Information Diffusion

Diffusion, Assortativity & Influence

— SL08 —

Philippe Cudré-Mauroux

pcm@unifr.ch

Table of contents — SL08

OVERVIEW

- Information diffusion:
  - Herd Behavior
  - Information Cascades
  - Diffusion of Innovations
  - Epidemics
- Assortativity
- Influence

Information Diffusion on Social Media (1/2)

- Example: Super Bowl XLVII blackout
- Tweet from cookie brand

- SuperBowl ads cost around $4 million for 30s
- Cookie brand got similar attention basically for free
  - 15k retweets, 10k likes, media exposure...
  - Wired: "How Oreo won the marketing Super Bowl"

Information Diffusion on Social Media (2/2)

- Information diffusion: process by which a piece of information (knowledge) is spread and reaches individuals through interactions.
- Information diffusion is a research arrea borrowing from multiple fields
  - Sociology, epidemiology, ethnography...
- Diffusion process typically involve three kinds of entities:
  - i) senders ii) receivers iii) a medium
- Today's focus: techniques that can model information diffusion.

Four Models of Diffusions

- Explicit network:
  - Herd behavior (individuals observe the actions of all others and act in an aligned form with them)
  - Information cascades (individuals observe their immediate neighbors)
- Implicit network:
  - Diffusion of Innovation (bird's-eye view of how an innovation spreads through a population assuming that interactions among individuals are unobservable)
  - Epidemics (infection is considered a random natural process where individuals are exposed to a pathogen)

TABLE OF CONTENTS — SL08

Herd Behavior

- Example: online auction
  - Individuals are connected through the auctions that are public
  - Individuals sometimes bid on items that might otherwise be considered unpopular as they trust others and assume that the high number of bids that the item received is a strong signal of its value
- Further example: choosing restaurant based on current attendance
- Herd behavior describes when a group of individuals performs actions that are aligned without previous planning
  - It has ben observed in flocks, herds, and in humans during sportings events, demonstrations, religious gatherings, etc.
  - It requires i) connections between individuals and ii) a method to transfer behavior among individuals or to observe their behavior

Solomon Asch Experiment

- 3% vs 32% of incorrect answers
- Wisdom of the crowd?

DESIGNING A HERDING EXPERIMENT

- Four conditions to satisfy:
  1. Decisions must be made
  2. Decisions must be sequential
  3. Individuals must have private information that helps them decide
  4. Individuals do not know the private information of others but can try to infer them from what they observe
- Example: Opaque urn with three marbles in it
  - Marbles can be blue (B) or red (R)
  - Guarantee to have at least one of of each color (so either BBR or RRB)
  - Students come in turn, pick one marble and check its color in private
  - Then make their prediction for the majority color on a blackboard in public
- When does herd behavior take place?

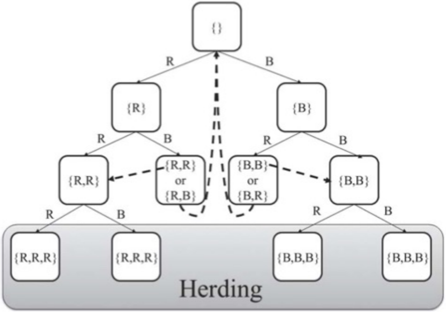
Bayesian Analysis of Herd Behavior (1/2)

- $P(BBR) = P(RRB) = 0.5$
- $P(B|BBR) = P(R|RRB) = 2/3$
- Let's imagine that the first student draws a B
  - $P(B) = P(B|BBR)P(BBR) + P(B|RRB)P(RRB) = 0.5$
  - $P(BBR|B) = P(B|BBR)P(BBR)P^{-1}(B) = 2/3$
  - So first student should rationally predict BBR

Bayesian Analysis of Herd Behavior (2/2)

- Now, imagine that the second student draws B also
- What will the third student predict?
  - $P(B, B, R|BBR) = 2/3 * 2/3 * 1/3 = 4/27$
  - $P(B, B, R) = P(B, B, R|BBR)P(BBR) + P(B, B, R|RRB)P(RRB) = 1/9$
  - $P(BBR|B, B, R) = 2/3$
- So the third student will predict B even if she draws red!
- Similar for all further students (... even if the urn if RRB!)

