Social Media Analytics

BITS Pilani

Garima Jindal
Faculty Department

innovate achieve lead

BITS Pilani

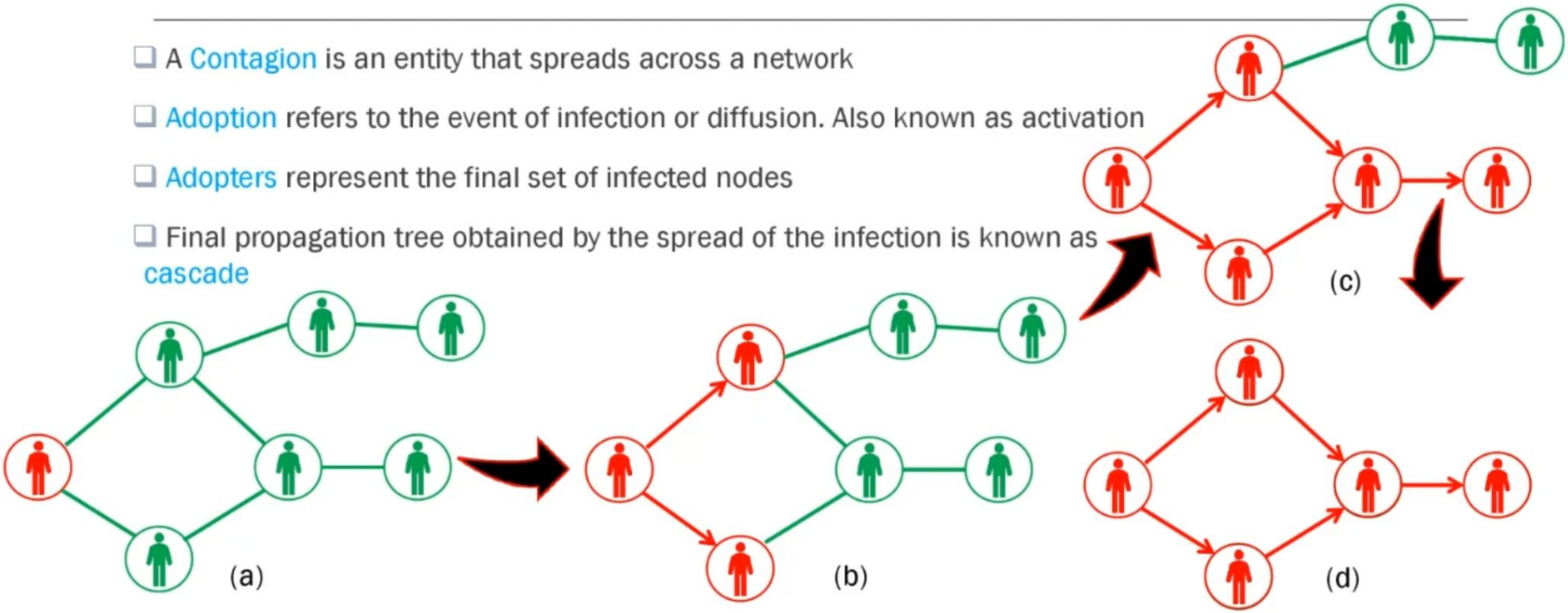
Social Media Analytics
Lecture No. 11

Recap

- Information discussion.
- Types of Diffusion MOdels
  - Decision-Based Diffusion Models
  - Probabilistic-Based Diffusion Models
- What are the advantage and disadvantage
- how they evolve over time

Information Diffusion: Terminologies

- A Contagion is an entity that spreads across a network
- Adoption refers to the event of infection or diffusion. Also known as activation
- Adopters represent the final set of infected nodes
- Final propagation tree obtained by the spread of the infection is known as cascade



