Social Media Analytics

BITS Pilani

Garima Jindal
Faculty Department

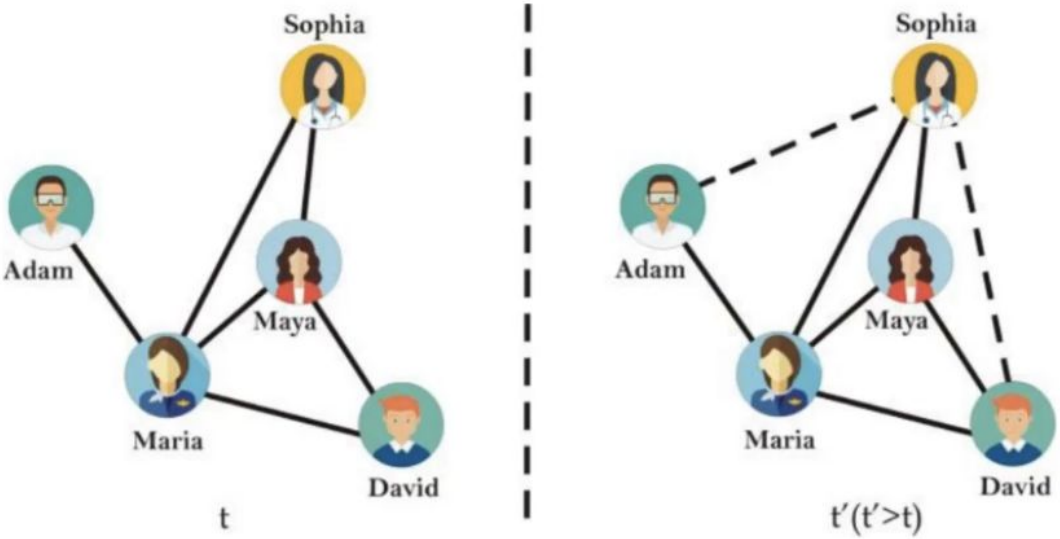
innovate achieve lead

BITS Pilani

Social Media Analytics
Lecture No. 10

What is Link Prediction?

- The problem of predicting the existence of a link between two entities in a network
- Involve several research communities ranging from statistics and network science to machine learning and data mining
- Help in predicting the state of a dynamic network at future timestamp



https://www.nature.com/articles/s41598-019-57304-y

Application Areas

Online Social Networks

- Recommend friends to connect
- Suggest users/pages to follow

E-commerce

- Recommend products/services

Police/Military

- Identify hidden groups of terrorists
- Spot criminals in security related applications

Bioinformatics/Biology

- Predict protein-protein interactions
- Infer interactions between drugs and targets