URN EXPERIMENT



- Blackboard predictions are in rectangles
- Edges represent what students observe

Intervention

- As herding converges to a consensus, one can intervene with the process
  - Typically by disclosing private information to the individuals
  - Example for the urn: i) disclosing the majority or ii) disclosing previous observations

Table of contents — SL08

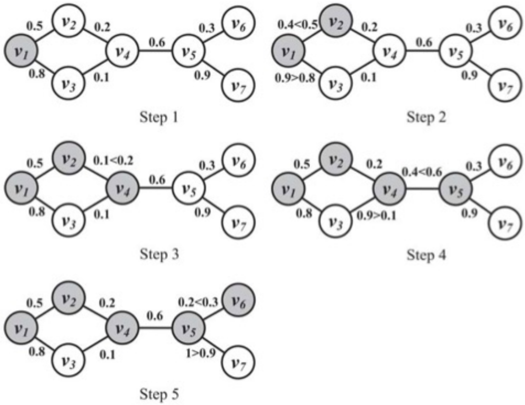
INFORMATION CASCADES

- On social media, individuals commonly repost content posted by others
- An information cascade is a piece of information being cascaded among a set of individuals where
  1. Individuals are connected by a network and
  2. Individuals are only observing decisions of their immediate neighbors
- Cascade users have less information available to them compared to herding users

Independent Cascade Model (ICM)

- Basic model that can help explain information cascades
- Underlying assumptions:
  - Directed graph with actors (nodes) and communication channels (edges)
  - Decisions are binary: nodes can either be active (adopting the behavior) or inactive
  - Once activated, a node can activate its neighbors
  - Activation is progressive: nodes cannot turn inactive once active
- Let v get activated at time t
  - v can activate its neighbors w with a probability $p_{v,w}$ at time $t+1$
  - v cannot activate its neighbors after that

ICM Example



ICM example; number on the edges represent p_{vw} ©SMM

Maximizing the Spread of Cascades

- One interesting question is which nodes to activate such that the final number of activated nodes is maximized?
  - Let S denote the seed set and $f(S)$ the final number of activated nodes
  - ICM is stochastic; it can however be made deterministic by pre-generating all random numbers at the beginning of the process
  - $f(S)$ is monotone: $f(S \cup \{v\}) \geq f(S)$
  - Unfortunately the solution is NP-hard
  - One can get at least a $(1 - 1/e) \approx 0.63$ approximation of the optimal value greedily by iteratively selecting nodes that maximize the total number of nodes being ultimately activated

Maximizing Cascades

Algorithm 7.2 Maximizing the spread of cascades – Greedy algorithm

Require: Diffusion graph $G(V, E)$, budget k

return Seed set S (set of initially activated nodes)

