Massive Graph Management & Analytics

Models of Influence & Diffusion

Nacéra Seghouani

Computer Science Department, CentraleSupélec
Laboratoire Interdisciplinaire des Sciences du Numérique, LISN
nacera.seghouani@centralesupelec.fr

2024-2025

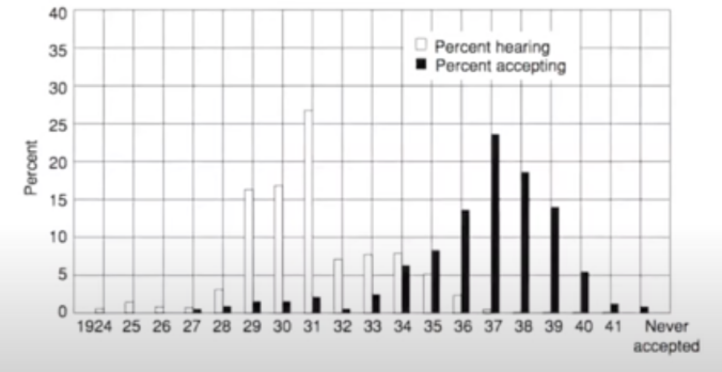
Outline

- ☞ Introduction & Motivations
- ☞ Influence & Diffusion Models
  - $\rightarrow$ Independent Cascade (IC) and Linear Threshold (LT) models
- ☞ Influence Maximisation Problem
- ☞ Research papers

CentraleSupélec

Introduction & Motivation

- B. Ryan & N. Gross published *Acceptance and Diffusion of Hybrid Corn Seed in two Communities*, 1950
- Information effect vs adopting innovation between 1924 and 1940
  - → The diffusion pattern made up of three periods: long period of slow initial growth, rapid rise in adoption and a brief decline as the most resistant adopters accepted the new technology
  - → innovators (starters), early adopters (usually small number), early majority, late majority, then laggards



Introduction & Motivation

- ☞ Studying and modelling the spread of beliefs or ideas or information (rumours, news) or virus
  →active research topic in various fields economics, epidemiology, social science, political science, computer science, etc.
- ☞ Influence models have been studied for years:
  → Original mathematical models: Schelling (1970, 1978) & Granovetter 1978;
  → Viral Marketing Strategies modelled by Domingos & Richardson 2001;
  → Network coordination games 2000;
  → D. Kempe, J. Kleinberg, and E. Tardos 2003, 2005
- ☞ Studying diseases or contagions, the most commonly used epidemic models: **S**usceptible-**I**nfected-**R**ecovered (SIR) and **S**usceptible-**I**nfected- **S**usceptible (SIS), Kendall 1956; Ross and Hudson in 1917; Kermack and McKendrick in 1927.
- ☞ Influence of the social environment on health: behaviors such as eating, practicing physical activities, drug use and seeking medical follow-up (House, Landis and Umberson, 1988)

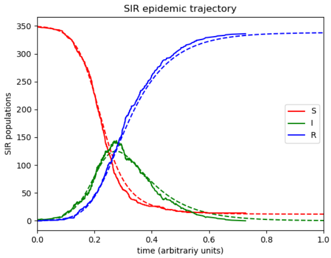
Compartmental models in epidemiology

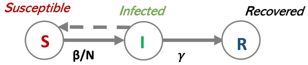
- General models for infectious diseases from human to human, many mathematical models since Spanish flu in 1918, applied first to SIDA in 1980 and recently to Covid 19.
- Epidemiological model is based on 2 concepts: (i) compartments to divide individuals and (ii) rules to specify the rate of transition between compartments like the force of infection.
- With or without the dynamics of birth and death, immunity period, . . .
- **S**usceptible-**I**nfected-**R**ecovered (SIR) is the simplest predictive model, and recovery confers lasting resistance (death is negligible).
  - ✓ **S**: number of susceptible individuals, when a susceptible individual and an infectious individual come into contact, the susceptible individual contracts the disease and transitions to the infectious compartment.
  - ✓ **I**: number of infectious individuals.
  - ✓ **R**: number of removed (and immune) or deceased individuals (negligible). Also called "recovered" or "resistant".