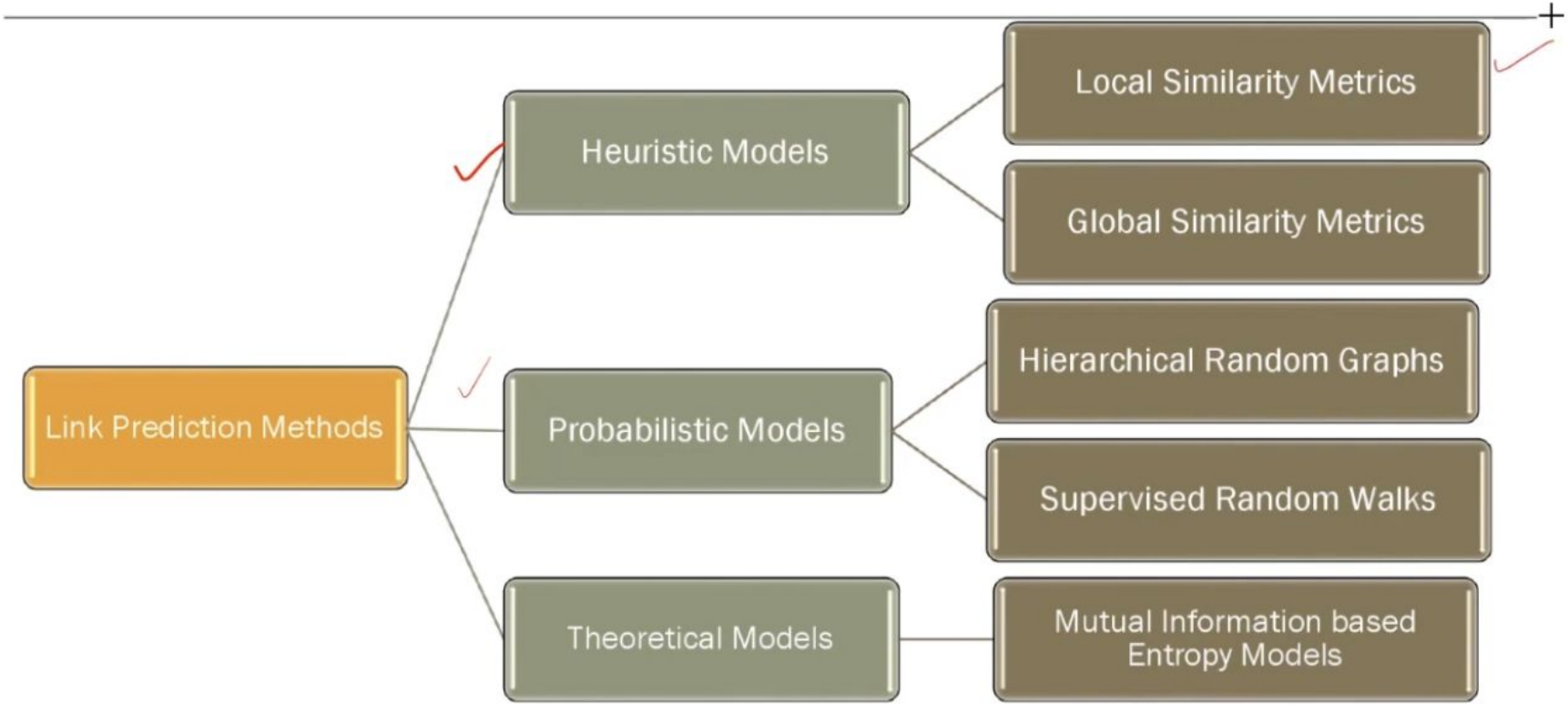
Network Reconstruction

- Remove spurious edges
- Predict missing links
- Predict new links

Citation Networks

- Predict missing citations
- Predict future collaboration

Link Prediction Methods



Link Prediction: Local Heuristic

- $G(V, E)$: an undirected dynamic network
- Three nodes $x, y, z \in V$ such that, at the current time instance
  - $(x, z) \in E$, $(y, z) \in E$
  - $(x, y) \notin E$
  - To decide the formation of the link (x, y) in near future
- Some local structural similarity base heuristic for the above
  - Common Neighbourhood
  - Jaccard Similarity
  - Preferential Attachment
  - Adamic Adar
  - Salton Index
  - Hub Promoted Index

Agenda

1. **Global Link Prediction Methods**
    - Motivation and overview
    - Examples: Katz Score, PageRank-based methods
2. **Probabilistic Methods**
    - Hierarchy and community prediction
3. **Information Diffusion**

What is a *global heuristic–based* link prediction algorithm?

A **global heuristic–based link prediction algorithm** predicts missing or future links by using **the entire structure of the network**, not just local neighborhoods.

- It assigns a **score** to every pair of nodes
  Higher score ⇒ higher likelihood of a link

"Global" = uses **paths, walks, or matrix operations over the whole graph**

"Heuristic" = rule-based, not learned from labeled data

What is the Katz score of an edge?

The **Katz score** is a **link prediction measure** used in networks.

It measures how likely **two nodes should be connected**, based on:

- **All possible paths** between them
- **Short paths matter more than long ones**

Global Heuristic: Katz Score

- Inspired by Katz centrality
- Takes into account the influence by neighbors beyond 1-hop
- However, longer the path length, less likely the end nodes influence each other
- Between two random nodes x and y, Katz score is given by

$$S_{KZ}(x,y) = \sum_{p=1}^{\infty} \alpha^p \cdot A_{x,y}^p$$

Here, $A_{x,y}^p$: number of paths of length p that exists between x and y

and α: damping factor that reduces the impact of longer paths

How do we predict a link from Katz score?

Step-by-step logic

1. **Compute Katz scores** for all *non-connected* node pairs

2. **Rank pairs by score** (highest → lowest)

3. **Predict links** for:
   - Top-k pairs
   - Or pairs with score > threshold

Global Heuristic: Hitting Time

- Based on random surfing model. A random surfer
  a) starts at node x
  b) moves to a neighbor of x chosen uniformly at random
  c) repeats step (a) till it reaches y
- Hitting time (HT_{xy}): Expected number of steps it takes for a random surfer starting at x to reach y
- The Hitting Time score between nodes x and y is given by

$$S_{HT}(x, y) = -HT_{xy}$$

- Smaller the hitting time between two nodes, closer in proximity the nodes, therefore higher the chances of their interaction in future
- The Normalized Hitting Time score between nodes x and y is given by

$$S_{HT}^{Norm}(x, y) = -HT_{xy} \cdot \pi_y$$

Here π: stationary distribution of PageRank for the network

Global Heuristic: Commute Time

- Extending upon the random walk model. A random walker
  a) starts at node x
  b) moves to a neighbor of x chosen uniformly at random
  c) repeats step (a) till it reaches y
  d) travels back to x (not simply jumps to x)

- The Commute Time score between nodes x and y is given by

$$S_{CT}(x, y) = -C_{xy} = -(HT_{xy} + HT_{yx})$$

- Smaller the commute time between two nodes, closer in proximity the nodes, therefore higher the chances of their interaction in future

- The Normalized Commute Time score between nodes x and y is given by

$$S_{CT}^{Norm}(x, y) = -(HT_{xy} \cdot \pi_y + HT_{yx} \cdot \pi_x)$$

Here π: stationary distribution of PageRank for the network

Local methods are **fast** and simple but limited, while global methods are powerful and accurate but **computationally costly.**

Aspect	Local Methods	Global Methods
Information used	Neighbor-level	Whole network
Path length	Short (1–2 hops)	All paths
Accuracy	Moderate	Higher
Computational cost	Low	High
Scalability	Excellent	Limited

Probabilistic Link Prediction Methods

What is a probabilistic method?

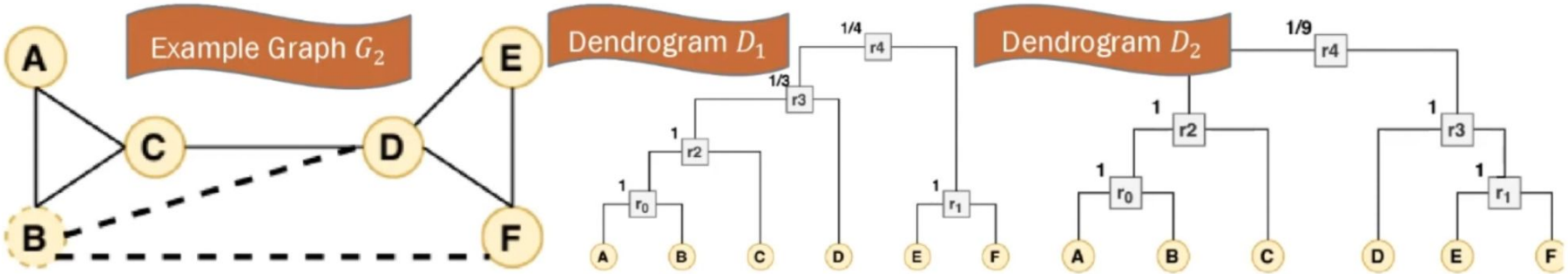
- Models **link existence as a probability**
- Assumes a **generative process** for how links are formed
- Uses **statistical or Bayesian models**
- Outputs P(link between u,v)