$i = 0;$

$S = \{\};$

while $i \neq k$ **do**

  $v = \arg\max_{v \in V \setminus S} f(S \cup \{v\});$

  `or equivalently` $\arg\max_{v \in V \setminus S} f(S \cup \{v\}) - f(s)$

  $S = S \cup \{v\};$

  $i = i + 1;$

end while

Return $S;$

INTERVENTION

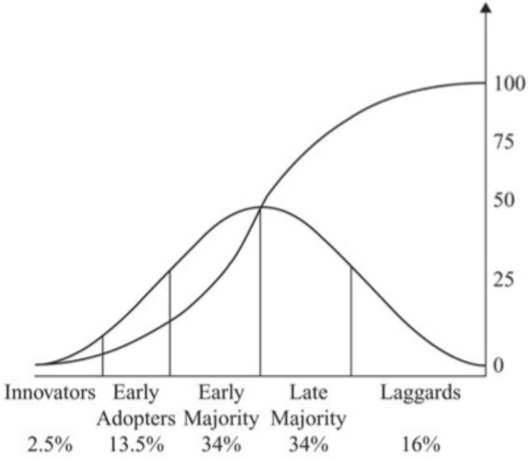
- There are basically three ways of stopping an information cascade (e.g., stopping the spread of a false rumor on social media)
  - Limiting the number of out-links of activated nodes
  - Limiting the number of in-links of inactive nodes
  - Decreasing the activation probability of a node $p_{v,w}$

TABLE OF CONTENTS — SL08

DIFFUSION OF INNOVATIONS

- An innovation is defined as an idea, practice, or object that is perceived as new by an individual
- Diffusion of innovation is a phenomenon that is commonly observed on social networks
  - Video going viral
  - Piece of news being retweeted largely
- Innovations abound, however only few of those largely spread through networks

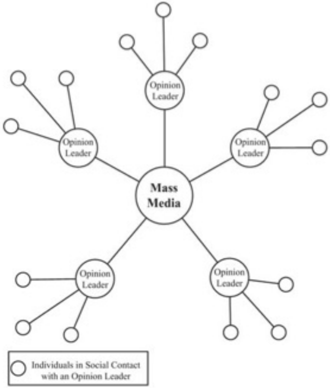
Types of Adopters



Types of adopters and s-shaped cumulative adoption curve ©SMM

Two-Step Flow Model

- Elihu Katz developed a two-step flow model to describe how information gets diffused through mass media



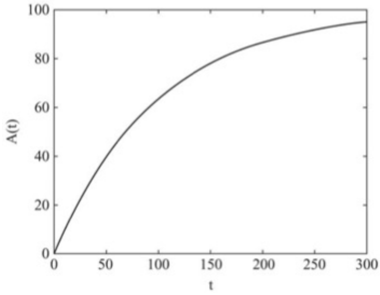
Two-step flow model ©SMM

MODELLING DIFFUSION

- Three notions are of particular importance: $A(t)$: the population that adopted the innovation at time t; P the total population; and $i(t)$ the coefficient of diffusion of the item (innovativeness)
- A simple diffusion model capturing that the rate at which the adopters grow directly depends on innovativeness:
  - $\frac{dA(t)}{dt} = i(t)[P - A(t)]$
  - The adoption rate only affects adopters who have not yet adopted the item
  - $i(t)$ can be defined in various ways depending on the model

External-Influence Model

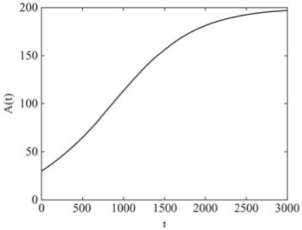
- The coefficient of diffusion is constant
  - Example: diffusion of a breaking news on social media
- $\frac{dA(t)}{dt} = \alpha[P - A(t)]$ which can be solved as
- $A(t) = P(1 - e^{-\alpha t})$



External-Influence for $P = 100$ and $\alpha = 0.01$ ©SMM

Internal-Influence Model

- The adoption depends on how many have adopted the item in the current time step (pure imitation model)
  - Example: peers joining a social networking site
- $\frac{dA(t)}{dt} = \beta A(t)[P - A(t)]$ which can be solved as
- $A(t) = \frac{P}{1+\frac{P-A_0}{A_0} e^{-\beta P(t-t_0)}}$



Internal-Influence for $P = 200$ and $\beta = 10^{-5}$ and $A_0 = 30$ ©SMM

- Mixed-Influence Model: combination of both models

Intervention

- Interventions to stop the diffusion can leverage the three main aspects of the model
  - Limiting the distribution of the item or the audience by reducing the population P
  - Reducing the interest in the item by influencing α
  - Reducing the interactions within the population and thus reducing β

