x} \propto A\mathbf{x} \quad \Rightarrow \quad A\mathbf{x} = \lambda \mathbf{x}.$$

- The centrality corresponds to the top eigenvector of the adjacency matrix A.
- A variant of this eigenvector centrality is the PageRank score.

Computation of Centrality Measures

Expensive except for degree centrality and eigenvector centrality

- Degree Centrality
  - easy
- Closeness Centrality
  - Time: $O(n^2)$; Space: $O(n^3)$ (Floy, 1962)
  - Time: $O(n^2 \log n + nm)$ (Johnson, 1977)
- Betweenness Centrality
  - $O(n^2)$ ($O(nm)$ with sparsity)
- Eigenvector Centrality
  - Power method (Golub and Van Loan, 1996)

Outline

Nodes, ties and influence

- Importance of nodes
- Strength of ties
- Influence modeling
- Influence maximization

Weak and Strong Ties

- In practice, connections are not of the same strength
- Interpersonal social networks are composed of strong ties (close friends) and weak ties (acquaintances).
- Strong ties and weak ties play different roles for community formation and information diffusion
- Strength of Weak Ties (*Granovetter, 1973*)
  - Occasional encounters with distant acquaintances can provide important information about new opportunities for job search

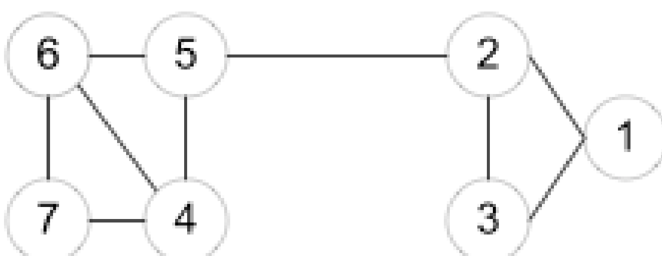
Connections in Social Media

- Social media allows users to connect to each other more easily than ever
  - One user might have thousands of friends online
  - Who are the most important ones among your 300 Facebook friends?

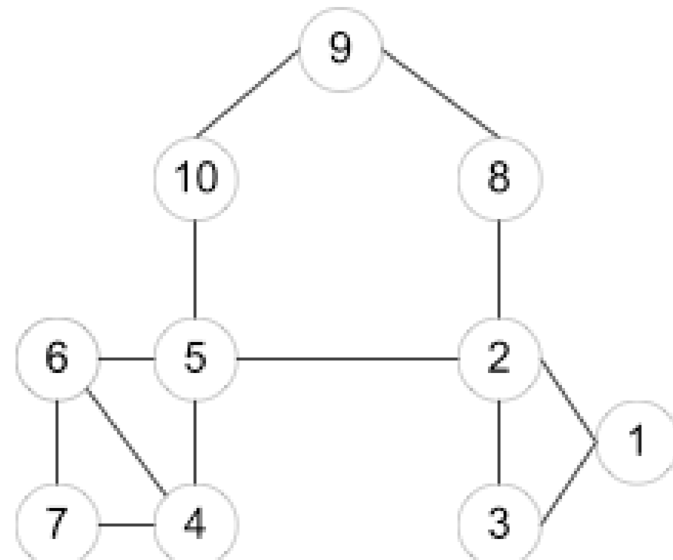
- Imperative to estimate the strengths of ties for advanced analysis
  - Analyze network topology
  - Learn from User Profile and Attributes
  - Learn from User Activities

Learning from Network Topology

- Bridges connecting two different communities are weak ties
- An edge is a *bridge* if its removal results in disconnection of its terminal nodes
- Bridges in a network are weak ties



e(2,5) is a bridge

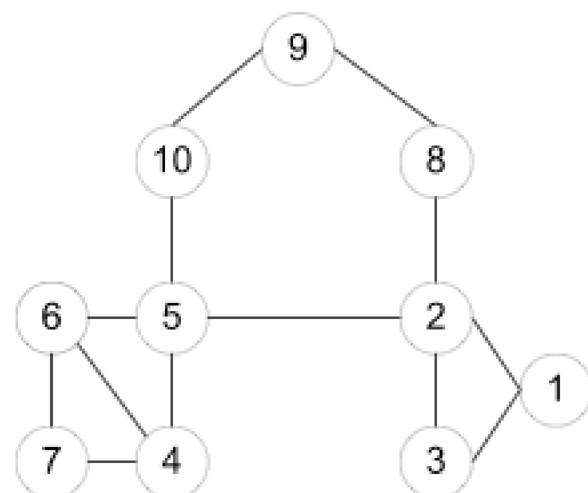


e(2,5) is NOT a bridge

“shortcut” Bridge

- Bridges are rare in real-life networks
- Alternatively, one can relax the definition by checking if the distance between two terminal nodes increases if the edge is removed
- The larger the distance, the weaker the tie is

- d(2,5) = 4 if e(2,5) is removed
- d(5,6) = 2 if e(5,6) is removed
- e(5,6) is a stronger tie than e(2,5)