Probabilistic Models: Hierarchical Networks

- A network is said to be a hierarchical network if
  - the vertices can be divided into groups,
  - each of these groups can further be subdivided into groups of groups, and so on
  - each group formed in a logical order corresponding to a granular functional/social unit
- Can easily be rendered as a tree or a dendrogram: Nodes of a network form the leaves of the dendrogram
- Smaller the height of the links between the groups or the nodes, the higher the similarity between them

Probabilistic Models: Dendrograms

- Dendrogram D for a graph $G(V, E)$ with n nodes is
  - a tree with n leaves (nodes of the G) and $n-1$ internal nodes $(r_0, r_1, \cdots, r_{n-1})$
  - each internal node corresponds to the group of vertices that directly descent from it
- Each internal node r has an association probability p_r
  - how likely two nodes/groups are to form a connection, given r as their least common ancestor



Probabilistic Models: Dendrograms

- E_r: number of edges in G whose endpoints has r as the lowest common ancestor in D
- L_r and R_r: Number of leaves in the left and in the right subtrees of r, respectively
- Probability that each of the original edges aggregated in E_r forms a connection in D $= p_r^{E_r}$
- Probability of success for internal node $r = p_r^{E_r}(1-p_r)^{L_r R_r - E_r}$

Probabilistic Models: Dendrograms

- Likelihood of the hierarchical graph:

$$\mathcal{L}(D, p_r) = \prod_{r \in D} p_r^{E_r} (1 - p_r)^{L_r R_r - E_r}$$

- Successive application of log likelihood, partial differentiation, and equating to zero yields

$$p_r^* = \frac{E_r}{L_r R_r}$$

Difference between heuristic based methods and probabilistic methods

Probabilistic methods are more **expressive and interpretable** than heuristic methods, but they are computationally more expensive and require model assumptions.

Aspect	Heuristic Methods	Probabilistic Methods
Output	Score	Probability
Uncertainty	❌	✅
Latent variables	❌	✅
Attribute usage	Limited	Natural
Interpretability	Medium	High
Theory	Heuristic	Statistical