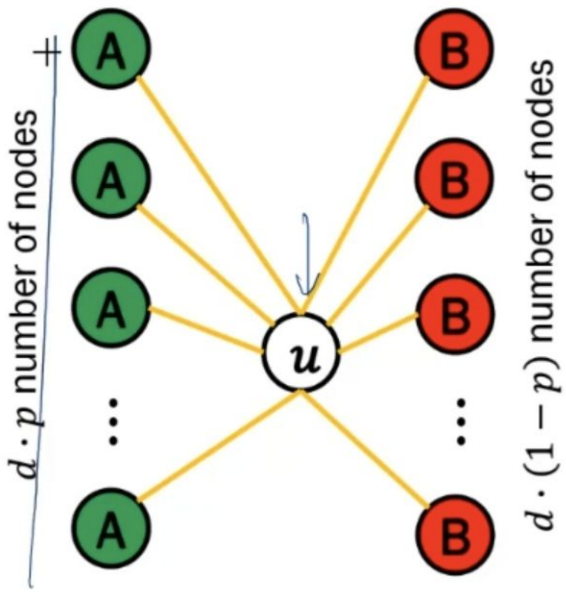
Decision-based Cascade Model: Two-player Coordination Game

u's decision	v's decision	Payoff
A	A	a*
B	B	b*
A	B	0
B	A	0

Payoff distribution for different adoption strategies
* a and b are positive constants

- A and B: two possible strategies that each node in network $G(V, E)$ could adopt
- Each node u will play its own independent game
- Final payoff is the sum of payoffs for all the games
- To calculate the required threshold at which a node u would decide to go with strategy A

Decision-based Cascade Model: Two-player Coordination Game



- Node u has d neighbours
  - p fraction of neighbours adopt strategy A
  - Rest adopts strategy B
- Total payoff for node u if it goes with strategy A $= a \cdot d \cdot p$
- Total payoff for node u if it goes with strategy B $= b \cdot d \cdot (1-p)$
- Node u would adopt contagion A if

$$p \geq \frac{b}{a+b}$$

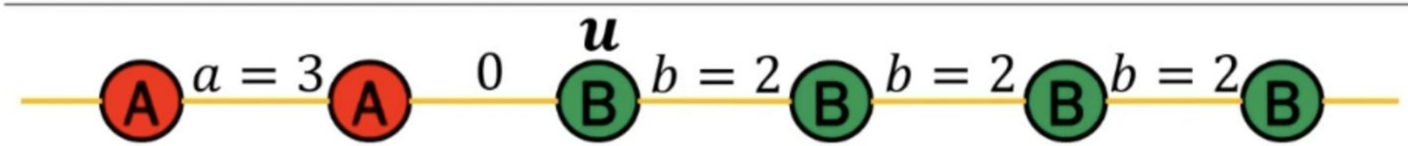
Multiple Choice Decision-based Cascade Model

- Allows a node to adopt more than one strategy/behavior
- In case a node prefers to go with both the strategies A and B, it would incur an additional cost c
- The revised payoff distribution:

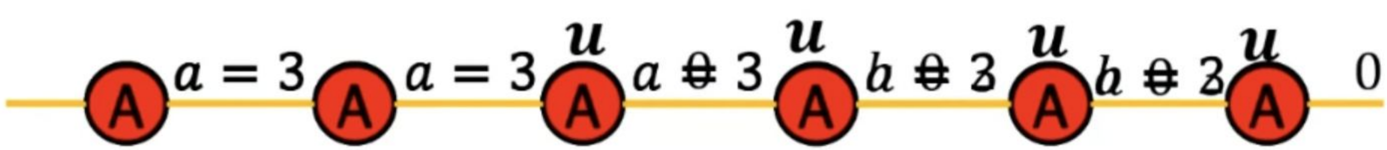
u's decision	v's decision	Payoff
AB	A	$a*$
AB	B	$b*$
AB	AB	$max(a,b)$

Payoff for a multiple choice decision model
* a and b are positive constants

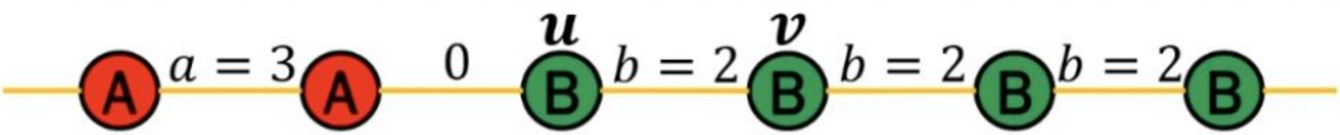
Cascades for Infinite Chain Networks: Single Choice