Table of contents — SL08

EPIDEMICS

- In an epidemic, a disease spreads widely within a population
  - The process consists of a pathogen (the disease being spread), a population of hosts (e.g., humans, animals, or plants) and a spreading mechanism (e.g., breathing, drinking, sexual activity)
- Many different ways of modeling epidemics
- Here we assume unknown connections among individuals and unknown process of infection
  - Focuses on global patterns
- Individuals usually do not decide whether to get infected or not

EXAMPLES

- Black Death in the 13th century
  - Plague that decimated more than 50% of Europe's population
- Computer viruses
  - *Stuxnet* infected more than 50% of computers in some countries in 2010

States of Individuals in Epidemics

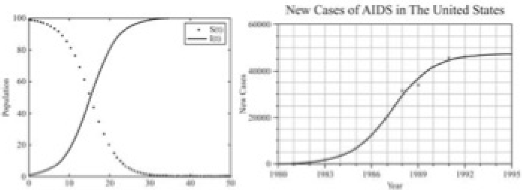
- Susceptible $S(t)$: population that can potentially be infected at time t
- Infected $I(t)$: infected population that can also infect susceptible individuals
- Recovered (or Removed) $R(t)$: population that either recovered (and is now immune) or was killed by the infection (cannot infect others and is not susceptible)
- Total population $N = S(t) + I(t) + R(t)\ \forall t$

SI Model (1/2)

- The most basic epidemic model, SI, consider that infected individuals never get cured
- We assume that the contact probability (prob. of individuals getting in contact) is β and that the disease is propagated with probability 100%
- An infected individual will infect βS individuals at each time step leading to: $dI/dt = \beta IS$, which can be rewritten as $dI/dt = \beta I(N - I)$

SI Model (2/2)

- The solution to this differential equation is called the logistic growth function
- $I(t) = \frac{N I_0 e^{\beta t}}{N + I_0(e^{\beta t} - 1)}$ where I_0 is the number of infected individuals at time 0.

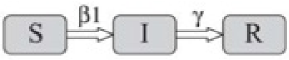


(a) SI Model Simulation (b) HIV/AIDS Infected Population Growth

SI simulation ($N = 100$, $I_0 = 1$, $\beta = 0.003$) compared to HIV growth

SIR Model (1/2)

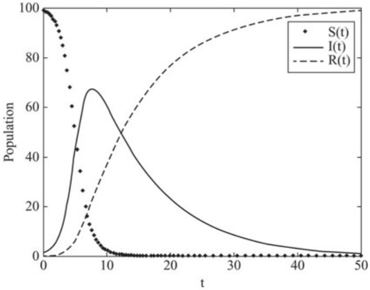
- A second model, SIR, considers as well that infected individuals can recover, with a probability γ
- This yields the following differential equations: $dS/dt = -\beta IS$; $dI/dt = \beta IS - \gamma I$; $dR/dt = \gamma I$.



The SIR model

SIR Model (2/2)

- The differential equations have no closed-form solution but results can be simulated



SIR simulation ($S_0 = 99$, $I_0 = 1$, $R_0 = 0$, $\beta = 0.01$ and $\gamma = 0.1$)

INTERVENTION

- Stopping the epidemic outbreak is usually a pressing question
- A standard solution is to vaccinate the population
  - Reduces the size of the population at risk, and hence of the infected
  - Typically requires that 96% gets vaccinated (herd immunity)
  - *If* we can identify highly-connected nodes, then 30% is enough
- Other techniques such as quarantine work as well

Table of contents — SL08

Assortativity

- Social forces connect individuals in different ways
- One of these ways is assortativity also known as social similarity
  - In assortative networks, similar nodes are connected to one another more often than dissimilar nodes
  - Friendship networks are typically assortative
- Assortativity can be quantified by measuring how similar nodes are connected

US High School Friendship (1994); 80% of the links exist between members of the same race