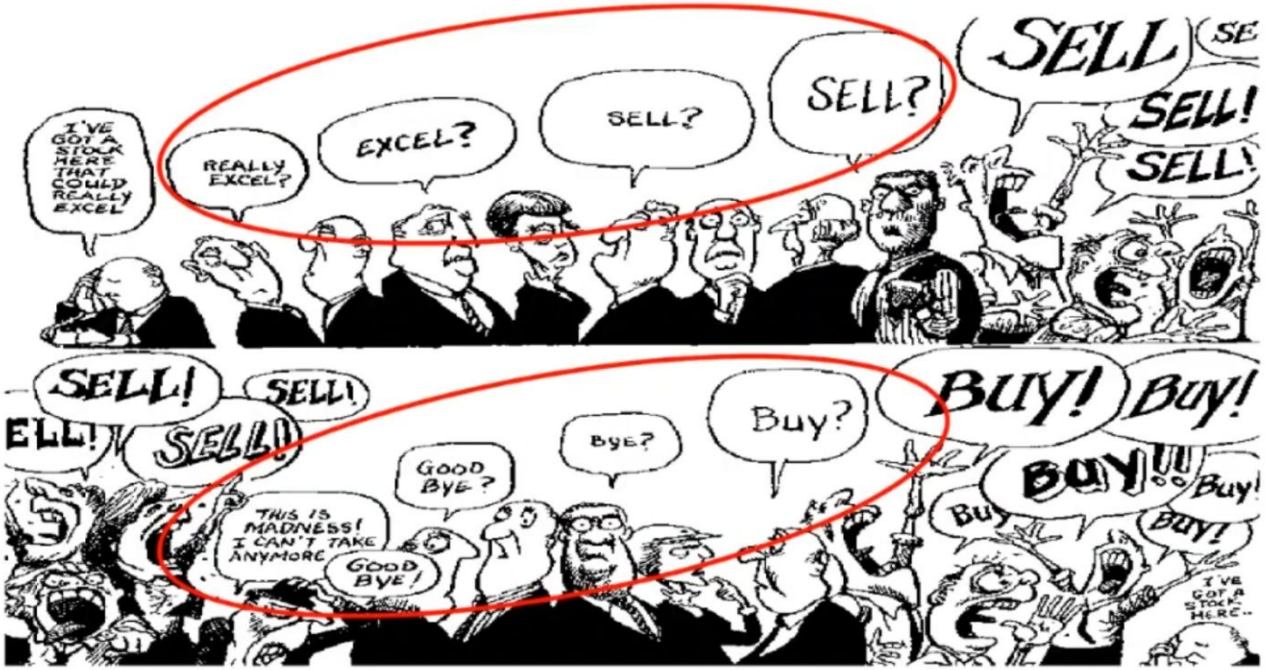
Information Diffusion

- Diffusion is the net movement of anything from a region of higher concentration to a region of lower concentration
- Driven by a gradient in concentration
- Information Diffusion is the process by which information is spread from one place to another through interactions
- Diffusion process involves three main elements:
  - Sender: An entity (or a group of entities) responsible for initiating the diffusion process
  - Receiver: An entity (or a group of entities) receives the diffusion information from the sender(s)
  - Medium: The channel through which the diffusion information is sent from the sender(s) to the receiver(s).

Cascade Behavior: Real-world Instances

- Healthcare
  - Disease Propagation
  - Epidemic Spreading
- Socio-political Cascades
  - Arab Spring Movement in 2010 – 2012
  - From small protests began in Leipzig to Fall of the Berlin Wall in 1989
  - #MeToo movement against sexual abuse and sexual harassment
- Financial Market Cascades
  - Market Bubble
  - A stock becomes overly popular among investors
  - Viral marketing
- Social network
  - Rumor spread
  - Belief Spread
  - Fake News virality

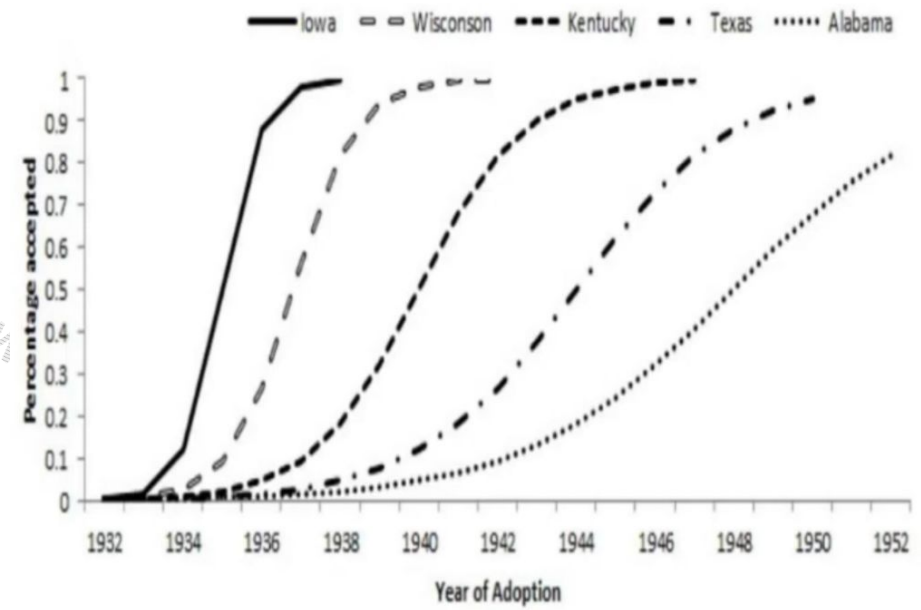
Information Diffusion: Information Cascade



https://blogs.cornell.edu/info2040/2011/11/17/ipos-and-information-cascades/

- An Information cascade is a phenomenon in which a number of people make the same decision in a sequential fashion.
- The phenomenon is found widely in behavioral economics and network theory
- Similar to, but not identical to herd behavior

Information Diffusion: Diffusion of Innovations



Adoption of hybrid seed corn by farmers in USA

http://homepage.cs.uiowa.edu/~sriram/196/spring12/lectureNotes/Lecture15.pdf

- Success/failure of an innovation is highly guided by the structure of the network formed by the initial adopters
- Adoption of hybrid seed corn by farmers
- Influenced by their neighbors in the community
- Adoption of a new drug by the doctors
- Assurance from social peer connections

Information Diffusion: Herd Behavior

http://www.kiffingish.com/2013/12/false-sense-of-security.html

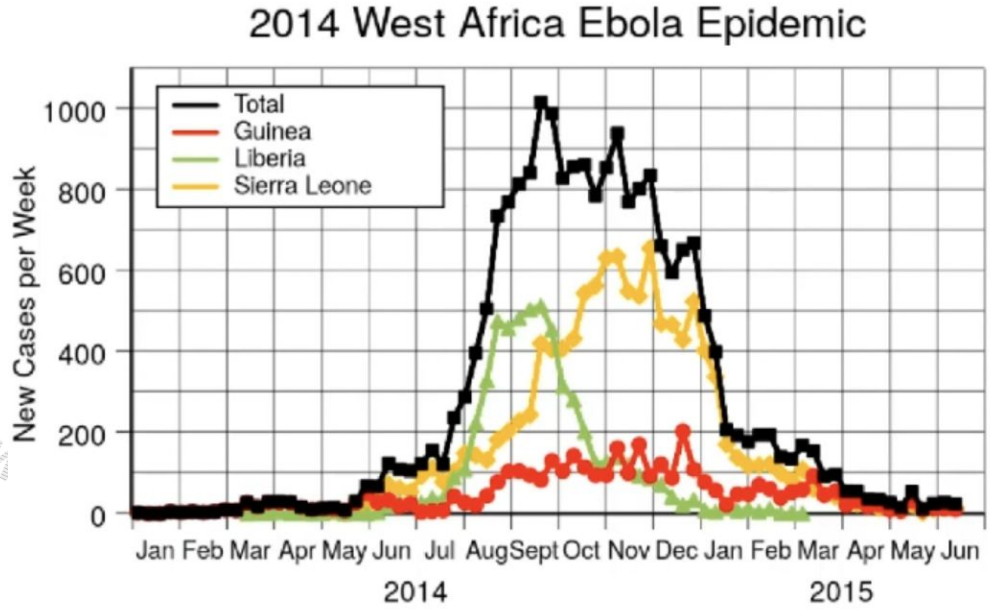
- The behavior of individuals in a group acting collectively without centralized direction
- Human based herd behaviour: demonstrations, riots, general strikes, religious gatherings, judgement and opinion-forming, etc.
- Often a useful tool in marketing; if used properly, can lead to increases in sales
- Herding behavior turns violent sometimes, particularly when confronted by an opposing ethnic or racial group

Information Diffusion: Echo Chambers

https://theconversation.com/the-problem-of-living-inside-echo-chambers-110486

- Situations in which beliefs are amplified or reinforced by communication and repetition inside a closed system
- A harmonious group of people amalgamate and develop tunnel vision
- Social communities become fragmented by echo chambers
- Causes powerful reinforcements of rumors and fake news due to the unchallenged trust in the evidence supplied by their peers

Information Diffusion: Epidemics

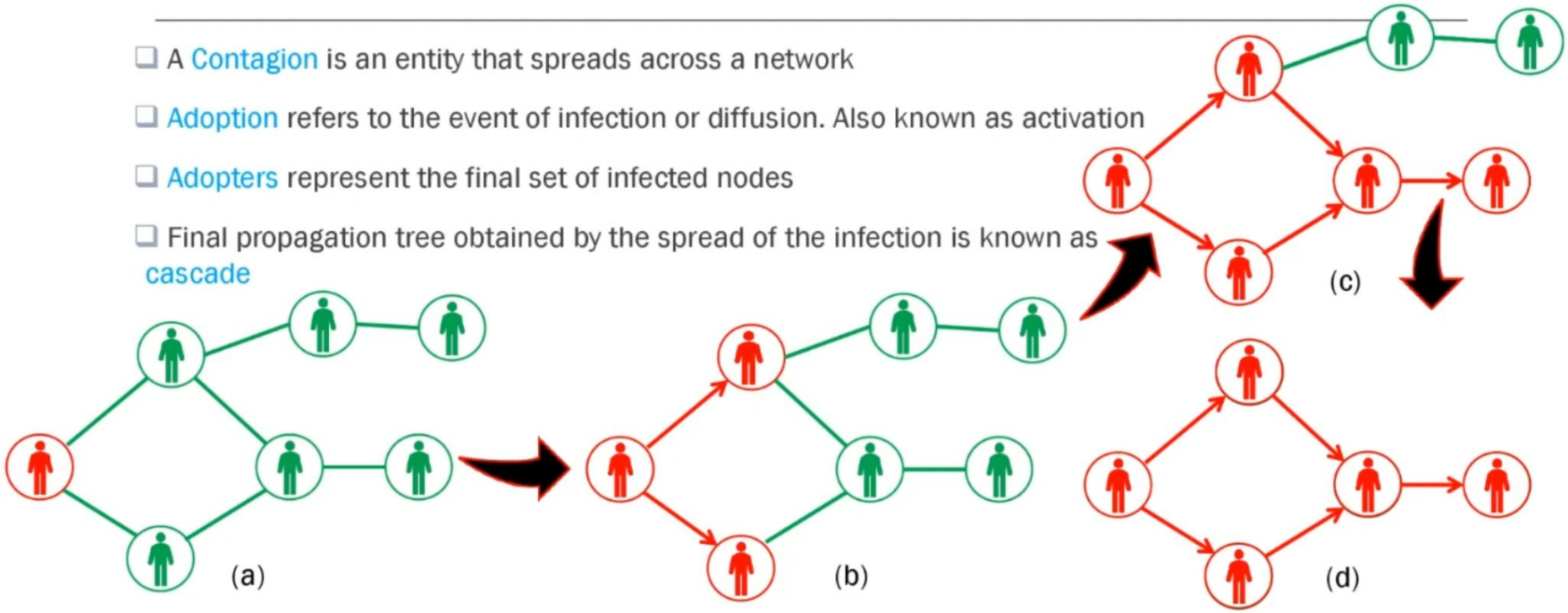


https://commons.wikimedia.org/wiki/File:2014_West_Africa_Ebola_Epidemic_-_New_Cases_per_Week.svg

- Rapid spread of disease to a large number of people in a given population within a short period of time
- Epidemic models are similar to diffusion of innovation models
- Only difference is: individuals do not decide whether to become infected or not

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



Cascade Model: Decision-based Model

- Given a network, each node has the freedom to decide whether to adopt a contagion or not
- Originated from the idea of local interaction models described by Morris in 2000
- Decision at each node is influenced by the behavior of nodes in its neighborhood
- Nodes decide to adopt a new contagion driven by a direct benefit or payoff
- The payoff by adopting a contagion is directly proportional to the number of its neighbors that have adopted the same contagion
- Can be explained using a two-player coordination game
  - Given a number of strategies, the end goal of the players is to coordinate on the same strategy to maximize their payoffs