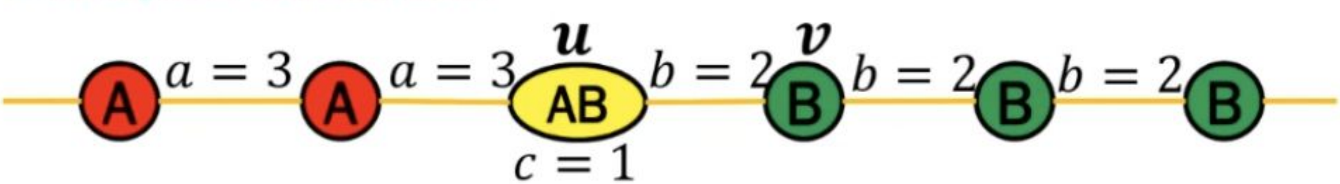
- Consider the case: $a = 3, b = 2$
- Two possible choice for node u
  - Stick with strategy B, total payoff: 0 + 2 = 2
  - Switch to strategy A, total payoff: **3 + 0 = 3**
- So, node u would adopt strategy A
- And the cascade continues...



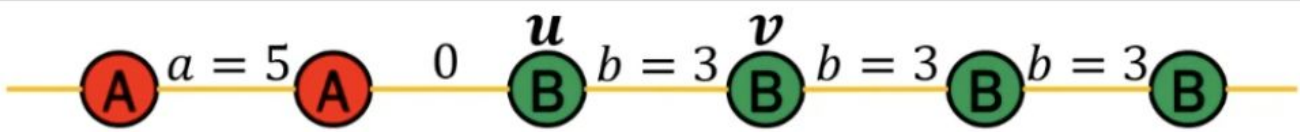
Cascades for Infinite Chain Networks: Multiple Choice: Case I



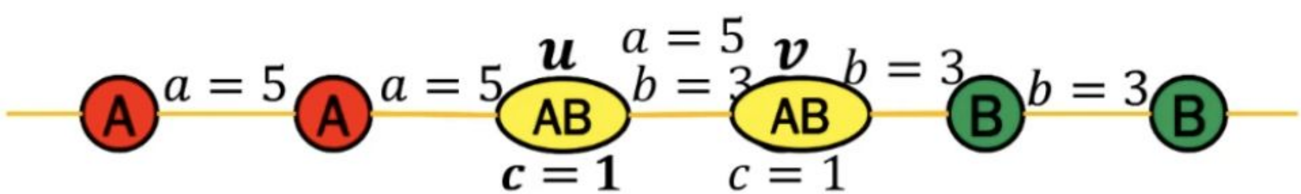
- Consider the case: $a = 3, b = 2, c = 1$
- Two possible choice for node u
  - Stick with strategy B, total payoff: 0 + 2 = 2
  - Switch to strategy A, total payoff: 3 + 0 = 3
  - Switch to strategy AB, total payoff: **3 + 2 – 1 = 4**
- So, node u would adopt strategy AB
- And system is stable now!!



Cascades for Infinite Chain Networks: Multiple Choice: Case II



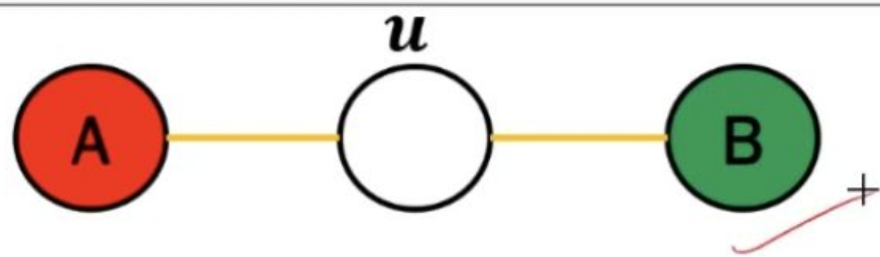
- Consider the case: $a = 5, b = 3, c = 1$
- Two possible choices for node u
  - Stick with strategy B, total payoff: 0 + 3 = 3
  - Switch to strategy A, total payoff: 5 + 0 = 5
  - Switch to strategy AB, total payoff: **5 + 3 – 1 = 7**
- So, node u would adopt strategy AB
- Two possible choices for node v
  - Stick with strategy B, total payoff: 3 + 3 = 6
  - Switch to strategy A, total payoff: 5 + 0 = 5
  - Switch to strategy AB, total payoff: **5 + 3 – 1 = 7**
- So, node v would adopt strategy AB
- And the cascade continues!!