Measuring Assortativity

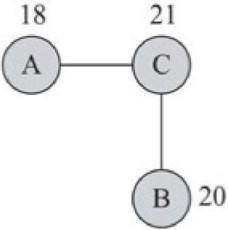
- For nominal attributes (e.g., race, nationality, gender) one simply has to consider the number of edges between nodes of the same type
  - If $t(v_i)$ denotes the type of a node, A the adjacency matrix, m the number of edges and $\delta(x, y)$ is 0 if $x \neq y$ and 1 otherwise then
  - $\frac{1}{2m} \sum_{ij} A_{ij} \delta(t(v_i), t(v_j))$
- A common technique is to subtract the expected assortativity to get the *assortativity significance*
  - The expected number of eges between two nodes v_i and v_j of degree d_i and d_j is $d_i d_j / 2m$; the expected number of edges of the same type is then
  - $\frac{1}{2m} \sum_{ij} \frac{d_i d_j}{2m} \delta(t(v_i), t(v_j))$

MODULARITY

- The resulting measure is called modularity Q
- $Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{d_i d_j}{2m}) \delta(t(v_i), t(v_j))$
- Modularity can be normalized by dividing by its maximum value (when all edges are connecting nodes of the same type)

Assortativity for Ordinal Attributes

- For ordinal values (when there is a clear ordering of the values), we are interested in how correlated are the values of connected nodes
- We construct two variables: X_L representing the ordinal values associated with the *left* node of the edges and X_R for the values of the *right* node of the edges



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix}, \quad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}.$$

Pearson Correlation

- The covariance is then simply $\sigma(X_L, X_R) = E[X_L X_R] - E[X_L]E[X_R]$
- Similar to modularity, we can normalize covariance by dividing by the standard deviation to obtain the Pearson correlation ρ
- $\rho(X_L, X_R) = \frac{\sigma(X_L,X_R)}{\sigma(X_L)\sigma(X_R)} = \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)x_i x_j}{\frac{1}{2m}\sum_{ij} A_{ij} x_i^2 - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m} x_i x_j}$

Table of contents — SL08

INFLUENCE

- A frequent question in assortative networks is to determine the influence of nodes.
- In a simple model, nodes make decision based on their neighbors who have already made the decision
- The Linear Threshold Model (LTM) is such a model where the weights of the edges between the nodes represent how much the nodes can affect each other
    - A node become active at time t if the sum of the weight of its incoming edges reaches its own threshold θ

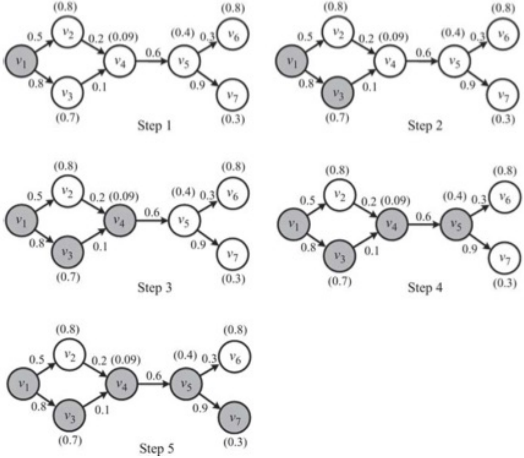
Linear Threshold Model

Algorithm 8.1 Linear Threshold Model (LTM)

Require: Graph $G(V, E)$, set of initial activated nodes A_0

return Final set of activated nodes A_∞
i=0;
Uniformly assign random thresholds θ_v from the interval $[0, 1]$;
while $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ **do**
    $A_{i+1} = A_i$
    inactive $= V - A_i$;
    for all $v \in$ inactive **do**
        if $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. **then**
            activate v;
            $A_{i+1} = A_{i+1} \cup \{v\}$;
        end if
    end for
    $i = i + 1$;
end while
$A_\infty = A_i$;
Return A_∞;

LTM Simulation



LTM simulation (values on the nodes represent thresholds) ©SMM