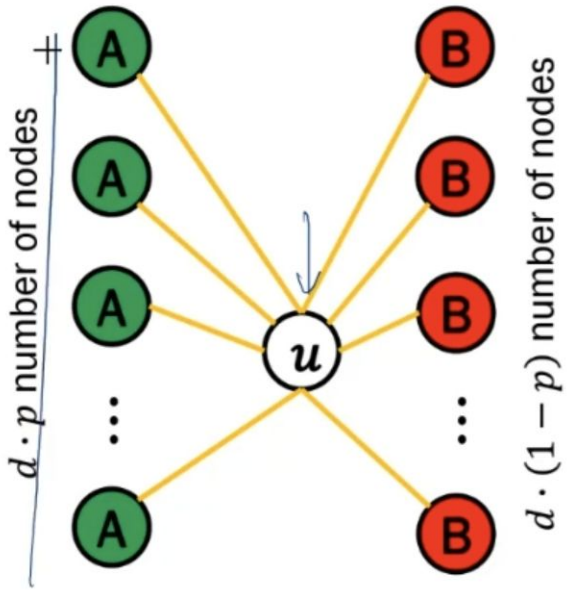
Decision-based Cascade Model: Two-player Coordination Game

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

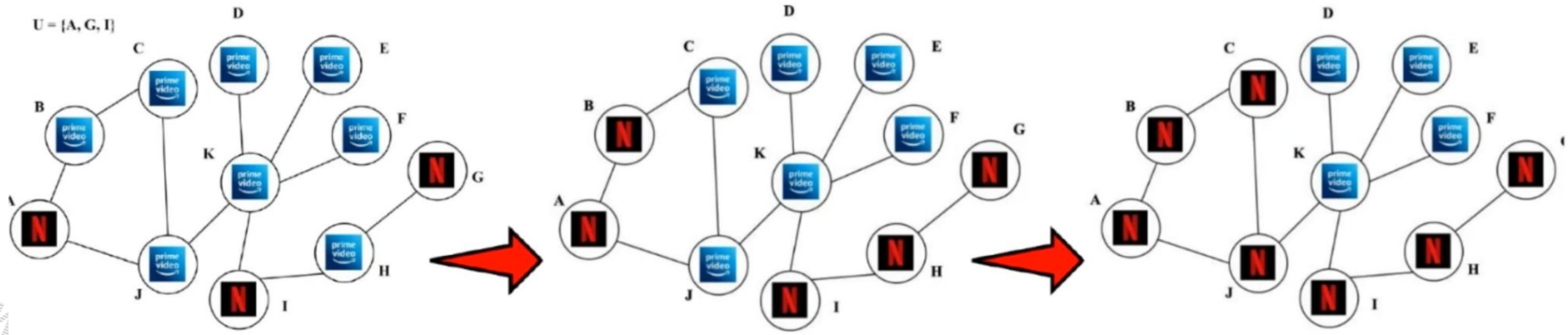
Decision-based Cascade Model: Two-player Coordination Game



- 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

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

Decision-based Cascade Model: Illustration



The threshold for a switch from Amazon Prime Video to Netflix at a node is 0.50

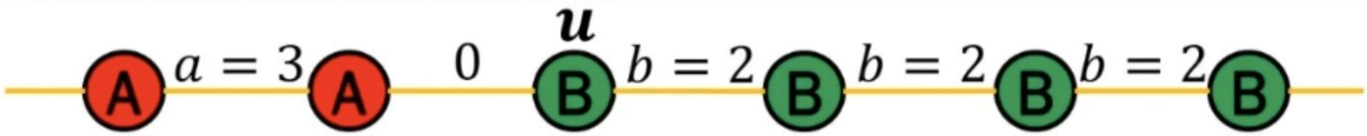
Multiple Choice Decision-based Cascade Model

- 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...