Cascade in Infinite Chain Networks: Generic Model

- Let us consider an infinite chain network with strategy set $\{A, B, AB\}$
- We consider the scenario: $a = a, b = 1, c = c$
- Two possible cases may arise:

Generic Model: Case A



❑ Three possible options for node u

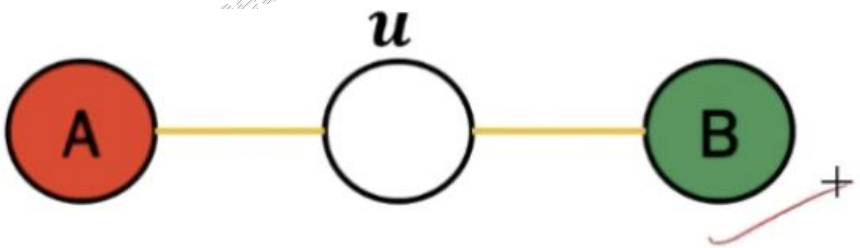
1. Adopt Behavior A; Payoff $= a + 0 = a$
2. Adopt Behavior B; Payoff $= 0 + 1 = 1$
3. Adopt Behavior AB; Payoff $= a + 1 - c$

❑ Breakpoint Equations:

a. B versus A: $a = 1, a < 1$: Prefer strategy B; $a > 1$: Prefer strategy A

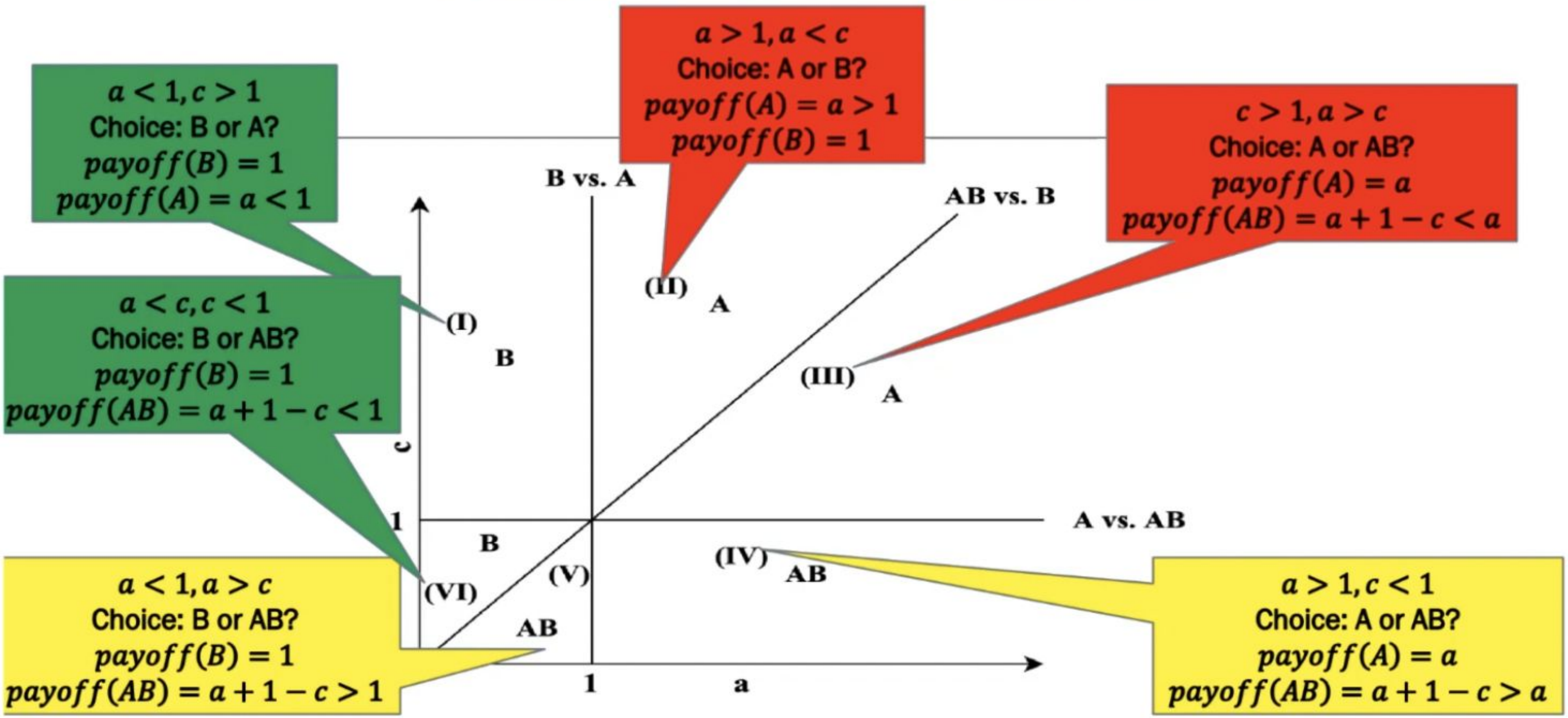
b. AB versus B: $a = c, a < c$: Prefer strategy B; $a > c$: Prefer strategy AB