Neighborhood Overlap

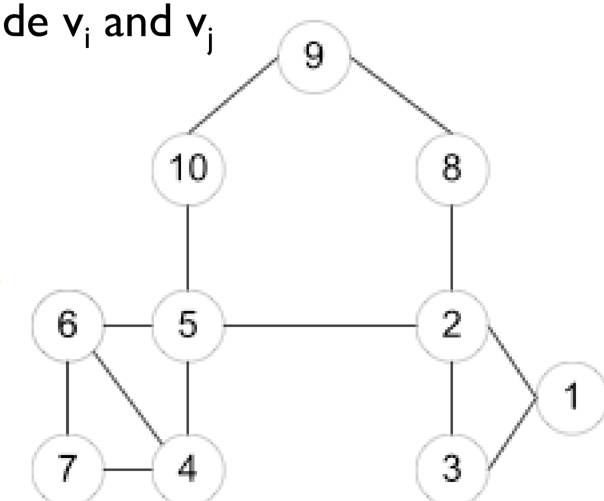
- Tie Strength can be measured based on neighborhood overlap; the larger the overlap, the stronger the tie is.

$$overlap(v_i, v_j) = \frac{\text{number of shared friends of both } v_i \text{ and } v_j}{\text{number of friends who are adjacent to at least } v_i \text{ or } v_j} = \frac{|N_i \cap N_j|}{|N_i \cup N_j| - 2}.$$

-2 in the denominator is to exclude v_i and v_j

$$overlap(2, 5) = 0,$$

$$overlap(5, 6) = \frac{|\{4\}|}{|\{2, 4, 5, 6, 7, 10\}| - 2} = 1/4$$



Learning from Profiles and Interactions

- Twitter: one can follow others without followee's confirmation
  - The real friendship network is determined by the frequency two users talk to each other, rather than the follower-followee network
  - The real friendship network is more influential in driving Twitter usage
- Strengths of ties can be predicted accurately based on various information from Facebook
  - Friend-initiated posts, message exchanged in wall post, number of mutual friends, etc.
- Learning numeric link strength by maximum likelihood estimation
  - User profile similarity determines the strength
  - Link strength in turn determines user interaction
  - Maximize the likelihood based on observed profiles and interactions

Learning from User Activities

- One might learn how one influences his friends if the user activity log is accessible
- Depending on the adopted influence model
  - Independent cascading model
  - Linear threshold model
- Maximizing the likelihood of user activity given an influence model

Outline

Nodes, ties and influence

- Importance of nodes
- Strength of ties
- Influence modeling
- Influence maximization

Influence Modeling

Influence modeling is one of the fundamental questions in order to understand the information diffusion（传播）, spread of new ideas, and word-of-mouth (viral) marketing

Well known Influence modeling methods

1. Linear threshold model (LTM)
2. Independent cascade model (ICM)

Common Properties of Influence Modeling Methods

- A social network is represented a *directed graph G=(V, E), with each actor being one node;*
- Each node is started as active or inactive;
- A node, once activated, will activate his neighboring nodes;
- Once a node is activated, this node cannot be deactivated.

Linear Threshold Model

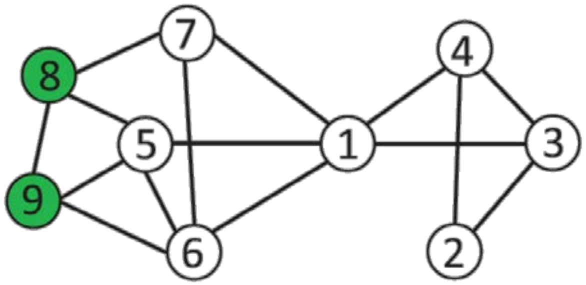
An actor would take an action if the number of his friends who have taken the action exceeds (reaches) a certain threshold

- Each node v chooses a threshold Θ_v randomly from a uniform distribution in an interval between 0 and 1.
- A neighbor w can influence node v with strength $b_{w,v}$
- In each discrete step, all nodes that were active in the previous step remain active
- The nodes satisfying the following condition will be activated

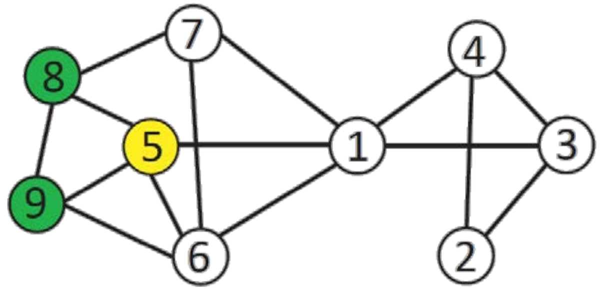
$$\sum_{w \in N_v, w \text{ is active}} b_{w,v} \geq \theta_v$$

Linear Threshold Model - Diffusion Process

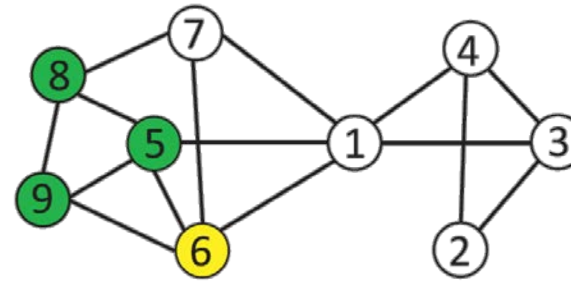
(Threshold = 50%)



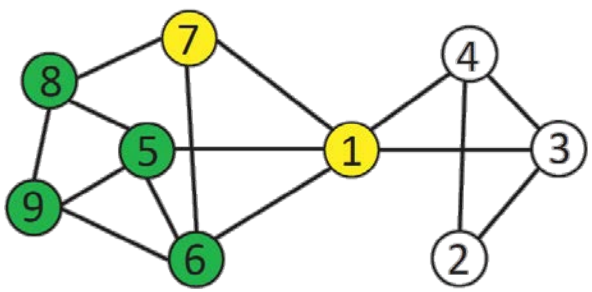
Step 0



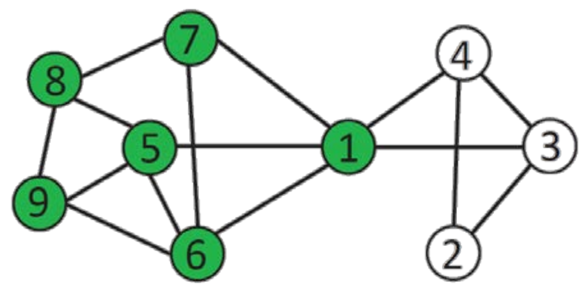
Step 1



Step 2



Step 3



Final Stage

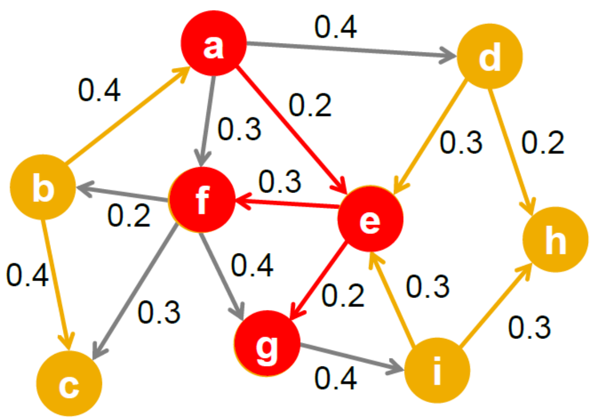
Independent Cascade Model (ICM)

The independent cascade model focuses on the sender's rather than the receiver's view

- A node *w*, once activated at step *t* , has one chance to activate each of its neighbors randomly
  - For a neighboring node (say, *v*), the activation succeeds with probability $p_{w,v}$(e.g. p = 0.5)
- If the activation succeeds, then *v* will become active at step *t*+1
- In the subsequent rounds, *w* will not attempt to activate *v* anymore.
- *The diffusion process*, starts with an initial activated set of nodes, then continues until no further activation is possible

Independent Cascade Model (ICM)

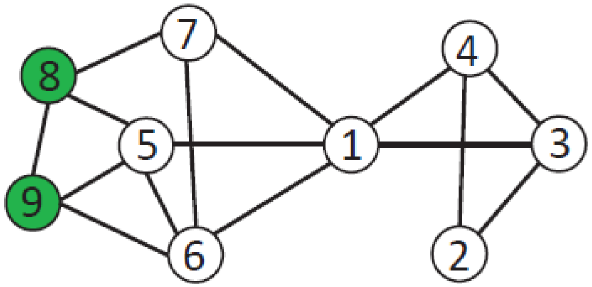
- Initially some nodes *S* are active
- Each edge (*w*, *v*) has probability (weight) $p_{w,v}$



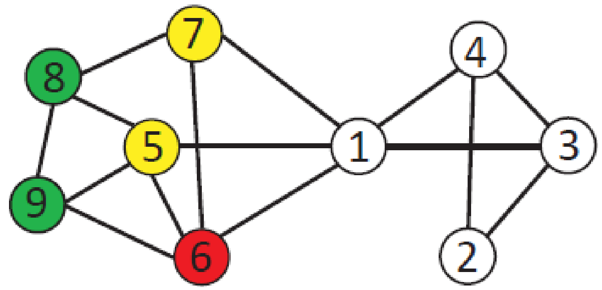
- When node *w* becomes active:
  - It activates each out-neighbor *v* with probability $p_{w,v}$

Activations spread through the network

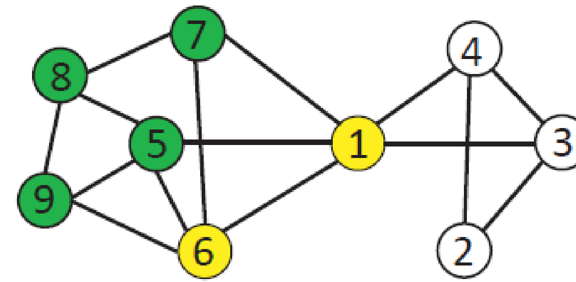
Independent Cascade Model- Diffusion Process ($p_{w,v}$ = 50%)



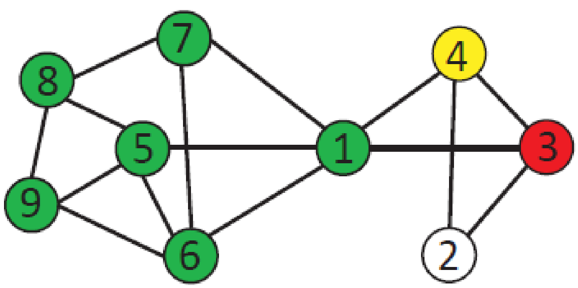
Step 0



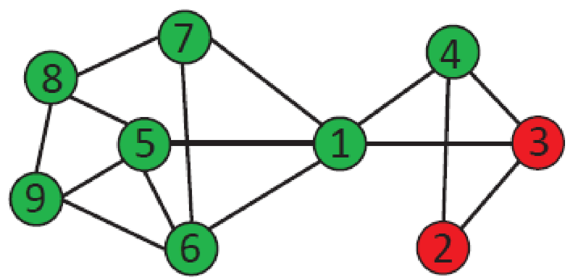
Step 1



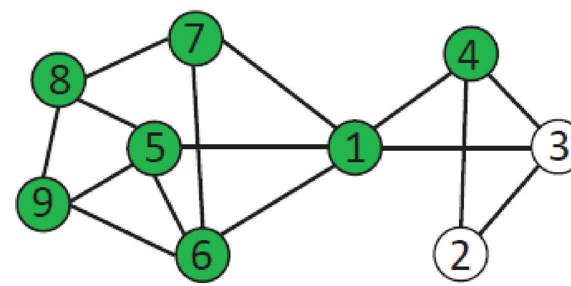
Step 2



Step 3



Step 4



Final Stage

Remarks on LTM vs. ICM

Two basic models used to study influence and information diffusion

- LTM: receiver-centered; ICM: sender-centered
- LTM's activation depends on the whole neighborhood of one node
- ICM activates each of its neighbors independently
- Once the thresholds are sampled, the LTM diffusion process is determined
- ICM varies depending on the cascading process

- Both are submodular

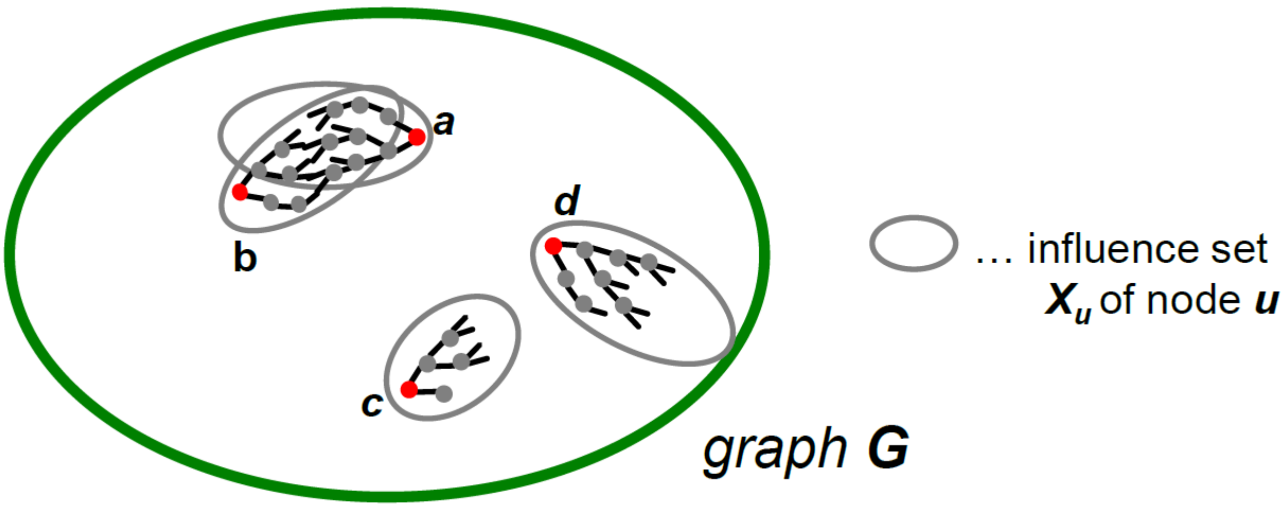
Outline

Nodes, ties and influence

- Importance of nodes
- Strength of ties
- Influence modeling
- Influence maximization

Most Influential Set of Nodes

- *S*: is initial active set
- *f(S)*: The expected size of final active set



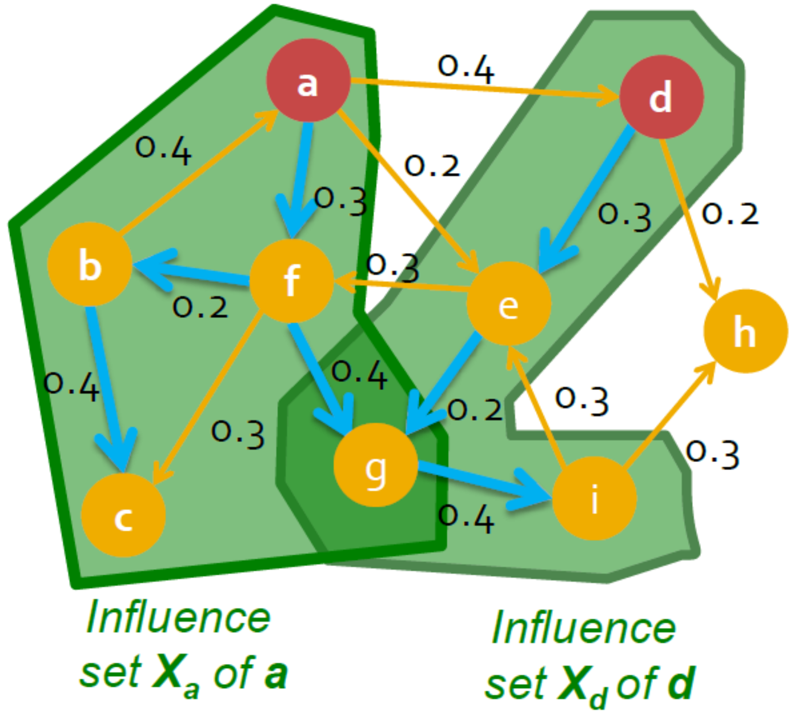
- Set *S* is more influential if *f(S)* is larger