Compartmental models in epidemiology: Susceptible-Infected-Revovered (SIR)

- $S(t)$, $I(t)$ and $R(t)$ functions defined to study the dynamic in a short infectious period.
  - ✓ N is the total population, β the average number of contacts per person per time, $\frac{\beta}{N}$ the transmission parameter
  - ✓ $\frac{dS}{dt}$ is the balance of individual number in **S**, negative means that the individuals leave **S**.
  - ✓ transition rate γ, between **I** and **R**, is proportional to the number of infectious individuals.
  - ✓ $\frac{dI}{dt}$ is the incidence in terms of infections.
  - ✓ $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$, this means that the size of population doesn't change.
- SIR system can be expressed using the differential equation system solution (dashed):

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta.S.I}{N}, \\ \dfrac{dI}{dt} = \dfrac{\beta S.I}{N} - \gamma I, \\ \dfrac{dR}{dt} = \gamma.I, \end{cases}$$

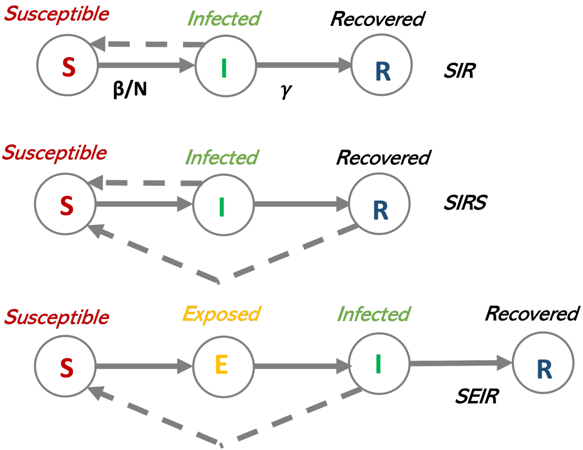




- Gillespie algorithm

  → used to simulate chemical or biochemical systems: Stochastic Simulation Algorithm which generates a statistically correct trajectory (possible solution) of a stochastic equation system for which the reaction rates are known.

Compartmental models in epidemiology: Many Variants



Compartmental models in epidemiology: Many Variants

- Many SIR model variants where:
  - ✓ upon recovery no immunity (SIS model), the common cold and influenza, do not confer any long-lasting immunity;
  - ✓ immunity lasts only for a short period of time (SIRS);
  - ✓ a latent period (Exposed) of the disease where the person is not infectious (SEIS and SEIR);
  - ✓ infants can be born with Maternally derived immunity (MSIR);
  - ✓ model differentiates between Recovered (individuals having survived the disease and now immune) and Deceased (SIRD);
  - ✓ Vaccinated susceptible population (SIRV).

Challenges in Social Networks

- ☞ Diffusion models are used to identify the way the information is transmitted in a network:
  - → how to model the information diffusion process ? in a social network?
  - → how to identify the influencers? which kind of graph-based measures ?
  - → how to maximise the influence? or how to minimise/stop the influence (which links to remove)
- ☞ From computer science view, we need to develop fast and efficient algorithms on large networks

Influence & Diffusion Models

- Probabilistic model: probability that someone do something based on its activated n neighbours to become activated (Goldenberg et al; 2001)

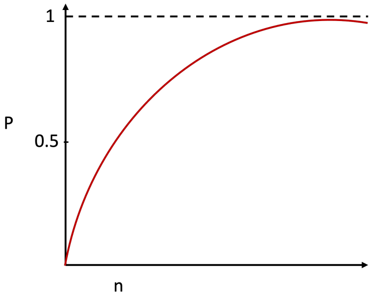
$$P(n) = 1 - (1 - p)^n$$

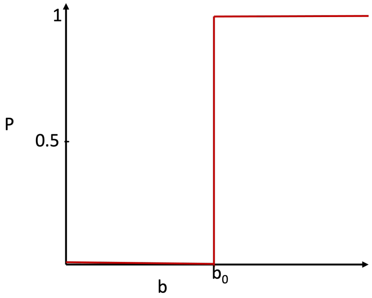
with p the activation probability of a neighbor

intuitively the high the number of neighbors do something the high the probability that you do the same thing

- Threshold model: nothing happens (no activation) until the threshold reached (critical mass) (Schelling & Granovetter 1978)

$$P(b) = \delta(b > b_0)$$

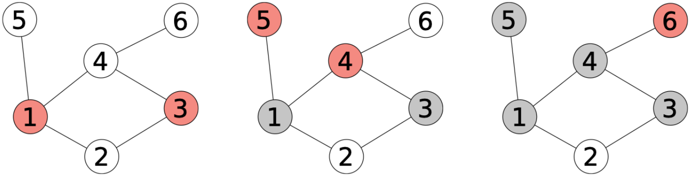




- ☞ Two models: Threshold, Independent Cascade (probabilistic)
  - ✓ In both the information diffusion occurs by the activation of nodes in discrete steps
- ☞ Main idea: define a diffusion process on the network originating from a set seed $\mathcal{S}$. The expected number of activated nodes at the end is the influence $\sigma(\mathcal{S})$ of $\mathcal{S}$.
  - ✓ Network $\mathcal{G}(V, E)$ represented as a directed graph
  - ✓ Individual nodes are active or inactive
  - ✓ Process:
    - → Start with initial set of active seed nodes $\mathcal{S}$
    - → Run t steps and end when no more possible activation

Influence & Diffusion Models: Independent Cascade Model

- ☞ When node u becomes active, it is given a single chance to activate each currently inactive neighbor v
- ☞ Succeeds with a probability $p_{u,v}$ (system parameter).
  - ✓ Independent of history
  - ✓ This probability is generally a coin flip $\mathcal{U}[0,1]$
  - ✓ If u succeeds, then v will become active in step $t+1$; but whether or not u succeeds, it cannot make any further attempts to activate v in subsequent rounds.
  - ✓ If v has multiple newly activated neighbors, their attempts are sequenced in an arbitrary order.



Influence & Diffusion Models: Independent Cascade Model

- A node v is activated by its incoming activated neighbors u independently with the probability $p_{u,v}$
- Let D_t be the set of active nodes at t. For $v \in \mathcal{N}(D_t)$, its probability of being active at $t+1$ is:

$$p_v(t+1) = 1 - \prod_{u \in \mathcal{N}(v) \cap D_t} (1 - p_{u,v})$$

- The sets I_t and S_t of resp. infected and not infected nodes at each discrete time are defined as follows:

$$I_0 = \mathcal{S}; \quad S_0 = V \setminus I_0; \quad S_{t+1} = S_t \setminus I_{t+1}$$

  Set of all infected nodes throughout a contagion process originating at $\mathcal{S}$

$$I(\mathcal{S}) = \cup_{t \geq 0} I_t$$

- The expectation is taken over the random infection attempts from the infected nodes.

$$\sigma(\mathcal{S}) = \mathbb{E}[|I(\mathcal{S})|]$$

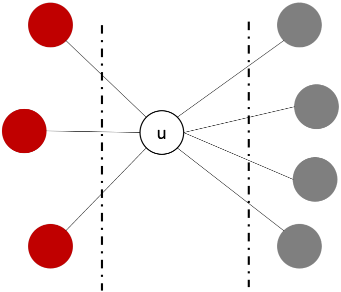
Influence & Diffusion Models: Independent Cascade Model

Determining influence probabilities:

- ☞ A commonly-used assigns to edge (u, v) $p_{u,v} = \frac{1}{d_v^-}$ (in degree).
- ☞ Some studies propose to learn influence probabilities from data, e.g., propagation actions (e.g. replies, forwards, etc.) in the social networks
- ☞ Saito et al. (2008) are the first to formalize the problem of learning edge probabilities from past propagation actions as a likelihood maximization problem.
- ☞ Deep Graph Representation Learning and Optimization for Influence Maximization (2023), studied in GNN extensions lectures.

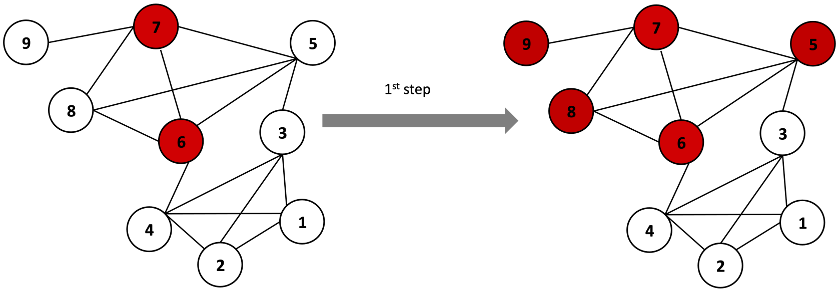
Influence & Diffusion Models: Threshold Model

- Node u of degree d_u has p proportion of red neighbours and $(1-p)$ of grey ones
- To accept the new technology red: $\rho p > \gamma(1-p)$ the threshold to accept is $p \geq \frac{\gamma}{\rho+\gamma}$, red rewards ρ and grey rewards γ
- What does mean: $\rho = \gamma$



Influence & Diffusion Models: Threshold Model

- Node u of degree d_u has p proportion of red neighbours and $(1-p)$ of grey ones
- To accept the new technology red: $\rho p d > \gamma(1-p)d$ the threshold to accept is $p \geq \frac{\gamma}{\rho+\gamma}$, red rewards ρ and grey rewards γ
- $\rho = 3$ and $\gamma = 2$, threshold = $\frac{2}{5}$, from nodes 6 and 7, nodes 5 $(\frac{2}{4})$, 9 $(\frac{1}{1})$, 8 $(\frac{2}{3})$
- Complete cascade until no possible activation



Influence & Diffusion Models: The Linear Threshold Model

- ☞ A node v is influenced by each incoming active neighbour u according to a weight $\omega_{u,v}$

$$\sum_{u \in \mathcal{N}(v)} \omega_{u,v} \leq 1$$

- ☞ Each v has a random acceptance threshold $\theta \sim \mathcal{U}[0,1]$: this represents the fraction of v's neighbors that must become active in order for v to become active.
  $\rightarrow$ Given random thresholds, and an initial set of active nodes S_0 (with all other nodes inactive), the diffusion process unfolds in discrete steps:
  $\rightarrow$ in step t, all nodes that were active in $t-1$ remain active, and activate any node v such that the total weight of its active neighbors is at least θ_v.

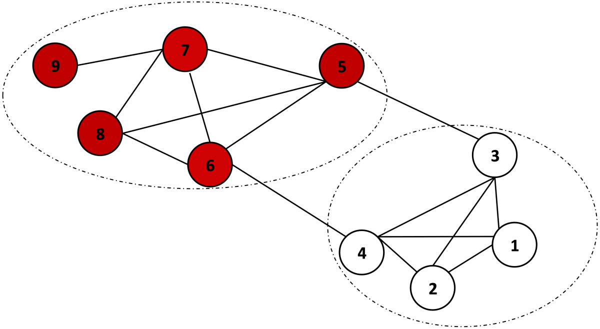
$$\sum_{u \in \mathcal{N}(v)} \omega_{u,v} \geq \theta_v$$

- ☞ Example to compute $\omega_{u,v}$ is to take into account the degrees.

$$\omega_{u,v} = \frac{1}{d_v}$$

Influence & Diffusion Models: Cascades and Clusters

- homophily can often serve as a barrier to diffusion, by making it hard for innovations to arrive from outside densely connected communities.
- how the cluster structure of a network might tell us something about the success or failure of a cascade.
- a cluster of density δ is a set of nodes such that each node in the set has at least a p fraction of its network neighbours in the set. Two clusters of density $\frac{3}{4}$ in the network
- To get cascade into cluster the threshold should be $\leq 1 - \delta$



Influence Maximization Problem Formulation

- Given $\mathcal{G}(V, E)$, let σ be a function such that
  $\sigma : \mathcal{S} \rightarrow \mathbb{N}$ maps a set of nodes $\mathcal{S} \in V$ to their influence value $\sigma(\mathcal{S})$
  number of activated nodes when propagation stops
- The Influence Maximization Problem asks: for a given k, called budget, find a k-node $\mathcal{S}$:

$$max_{|\mathcal{S}| \leq k} \sigma(\mathcal{S})$$

- Solving a constrained maximization problem with $\sigma(\mathcal{S})$ as the objective function is NP-hard. Consider an instance of the NP-complete Set Cover problem

Influence Maximization Problem: Greedy Framework

☞ Approximation Greedy Algorithm($\mathcal{G}(V,E),k$): Each iteration add to S the node providing the maximum marginal gain in spread

$S \leftarrow \emptyset$

for $i = 1 : k$ do

select $u^* = argmax_{u \in V \setminus S} \sigma(S \cup \{u\}) - \sigma(S))$

$S \leftarrow S \cup \{u^*\}$

CentraleSupélec

Submodular Functions

- Set function f is submodular if for sets R et T and $R \subseteq T$, $\forall v \notin T$ and R

$$f(R \cup \{v\}) - f(R) \geq f(T \cup \{v\}) - f(T)$$

  - → Function of diminishing returns: the marginal gain from adding an element to a set R is at least as high as the marginal gain from adding the same element to a superset of R
  - → Function is monotone $f(R \cup \{v\}) \geq f(R)$
- **Theorem** **:** For a non-negative, monotone, submodular function f, let $\mathcal{S}$ be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let $\mathcal{S}^*$ be a set that maximizes the value of f over all k-element sets. Then,

$$f(\mathcal{S}) \geq 1 - (1 - \frac{1}{k})^k f(\mathcal{S}^*)$$

  in other words, $\mathcal{S}$ provides a $(1 - \frac{1}{e})$ = $lim_{k \to \infty} 1 - (1 - \frac{1}{k})^k$ approximation using $e \approx 2,718$.
- $\sigma()$ is a submodular function
  $\sigma(S) \geq (1 - 1/e)\sigma(S^*)$
- Greedy algorithm for maximum influence set finds a set S such that its influence set $\sigma(S)$ is within $1/e \approx 0.367$ from the optimal set $\sigma(S^*)$ $\sigma(S) \geq 0.629\sigma(S^*)$

Research Paper Study

Maximizing the Spread of Influence through a Social Network. David Kempe, Jon Kleinberg and Eva Tardos. Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 137–146, ACM 2003

- ☞ Proof that for an arbitrary instance of IC model, the resulting $\sigma(.)$ is submodular For any A and elements v
  - ✓ Difficult to analyse the margin $\sigma(A \cup \{v\}) - \sigma(A)$: IC process is underspecified, no order in which newly activated nodes in a given t will attempt to activate their neighbors.
  - ✓ it does not matter whether the coin was flipped at the moment that v became active, or whether it was flipped at the very beginning of the whole process and is only being revealed now.
    $\rightarrow$ Compute all pairs (blocked or live)
    $\rightarrow$ v ends up active if and only if there is a path of live edges from a node in A to v.
  - ✓ $\sigma_X(A)$: the total number of nodes activated by the process originating from A, and X is the set of outcomes of all coin flips on edges ($\sigma_X(A)$ is a deterministic quantity).
  - ✓ $\sigma_X(A) = |\cup_{v \in A} R(v, X)|$ where $R(v, X)$ is the set of all nodes that can be reached from v on a live-edges path.
    $\rightarrow$ need to prove that the function $\sigma_X(A)$ is submodular.
    $S \subseteq T$, consider $\sigma_X(S \cup \{v\}) - \sigma_X(S) = |R(v, X) / \cup_{u \in S} R(u, X)|$ it is at least as large as the number of elements in $R(v, X)$ that are not in the (bigger) union $\cup_{v \in T} R(v, X)$.

$$\sigma_X(S) \cup \{v\}) - \sigma_X(S) \geq \sigma_X(T) \cup \{v\}) - \sigma_X(T)$$

  - ✓ Finally, $\sigma(A) = \sum_X \sigma_X(A) P(X)$ a non-negative linear combination of submodular functions is also submodular, and hence $\delta(A)$ is submodular

CentraleSupélec

Research Paper Study

- Proof the influence maximization problem is NP-hard for IC model.
  Consider an instance of the NP-complete Set Cover problem: set of subsets $S = \{S_1, S_2, ..., S_m\}$ and set $U = \{u_1, u_2, ..., u_n\}$
  question: $\exists k \cup_{S_i \in S'} S_i = U, S' \subseteq S, |S'| = k$ with $k < n < m$
- Given an arbitrary instance of the Set Cover problem, we define a corresponding directed bipartite graph with $n+m$ nodes.
  $\rightarrow$ There is a node i corresponding to each set S_i, a node j corresponding to each element u_j, and a directed edge (i, j) with activation probability $p_{i,j} = 1$ whenever $u_j \in S_i$. The Set Cover problem is equivalent to deciding if there is a set A of k nodes in this graph with $\sigma(A) \geq n+k$
- Initially activating the k nodes corresponding to sets in a Set Cover solution results in activating all n nodes corresponding to the ground set U, and if any set A of k nodes has $\sigma(A) \geq n+k$, then the Set Cover problem must be solvable.