c. A versus AB: $c = 1, c < 1$: Prefer strategy AB; $c > 1$: Prefer strategy A

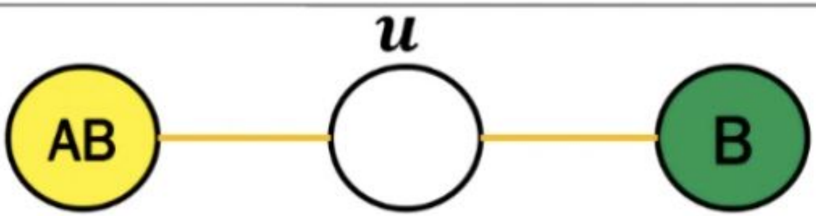


- Three possible options for node u
  1. Adopt Behavior A; Payoff $= a + 0 = a$
  2. Adopt Behavior B; Payoff $= 0 + 1 = 1$
  3. Adopt Behavior AB; Payoff $= a + 1 - c$

Generic Model: Case A



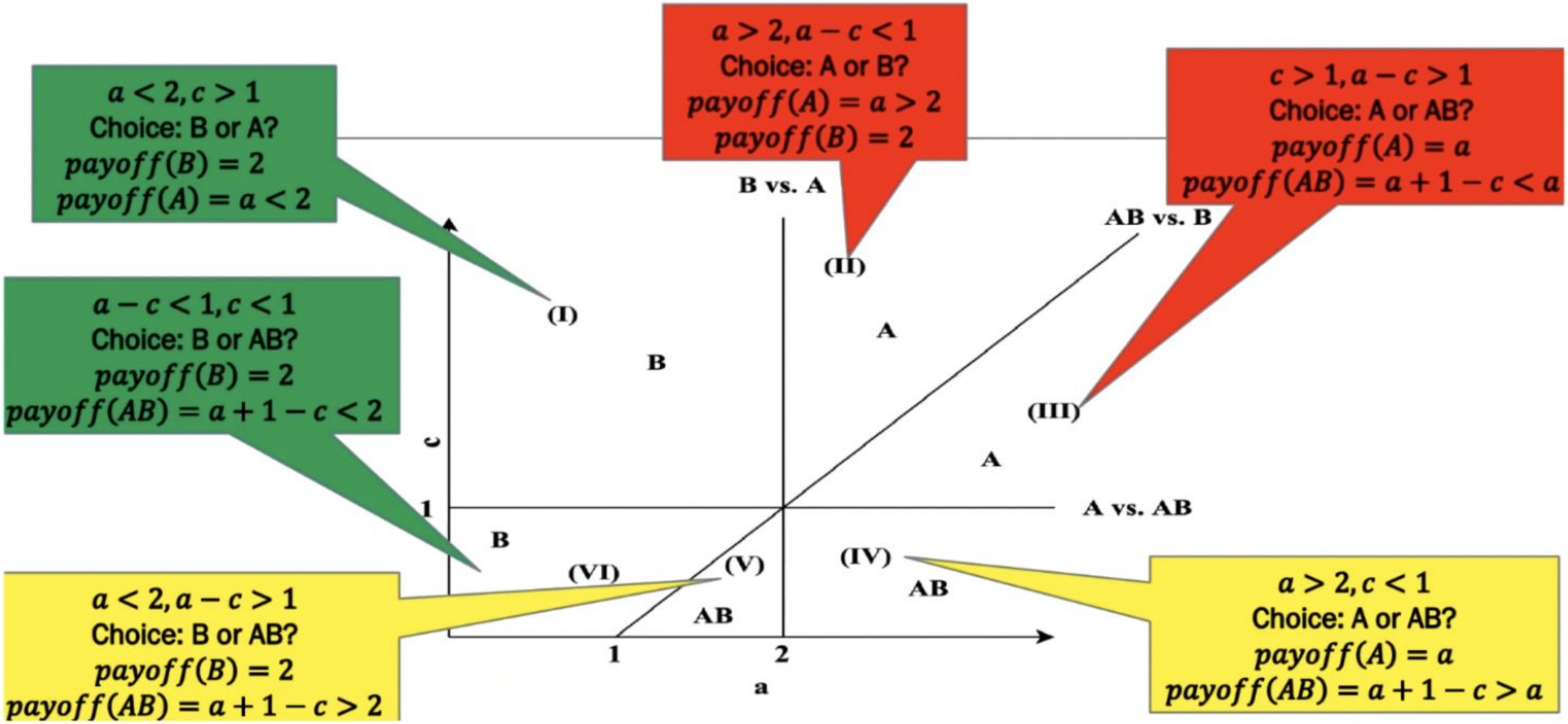
Generic Model: Case B



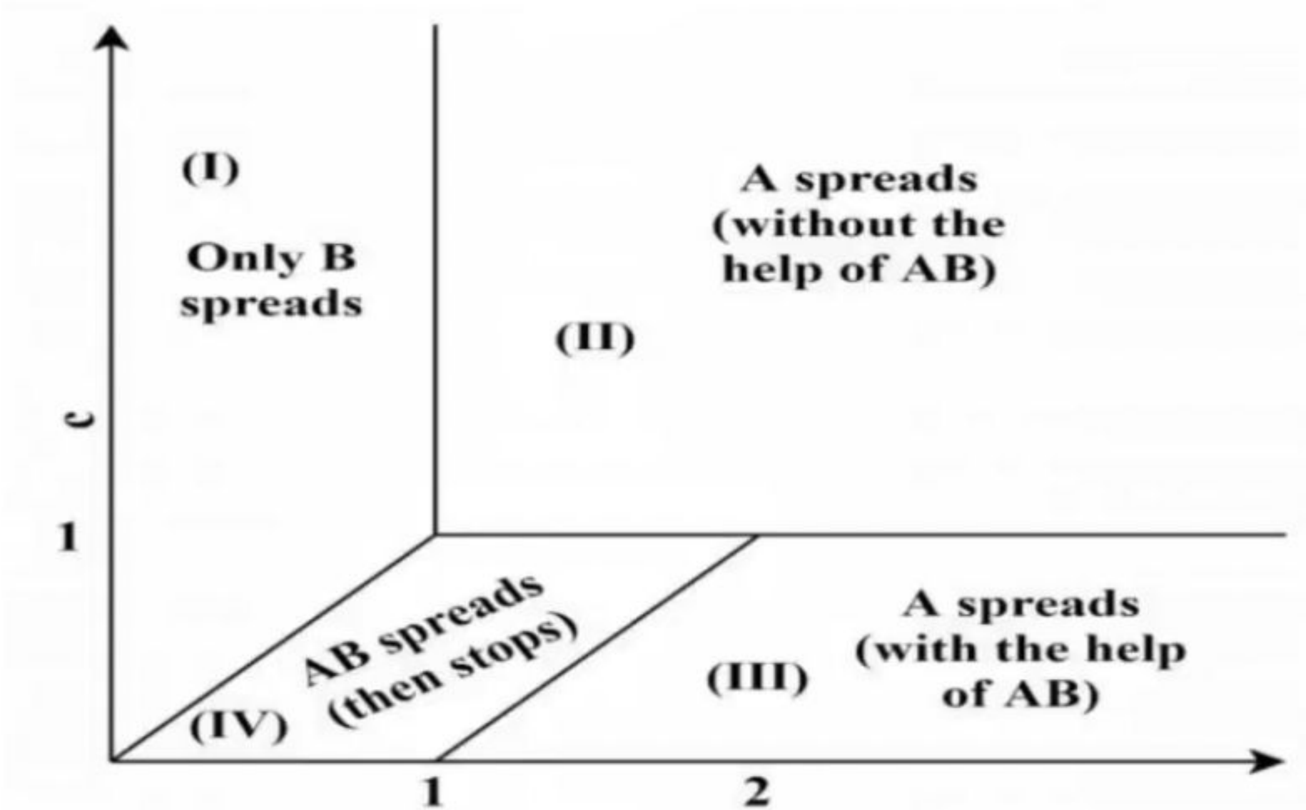
- Three possible options for node u
  1. Adopt Behavior A; Payoff $= a + 0 = a$
  2. Adopt Behavior B; Payoff $= 1 + 1 = 2$
  3. Adopt Behavior AB; Payoff $= a + 1 - c$, if $max(a, 1) = a$