$$f(\{a, b\}) < f(\{a, c\}) < f(\{a, d\})$$

Most Influential Set

Problem: (k is user specified parameter)

- **Most influential set of size *k*:** set ***S*** of ***k*** nodes producing largest **expected** cascade size ***f(S)*** if activated [Domingos-Richardson '01]
- **Optimization problem:** $\max_{S \text{ of size } k} f(S)$



Why "expected cascade size"? X_a is a result of a random process. So in practice we would want to compute X_a for many realizations and then maximize the "average" value $f(S)$

$$f(S) = \sum_{\substack{\text{Random} \\ \text{realizations } i}} f_i(S)$$

Most Influential Subset of Nodes

- Most influential set of *k* nodes

  Set *S* on *k* nodes producing largest expected cascade size f(S) is activated
- The optimization problem:

$$\max_{S \text{ of size k}} f(S)$$

- How hard is this problem?
  - NP-COMPLETE
    - Finding most influential set is at least as hard as a vertex cover

Influence Maximization

☹Bad news:

- Influence maximization is NP-complete

☺Next, good news:

- There exists an approximate algorithm

- Consider the Hill Climbing algorithm to find *S*
  - Input:
    Influence set of each node *u*: $X_u = \{v_1, v_2, ...\}$
    - If we activate *u*, nodes $\{v_1, v_2, ...\}$ will eventually get active
  - **Algorithm:** At each iteration *i* take the node *u* that gives best marginal gain: $\max_u f(S_{i-1} \cup \{u\})$

f(S): Size of the union of $X_u, u \in S$

(Greedy) Hill Climbing

Algorithm:

- Start with $S_0 = \{\}$
- For *i*=1,…*k*
  - Take node *u* that $\max_u f(S_{i-1} \cup \{u\})$
  - Let $S_i = S_{i-1} \cup \{u\}$

Approximation Guarantee

- Claim: Hill climbing produces a solution *S*
- Where *f(S)* ≥(1-1/e)*OPT *(f(S)*>0.63*OPT*)*

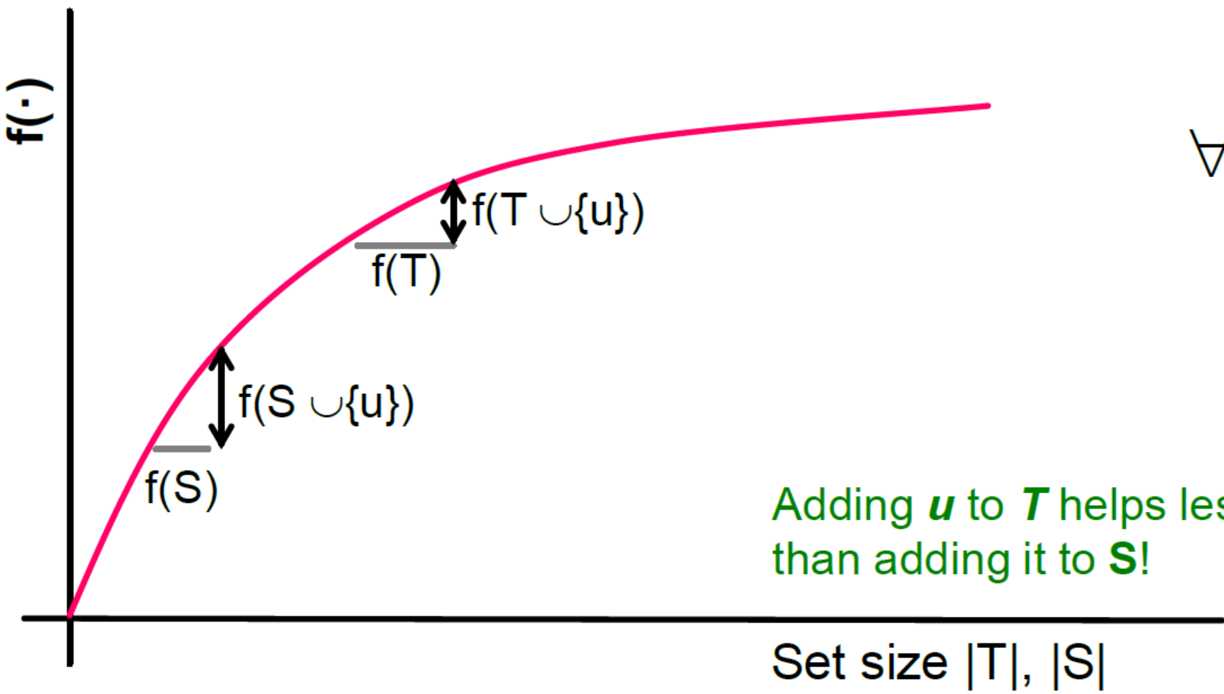
  [Nemhauser, Fisher, Wolsey '78, Kempe, Kleinberg, Tardos '03]
- Claim holds for function *f(.)* with 2 properties:
- *f* is monotone（单调）: (activating more nodes doesn't hurt)

  $$\textbf{if } S \subseteq T \text{ then } f(S) \leq f(T) \text{ and } f(\{\}) = 0$$
- *f* is submodular（子模性）: (activating each additional node helps less)
- Adding an element to a set gives less improvement than adding it to one of its subsets: $\forall S \subseteq T$

$$\underbrace{f(S \cup \{u\}) - f(S)}_{\text{Gain of adding a node to a small set}} \geq \underbrace{f(T \cup \{u\}) - f(T)}_{\text{Gain of adding a node to a larger set}}$$

Submodularity – Diminishing Returns

- Diminishing returns



$\forall \boldsymbol{S} \subseteq \boldsymbol{T}$

Adding ***u*** to ***T*** helps less than adding it to **S**!

$$f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$$

Gain of adding a node to a small set

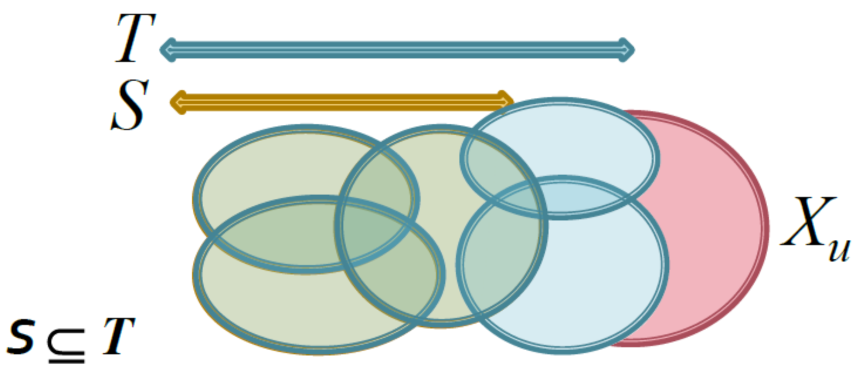
Gain of adding a node to a larger set

Two Claims

- Sets $X_1, \ldots, X_m$

  $f(S) = |\cup_{i \in S} X_i|$ (size of the union of sets $X_i, i \in S$)

Claim 1: *f(S)* is submodular



The more sets you already have the less new area a given set will cover

Claim 2: (Greedy) Hill climbing gives near-optimal solutions

Solution Quality

- Hill climbing finds solution *S* which

  $$f(S) \geq (1-1/e)*OPT \quad \textit{i.e.,} \quad f(S) > 0.63*OPT$$

- This is data independent bound
  - This is a worst case bound
  - No matter what is the input data, we know that the Hill-Climbing will never do worse than 0.63*OPT

Simulation Experiments

- A collaboration network: co-authorships in papers of the arXiv high-energy physics theory:
  - 10,748 nodes
  - 53,000 edges

- Independent Cascade Model:
  - Case 1: Uniform probability *p* on each edge
  - Case 2: Edge from *u* to *v* has probability *1/deg(v)* of activating *v*.

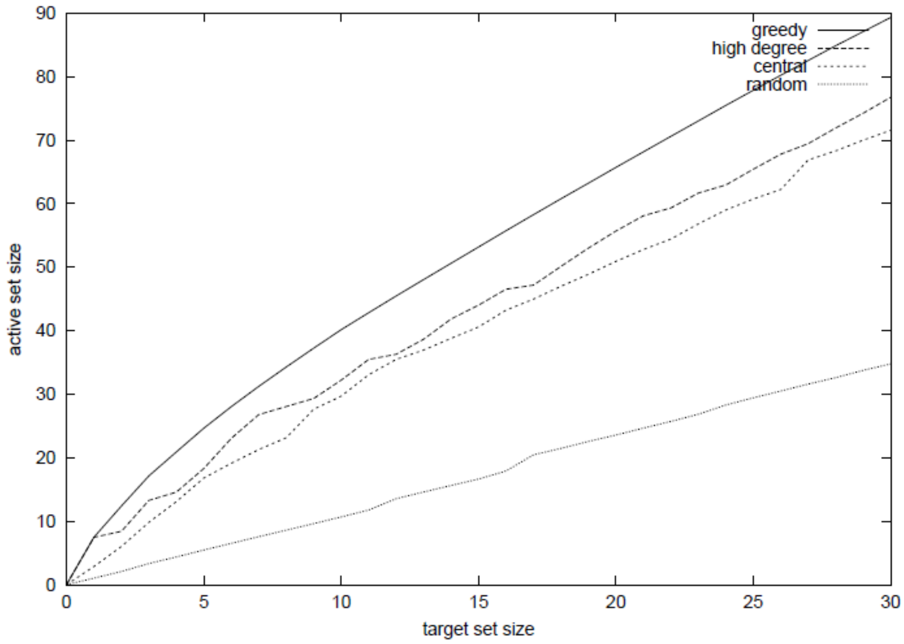
Experiment Settings

- Simulate the process 10,000 times for each targeted set
  - Every time re-choosing edge outcomes randomly

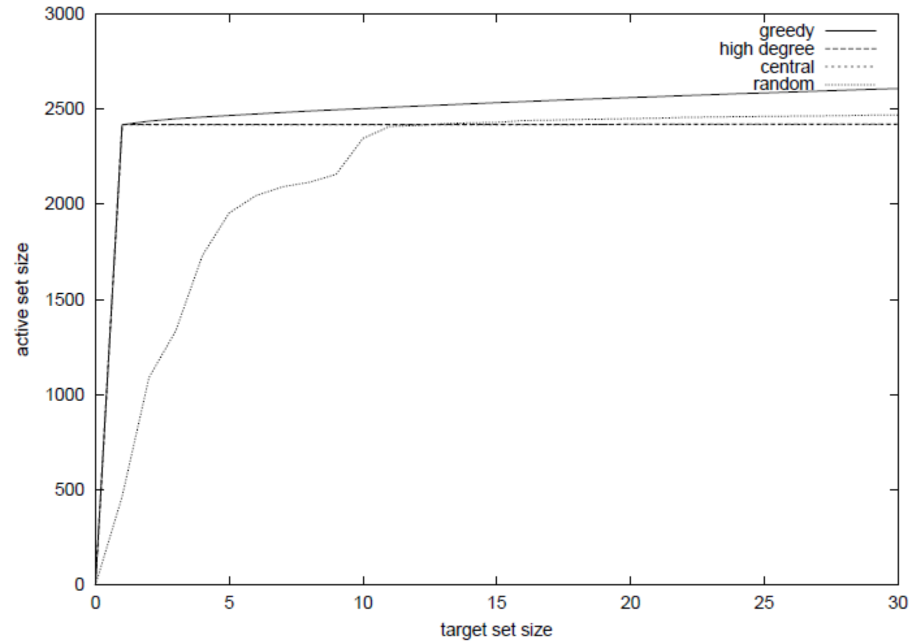
- Compare with other 3 common heuristics
  - Degree centrality
  - Distance centrality
  - Random nodes

Independent Cascade Model

p_{uv}=1%

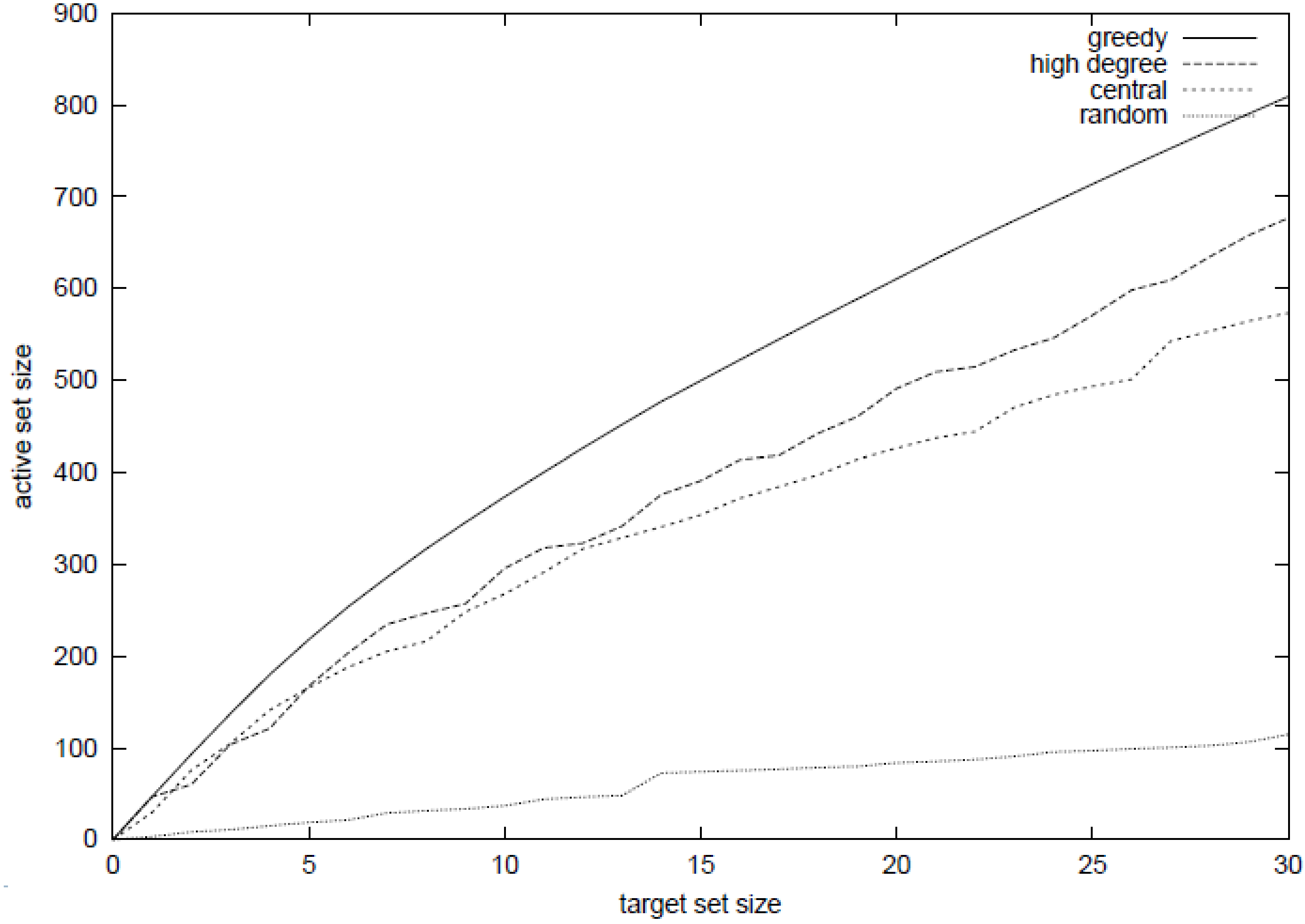


p_{uv}=10%



Independent Cascade Model

$p_{uv}=1/deg(v)$



Web/Networks Structure Mining: Challenges in Information Networks

Challenges

- Scalability
  - Social networks are often in a scale of millions of nodes and connections
  - Traditional Network Analysis often deals with at most hundreds of subjects
- Heterogeneity
  - Various types of entities and interactions are involved
- Evolution
  - Timeliness is emphasized in social media
- Collective Intelligence
  - How to utilize wisdom of crowds in forms of tags, wikis, reviews
- Evaluation
  - Lack of ground truth, and complete information due to privacy

Social Computing Tasks

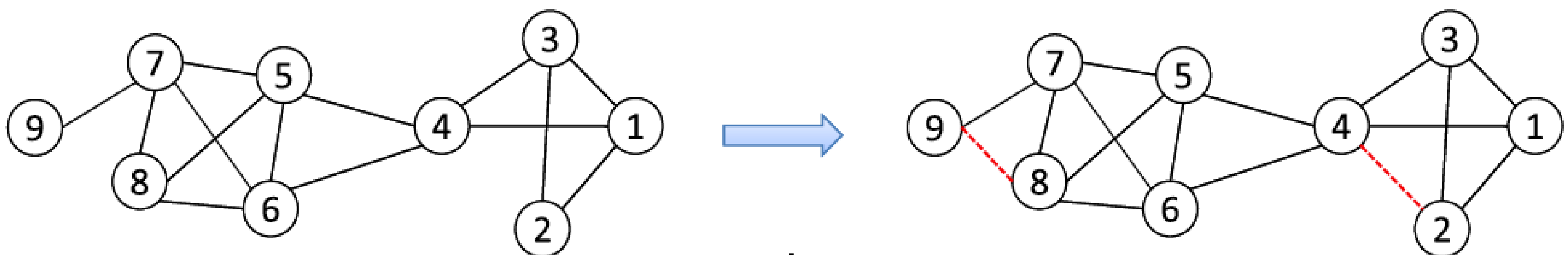
- Social Computing: a young and vibrant field
- Many new challenges
- Tasks
  - Network Modeling
  - Centrality Analysis and Influence Modeling
  - Community Detection
  - Classification and Recommendation
  - Privacy, Spam and Security

Privacy, Spam and Security

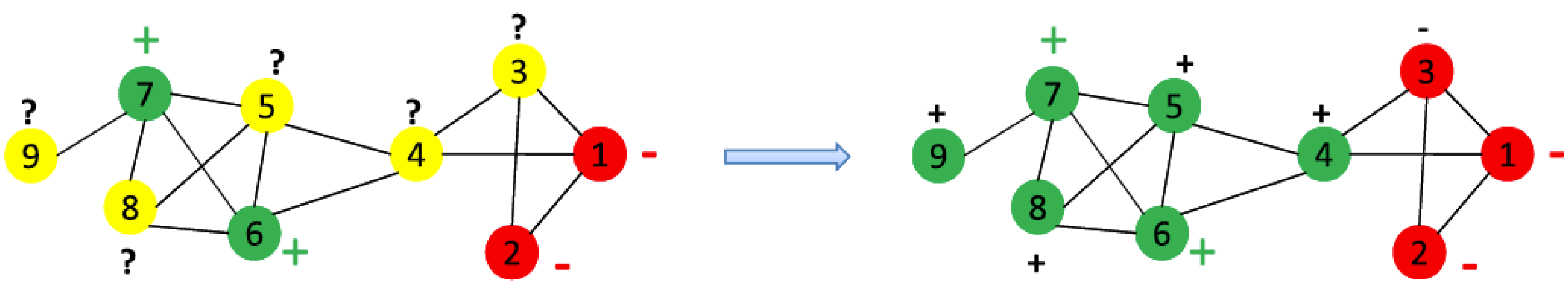
- Privacy is a big concern in social media
  - Facebook, Google buzz often appear in debates about privacy
  - NetFlix Prize Sequel cancelled due to privacy concern
  - Simple annoymization does not necessarily protect privacy
- Spam blog (splog), spam comments, Fake identity, etc., all requires new techniques
- As private information is involved, a secure and trustable system is critical
- Need to achieve a balance between sharing and privacy

Classification and Recommendation

- Common in social media applications
  - Tag suggestion, Friend/Group Recommendation, Targeting



Link prediction



Network-Based Classification