- Breakpoint Equations:
  a. B versus A: $a = 2, a < 2$: Prefer strategy B; $a > 2$: Prefer strategy A
  b. AB versus B: $a - c = 1, a - c < 1$: Prefer strategy B; $a - c > 1$: Prefer strategy AB
  c. A versus AB: $c = 1, c < 1$: Prefer strategy AB; $c > 1$: Prefer strategy A

Generic Model: Case B



Generic Model: Combined



Decision-based Cascade Model: Limitations

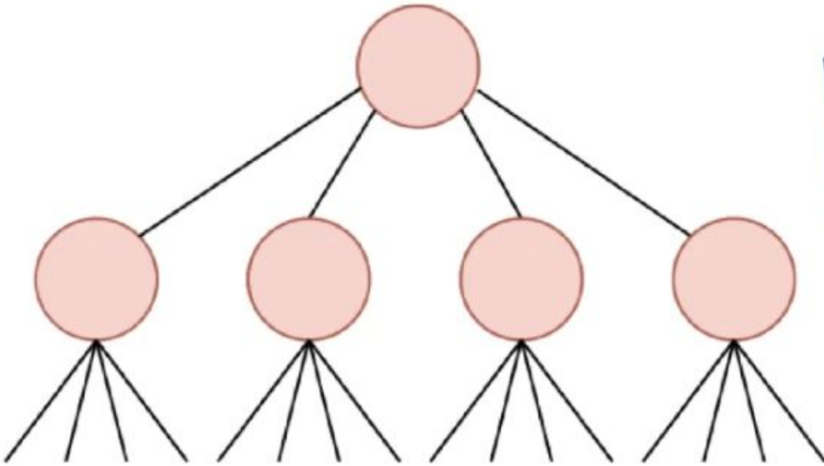
- Cascade will continue to grow only when its growth is associated with the highest reward amongst each of the nodes

- In many real-world scenarios, such hard decision making criteria or payoff functions are not available

- Infection spreading mechanism of a virus
  - We can model the spread of the virus as a cascade
  - Cascade growth is not in the hands of the node.

- Alternative Approach (Probabilistic Cascade Model)

Probabilistic Cascade Model: Random Tree



A random tree with $d = 4$

- Basic assumptions
  - Person at the root node of the random tree is always infected
  - Each person in the random tree meets d new people. So, the random tree is a d-nary tree
  - Each person, on meeting an infected person, has the probability of getting infected as $q \ (q > 0)$
- For the virus to stay active and keep on spreading (cascade)
  - probability that a node at a depth h will be infected should be a positive real number
  - Same must hold for all h

$$\lim_{h\to\infty} P[a\ node\ at\ depth\ h\ is\ infected] > 0$$

- The cascade would die out if

$$\lim_{h\to\infty} P[a\ node\ at\ depth\ h\ is\ infected] = 0$$

Probabilistic Cascade Model: Random Tree

- If p_h be the probability of a node being infected at level h, then

$$p_h = 1 - (1 - q \cdot p_{h-1})^d$$

- The recurrence relation can have the following functional form:

$$f(x) = 1 - (1 - qx)^d$$

- The properties of f
  - $f(x)$ is monotonic function
  - $f'(x)$ is non-increasing
  - $f'(x)$ is monotonic, non-increasing in $[0,1]$
- $f(0) = 0$ and $f'(0) = q \cdot d$

Probabilistic Cascade Model: Random Tree

- Since $f'(x)$ is monotonic non-increasing, $f'(x) \leq q \cdot d$
- For epidemic to die out, $f(x) < x \Rightarrow q \cdot d < 1$
- The quantity $q \cdot d$ is called Reproductive number in the literature, denoted R_0
- If $R_0 \geq 1$, the epidemic grows in an exponential manner
- If $R_0 < 1$, the epidemic spread reduces constantly and eventually dies out

- two methods to contain the spread of the epidemic
  - reduce the value of d ⇒ keep the already-infected nodes in isolation
  - reduce the value of q ⇒ reduce transmission rate by promoting better hygiene practices

R_0 shows how quickly a disease spreads.

- Measles: ~12–18 (very contagious)
- COVID-19 (early): ~2–3

Anomaly detection

Famous Survey papers, Books

- Graph based Anomaly Detection and Description: A Survey (600+ citations)

- Anomaly detection in dynamic networks: a survey (~200 citations)

Introduction

- We are drowning in the deluge of data that are being collected world-wide, while starving for knowledge at the same time.
- Anomalous events occur relatively infrequently
- However, when they do occur, their consequences can be quite dramatic and quite often in a negative sense

"Mining needle in a haystack. So much hay and so little time"

Introduction

- Anomaly is a pattern in the data that does not conform to the expected behaviour.
- Also referred to as **outliers**, **exceptions**, **peculiarities**, **surprises**, etc.
- The branch of data mining concerned with discovering rare occurrences in datasets is called ***anomaly detection***.
- This problem domain has numerous high-impact.

Real World Anomalies

- Credit Card Fraud
  - An abnormally high purchase made on a credit card

- Cyber Intrusions
  - A web server involved in *ftp* traffic

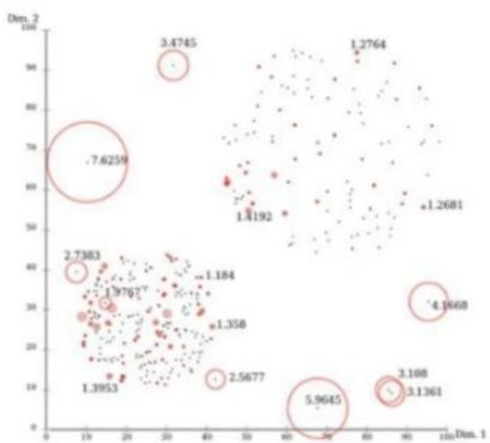
- Fake followers/retweers
  - Blackmarket based activities

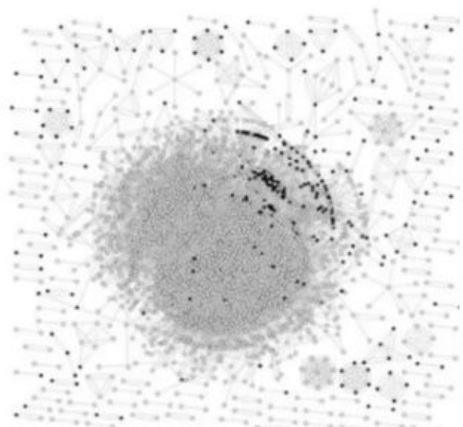
Other Applications

- Calling card and telecommunications fraud
- Auto insurance fraud
- Email and Web spam
- Opinion deception and reviews spam
- Auction fraud
- Tax evasion
- Customer activity monitoring and user profiling
- Click fraud
- Securities fraud
- Malware/spyware detection
- False advertising
- Image/video surveillance

Outliers vs. Graph Anomalies

- Most outlier detection techniques treat objects as points lying in a multi-dimensional space independently.
- In contrast, they may exhibit ***inter-dependencies***





In a reviewer-product review graph, the extent a reviewer is fraudulent depends on what ratings s/he gave to which products

- as well as how other reviewers rated the same products

Challenges/Opportunities: Ill-defined problem

- No unique definition for the problem of anomaly detection exists.
- The definition becomes meaningful only under a **given context or application.**

Definition: (Hawkins' Definition of Outlier, 1980)
"An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism."