Mathematical models for the diffusion of innovation in social networks

Alessandro Giua
DIEE, University of Cagliari, Italy
Email: giua@unica.it

version 2 May 2020

Contents

1 Introduction

In the last decades many researchers from different fields have been interested in the study of how innovation spreads in social networks. What is the mechanism that convinces an individual to follow a new idea or to buy a new product? What is the best marketing strategy which a company should adopt to take a competitive advantage? How does viral marketing works?

The studies on diffusion of an innovation in social networks began in the middle of the 20th century and several mathematical models have been proposed to give an answer to questions of this type. Among them we can mention *threshold models* and *epidemic models*. All these models are based on the same concept: in a social network the behaviour of each individual is highly influenced by the behaviour of its neighbors.

Models based on *threshold* assume that an individual adopts a behaviour if a certain ratio of its social contacts have already adopted it. Granovetter [3] and Schelling [7] firstly introduced the threshold models in 1978. Their approach was based on the use of node-specific thresholds and different mathematical models for diffusion has been proposed subsequently. Epidemic models assume that a node adopts a behaviour with a certain probability if at least one of its neighbors has adopted it. Threshold models are more suitable to describe social influence phenomena and individual behaviours, while epidemic models are more used for mass behaviours. Kempe *et al.* [4] showed that both models can be generalized and their generalized versions become equivalent. However, the basic linear threshold model and independent cascade model stay as two distinct models.

When a node adopts the innovation we say that it becomes *active*, otherwise is said to be *inactive*. It is implicitly assumed that a node can adopt the innovation, but once adopted, it cannot abandon it, i.e., a node can switch its state from inactive to active but cannot switch it from active to inactive. This model can be used to represent systems in which the adoption of an innovation is permanent and in the literature is called *progressive*. For instance, a progressive model can be suitable to describe the behavior of individuals that buy a certain item: once they spend money to buy that, i.e., once they adopt the innovation, usually it is not possible to return it and have the money back, thus we can say that the adoption of the innovation has a permanent nature.

In this note we discuss two models.

- The *Independent Cascade Model*: a stochastic epidemic model.
- The *Linear Threshold Model*: a deterministic threshold model.

2 Network structure

As in most of the models appeared in the literature, a social network is represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges.

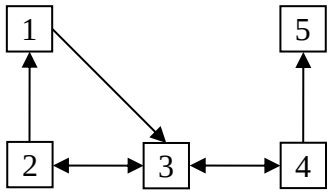

Figure 1: A directed graph with 5 nodes.

Each node $i \in \mathcal{V}$ represents an individual and in this notes we use the terms *individual* or *node* interchangeably. An oriented edge $(i, j) \in \mathcal{E}$ denotes that node j is influenced by node i (no selfloops, i.e., edges from one node to itself, are allowed).

Example 1 Consider the directed graph in Figure 3. The set of nodes is $\mathcal{V} = \{1, 2, 3, 4, 5\}$. The set of edges is $\mathcal{E} = \{(1, 3), (2, 1), (2, 3), (3, 2), (3, 4), (4, 3), (4, 5)\}$. $\diamond$

Node i is an in-neighbor of node j if $(i, j) \in \mathcal{E}$, i.e., there exists an edge going from i (tail node) to j (head node). The *set of in-neighbors* of a node $j \in \mathcal{V}$ is[1]:

$$\mathcal{N}_j = \{i \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}.$$

Example 2 Consider again the directed graph in Figure 3. The set of in-neighbors of its nodes are: $\mathcal{N}_1 = \{2\}, \mathcal{N}_2 = \{3\}, \mathcal{N}_3 = \{1, 2, 4\}$, etc. $\diamond$

3 Independent Cascade Model (ICM)

In the *Independent Cascade Model*, the mechanism according to which the innovation propagates is based on the assumption that when an individual observes the actions of other persons may engage in the same acts. This decision occurs in a random way, with different probabilities according to the influence that the observed person has on the considered individual. It has been observed that his is also the way in which contagious diseases usually spread from one person to another, and for this reason models of this type are called epidemics models.

As an example, consider an individual that so far has not yet bought a tablet. By observing different friends that already use tablets, she may be convinced that the new device may be useful to her for work, for accessing information while traveling, for reading books and newspapers, etc.

3.1 Cascade network and evolution

As discussed in Section 2, the network is described by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In addition, an *edge probability function* by $p : \mathcal{E} \to [0, 1]$ associates to each edge (i, j) a

[1]Is is usual to define the set of in-neighbor of a node j as $\mathcal{N}_j^{in}$ but here use the simpler notation $\mathcal{N}_j$.

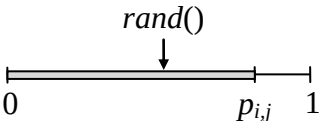

Figure 2: Simulating the propagation of innovation along edge (i, j): innovation propagates if $rand() \leq p_{i,j}$.

number $p_{i,j} \in [0, 1]$ representing the *probability* that node j adopts the innovation if node i has adopted it. Thus, we denote an *independent cascade network* by a triple

$$\mathcal{G}_{IC} = (\mathcal{V}, \mathcal{E}, p).$$

Let us define ϕ_0 as the *seed set*, i.e., the set of nodes which have adopted the innovation at time $k = 0$. From the seed set the innovation spreads through the social network, and we denote by ϕ_k the set of nodes *activated at time* k, i.e., those nodes that are not active at time $k - 1$ but adopt the innovation at time k. We also denote by $\Phi_k = \bigcup_{j=0}^{k} \phi_j$ the set nodes that are *active at time* k, i.e., those which have adopted the innovation during the time interval $[0, k]$.

When a node $i \in \mathcal{N}_j$ in the in-neighborhood of an inactive node j adopts the innovation at time k, it may propagate the innovation to node j at time $k + 1$ with probability $p_{i,j}$. To simulate this stochastic event, we may may generate a random number $rand()$ uniformly distributed in the interval $[0, 1]$ and assume that the innovation propagates when $rand() \leq p_{i,j}$ as shown in Figure 2.

Note that a node i activated at time k may only propagate the innovation to its inactive in-neighbors at time $k + 1$. In other words, once the arc (i, j) has been selected to propagate the innovation from node i to node j at time k, regardless of the fact that j adopts the innovation or not, it cannot be selected any more at a future time.

The innovation spreads in the network until no more individual can adopt it, and we denote the set of final adopters as:

$$\Phi^* = \bigcup_{k=0}^{\infty} \phi_k.$$

Algorithm 1 describes the dynamic of the network and returns as output the set Φ^*, computing also the set of nodes ϕ_k which adopt the innovation at each step k .

Algorithm 1 (Evolution of an independent cascade network)

INPUT: An independent cascade network $\mathcal{G}_{IC} = (\mathcal{V}, \mathcal{E}, p)$. A set of initial adopters $\phi_0 \subseteq \mathcal{V}$.

OUTPUT: The set of final adopters Φ^*.

1. *Let $\Phi = \phi_0$, $\bar{\Phi} = \mathcal{V} \setminus \phi_0$;*

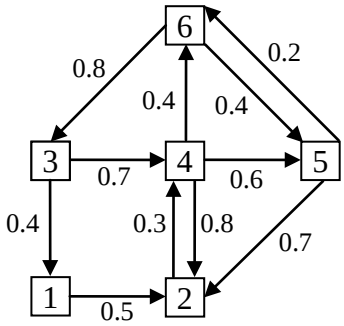

Figure 3: Network in Example 3.

2. *Let* $k = 0$*;*

3. ***while*** $\phi_k \neq \emptyset$*,*

   (a) Let $k = k + 1,\ \phi_k = \emptyset$*;*

   (b) ***for*** $j \in \bar{\Phi}$*,*
   for $i \in \mathcal{N}_j \cap \phi_{k-1}$*,*
   if $rand() \leq p_{i,j}$*,* ***then*** $\phi_k = \phi_k \cup \{j\}$*;*
   end for*;*
   end for*;*

   (c) Let $\Phi = \Phi \cup \phi_k,\ \bar{\Phi} = \bar{\Phi} \setminus \phi_k$*;*

   end while*;*

4. *Let* $\Phi^* = \Phi$. ◇

Example 3 Consider the network in Figure 3 with seed set $\phi_0 = \{6\}$. Next to each arc (i, j) is shown the value of the probability $p_{i,j}$.

A first evolution of the system is shown in Table 1. Here at step $k = 1$ we test edges $(6, 3)$ and $(6, 5)$: only node 3 adopts the innovation. At step $k = 2$ we test edges $(3, 1)$ and $(3, 4)$: only node 1 adopts the innovation. Finally at step $k = 3$ we test edge $(1, 2)$: node 2 does not adopt the innovation and we stop. The set of final adopters is $\Phi^* = \{1, 3, 6\}$.

As we have pointed out, the diffusion of innovation in the ICM is stochastic. This means that the set of final innovators may randomly change from one simulation to another one starting with the same seed set. As an example, a second evolution of the system is shown in Table 2. Here at step $k = 1$ we test edges $(6, 3)$ and $(6, 5)$: both nodes 3 and 5 adopt the innovation. At step $k = 2$ we test edges $(3, 1)$, $(3, 4)$ and $(5, 2)$: only node 4 adopts the innovation. At step $k = 3$ we test edge $(4, 2)$: node 2 adopts the innovation. At step $k = 4$ there are no edges to test — since the single edge outputting node 2 goes to node 4 that has already adopted the innovation — and we stop. The set of final adopters is $\Phi^* = \{2, 3, 4, 5, 6\}$. ◇

k	edge	$p_{i,j}$	$rand()$	adopt	ϕ_k	Φ_k
0					$\{6\}$	$\{6\}$
1	$(6,3)$	0.8	0.73	Y	$\{3\}$	$\{3,6\}$
	$(6,5)$	0.4	0.52	N		
2	$(3,1)$	0.4	0.37	Y	$\{1\}$	$\{1,3,6\}$
	$(3,4)$	0.7	0.82	N		
3	$(1,2)$	0.5	0.66	N	$\emptyset$	$\{1,3,6\}$

Table 1: First simulation in Example 3.

k	edge	$p_{i,j}$	$rand()$	adopt	ϕ_k	Φ_k
0					$\{6\}$	$\{6\}$
1	$(6,3)$	0.8	0.73	Y	$\{3,5\}$	$\{3,5,6\}$
	$(6,5)$	0.4	0.32	Y		
2	$(3,1)$	0.4	0.64	N	$\{4\}$	$\{3,4,5,6\}$
	$(3,4)$	0.7	0.53	Y		
	$(5,2)$	0.7	0.91	N		
3	$(4,2)$	0.8	0.48	Y	$\{2\}$	$\{2,3,4,5,6\}$
4	–	–	–	–	$\emptyset$	$\{2,3,4,5,6\}$

Table 2: Second simulation in Example 3.

3.2 Computing the activation probabilities

In an independent cascade network, given a seed ϕ_0 we may be interested in computing the probability π_i that each node $i \in \mathcal{V}$ adopts the innovation. Obviously this probability is equal to 1 for all nodes in the seed set but unknown for all other nodes.

We define the *node activation probability vector* as

$$\boldsymbol{\pi} = [\pi_1 \; \pi_2 \; \ldots \; \pi_n].$$

When it is necessary to explicitly point out the dependence of this vector from the seed set we may also denote it by $\boldsymbol{\pi}(\phi_0)$.

In [8] it was presented an algorithm, called *Path Method*, to compute the (exact) node activation probability vector. However, the time complexity of such algorithm is $\mathcal{O}(6^N$, i.e., exponential on the number of nodes N. For this reason, the Path Method can only be used for small networks.

Example 4 Consider the network in Figure 3 with seed set $\phi_0 = \{6\}$. Using the path based method in [8] one can find that the node activation probability vector is $\boldsymbol{\pi}$ whose components are given in the following table:

node i	1	2	3	4	5	6
π_i	0.32	0.6758	0.8	0.6073	0.6068	1

$\diamond$

3.2.1 Computing the activation probabilities by simulation

One possible way to experimentally compute the node activation probabilities consists in simulating the evolution of the network for a large number of runs. Assume that we run the simulation N times, and observe that node j adopts the innovation $r_j \leq N$ times. Then the observed frequency of activation is r_j/N and as N goes to infinity we expect that the observed frequency will go to the expected probability.

Example 5 Consider the network in Figure 3 with seed set $\phi_0 = \{6\}$. By simulating it evolution for an increasing number of runs we obtain the results in Figure 4, where we have reported for all nodes the observed frequencies as N goes from 1 to 5000. Note that even with a large number of runs the results of the simulation are affected by small random fluctuations.

The node activation probability vector computed via simulation is $\boldsymbol{\pi}^{sim}$ whose components are given in the following table:

node i	1	2	3	4	5	6
π_i^{sim}	0.318	0.674	0.799	0.610	0.603	1.000

Note that in this simulation for all $i \in \mathcal{V}$ if holds that

$$\pi_i - 0.005 \leq \pi_i^{sim} \leq \pi_i + 0.005.$$

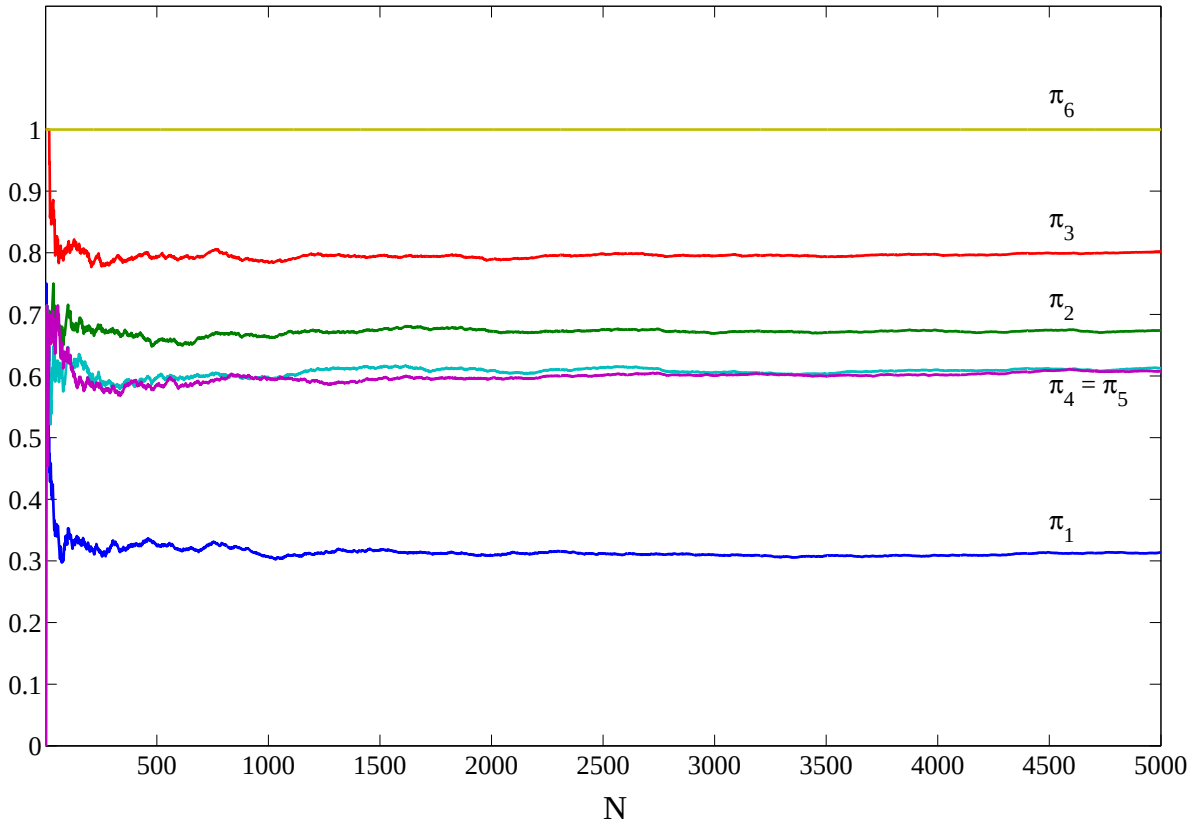

Figure 4: Computation of the node activation probabilities by simulation for the network in Example 3.

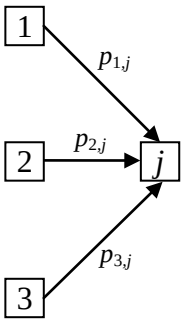

Figure 5: A node j with three neighbors.

3.2.2 Computing the activation probabilities by fixed-point iteration

The computation of the node activation probabilities by simulation is not very practical and requires a large number of simulation runs. For this reason, in this subsection we present a procedure based on *fixed-point iteration* which determines an upper bound of the activation probabilities.

To show how this technique can be applied to the problem at hand, let us first consider a network node $j \notin \phi_0$ that has neighbors $\mathcal{N}_j = \{1, 2, 3\}$ as shown in Figure 6. If the events "node i is activated" (for $i \in \mathcal{N}_j \cup \{j\}$) are independent events, then the probability $1-\pi_j$ that node j *does not adopt the innovation* is equal to :

$$(1 - \pi_j) = (1 - \pi_1\, p_{1,j})(1 - \pi_2\, p_{2,j})(1 - \pi_3\, p_{3,j}) = \prod_{i=1}^{3}(1 - \pi_i\, p_{i,j})$$

since this requires that all three neighbors fail to activate j. To justify this expression consider, say, the first neighbor: assuming no other node activates j, the probability that j is activated by node 1 is $\pi_1\, p_{1,j}$. Hence $(1 - \pi_1\, p_{1,j})$ is the probability that node j is not activated by node 1.

This expression can be rewritten and generalized so as to express the activation probability of node j as a function of its neighbors activation probabilities, i.e.,

$$\pi_j = 1 - \prod_{i \in \mathcal{N}_j} (1 - \pi_i\, p_{i,j}). \tag{1}$$

Assuming that the activation probability of each node in a networks can be computed using eq. (1), we can define the following vector.

Definition 1 *Given an independent cascade network $\mathcal{G}_{IC} = (\mathcal{V}, \mathcal{E}, p)$ let ϕ_0 be a seed set. The* SteadyStateSpread *probability vector $\boldsymbol{\pi}^{sss}(\phi_0) = [\pi_1^{sss}\ \pi_2^{sss}\ \ldots\ \pi_n^{sss}]$ is the solution*

of the following system of equations:

$$\pi_j^{sss} = \begin{cases} 1 & \text{if } j \in \phi_0, \\ 1 - \prod_{i \in \mathcal{N}_j} (1 - \pi_i^{sss}\, p_{i,j}) & \text{if } j \notin \phi_0; \end{cases} \qquad \forall j \in \mathcal{V} \tag{2}$$

◇

We can finally state the following general result (see also [2, 8]).

Fact 1 *Given an independent cascade network $\mathcal{G}_{IC} = (\mathcal{V}, \mathcal{E}, p)$ let ϕ_0 be a seed set. Let $\boldsymbol{\pi}(\phi_0)$ be the activation probability vector and $\boldsymbol{\pi}^{sss}(\phi_0)$ the* SteadyStateSpread *probability vector. It holds:*

$$\boldsymbol{\pi}(\phi_0) \leq \boldsymbol{\pi}^{sss}(\phi_0),$$

where the relation $\leq$ holds componentwise. ◇

The proof of the result follows from the fact that eq. 1 provides a correct value of the activation probabilities for node j if and only if the events "node i is activated" (for $i \in \mathcal{N}_j \cup \{j\}$) are independent events. When this does not happen, it can be shown that $\pi_j^{sss} > pi_j$, i.e., the *SteadyStateSpread* probability of node j is in general larger than the correct activation probability [2, 8].

The previous result provides an analytical characterization of the *SteadyStateSpread* probabilities but it does not offer a computational viable approach to determine them, because the system of equations is not linear. For this reason we present an algorithm to numerically compute the solution of (2).

The general idea behind this algorithm is briefly outlined. Given a real function of a real variable[2] $f : \mathbb{R} \to \mathbb{R}$, we say that a real number x is a *fixed point of f* if it satisfies

$$x = f(x).$$

Given a point x_0 in the domain of f, the *fixed-point iteration* is

$$x_{k+1} = f(x_k),\ k = 0, 1, 2, \ldots$$

which generates the sequence x_0, x_1, x_2, ... If the sequence converges to a point x and f is continuous, then one can prove that x is a fixed point of f.

In the following $\mathbf{0}_n$ denotes a vector of zero's of length n and given a vector $\boldsymbol{v} \in \mathbb{R}^n$ we denote its 2-norm by

$$||\boldsymbol{v}|| = \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2}.$$

Algorithm 2 (Computing the *SteadyStateSpread* probabilities)

INPUT: An independent cascade network $\mathcal{G}_{IC} = (\mathcal{V}, \mathcal{E}, p)$. A set of initial adopters $\phi_0 \subseteq \mathcal{V}$. A desired tolerance $\varepsilon > 0$.

OUTPUT: The SteadyStateSpread *probability vector $\boldsymbol{\pi}^{sss}(\phi_0)$.*

[2]More generally, the function f can be defined on any metric space (e.g., $\mathbb{R}^n$) with values in that same space.

1. *Let* $\boldsymbol{\pi}^{(-1)} = \mathbf{0}_n$.

2. *Let* $\boldsymbol{\pi}^{(0)} = [\pi_1^{(0)}\ \pi_2^{(0)}\ \ldots\ \pi_n^{(0)}]$ *such that* $\forall j \in \mathcal{V}$

$$\pi_j^{(0)} = \begin{cases} 1 & \text{if } j \in \phi_0, \\ 0 & \text{if } j \notin \phi_0; \end{cases}$$

3. *Let* $k = 0$.

4. ***while*** $||\boldsymbol{\pi}^{(k)} - \boldsymbol{\pi}^{(k-1)}|| > \varepsilon$,

   (a) *Let* $k = k + 1$;

   (b) ***for*** $j \in \mathcal{V}$, *let*

$$\pi_j^{(k)} = \begin{cases} 1 & \text{if } j \in \phi_0, \\ 1 - \prod_{i \in \mathcal{N}_j} (1 - \pi_i^{(k-1)}\, p_{i,j}) & \text{if } j \notin \phi_0; \end{cases}$$

   end while;

5. *Let* $\boldsymbol{\pi}^{sss}(\phi_0) = \boldsymbol{\pi}^{(k)}$. ◇

It is easy to show that the previous algorithm produces a non decreasing sequence of vectors, i.e., $\boldsymbol{\pi}^{(k)} \geq \boldsymbol{\pi}^{(k-1)}$. Also this sequence is upper bounded because each component of $\boldsymbol{\pi}^{(k)}$ is smaller than or equal to 1. Hence it converges to a fixed point, which is the solution of the system of equations (2).

Example 6 Consider again the network in Figure 3 with seed set $\phi_0 = \{6\}$. By means of Algorithm 2 with a tolerance $\varepsilon = 10^{-5}$ we compute in 13 steps the *SteadyStateSpread* probability vector $\boldsymbol{\pi}^{sss}(\phi_0)$. In Figure 6 we have plotted the components of vector $\boldsymbol{\pi}^{(k)}$ for $k = 0, \ldots, 13$. Comparing with Figure 4, we can immediately see that Algorithm 2 converges in a small number of steps in comparison the approach based on simulation.

The following table compares the *SteadyStateSpread* probability vector $\boldsymbol{\pi}^{sss}(\phi_0)$ with the exact node activation probability vector $\boldsymbol{\pi}(\phi_0)$ previously computed in Example 4 .

node i	1	2	3	4	5	6
π_i^{sss}	0.32	**0.782**	0.8	**0.663**	**0.639**	1
π_i	0.32	0.676	0.8	0.607	0.607	1

In the case of nodes 2, 4 and 5 the *SteadyStateSpread* probabilities are significantly larger than the actual activation probabilities. ◇

Several approaches have been proposed to modify the fixed-point approach so that a better estimate of the node activation probability vector can be computed: see [8] for more details.

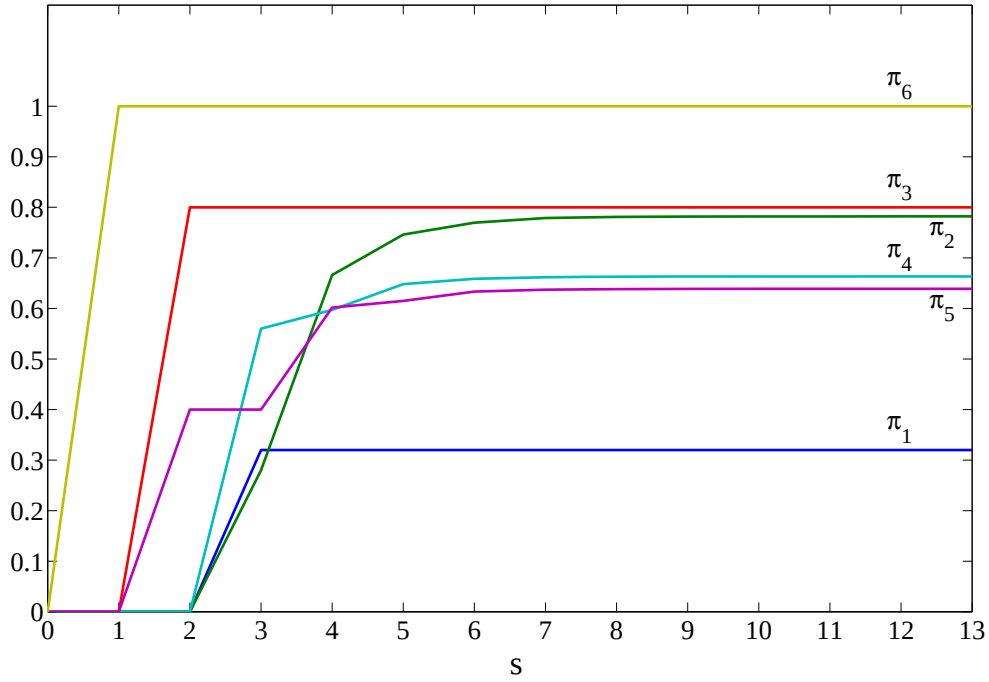

Figure 6: Computation of the *SteadyStateSpread* probabilities by fixed-point iteration in Example 6.

4 Linear Threshold Model (LTM)

In the simplest version of the *Linear Threshold Model*, a threshold value is assigned to each individual and all its neighbors have the same influence on it. An individual adopts the innovation as soon as the ratio of its neighbors who have already adopted it is above its threshold value. The origin of the previous rule is the following: in several competitive games such an individual decision rule has been proved to be the best response to the actions of one's neighbors. As an example, consider an individual belonging to a group of friends; if most of the friends use a particular messaging software (say *whatsapp*), the individual feels the need to use the same software to avoid being emarginated.

4.1 Threshold network and evolution

As discussed in Section 2, the network is described by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In addition, a *threshold function* $\lambda : \mathcal{V} \to [0, 1]$ associated to each node $j \in \mathcal{V}$ a number $\lambda_j \in [0, 1]$ called its *threshold value*. Thus, we denote a *linear threshold network* by a triple

$$\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda).$$

Let us define the *seed set* ϕ_0 as the set of initially active notes, i.e., the set of nodes which

adopt the innovation at time $k = 0$. From the seed set the innovation spreads through the social network, and we denote as ϕ_k the set of nodes which become active at time k. All the individuals that are active at time k, i.e., those that have adopted the innovation during the time interval $[0, k]$, belong to the set $\Phi_k = \bigcup_{i=0}^{k} \phi_i$.

Consider a node $j \notin \Phi_k$ which is not active at time k. This node becomes active at time $k + 1$, i.e., $j \in \phi_{k+1}$, if the following holds:

$$\frac{|\Phi_k \cap \mathcal{N}_j|}{|\mathcal{N}_j|} \geq \lambda_j. \tag{3}$$

Note that $|\Phi_k \cap \mathcal{N}_j|$ denotes the *number of in-neighbors of node j active at time k*, while $|\mathcal{N}_j|$ denotes the *total number of in-neighbors of node j*. Thus a node adopts the innovation if the fraction of it in-neighbors that are active reaches or exceeds the node threshold value.

The innovation spreads in the network until no more individual can adopt it, and we denote the set of final adopters as:

$$\Phi^*(\phi_0) = \bigcup_{k=0}^{\infty} \phi_k.$$

Here we write $\Phi^*(\phi_0)$ to explicitly denote that the set of final adopters for a given linear threshold network depends only on the seed set ϕ_0. When it will not be necessary to show this dependence, we will use the simpler notation $\Phi^*(\phi_0)$.

Algorithm 3 describes the dynamic of the network where the innovation starts from a seed set ϕ_0 and and computes the set $\Phi^*(\phi_0)$. The algorithm also computes the set of nodes ϕ_k which adopt the innovation at each step k .

Algorithm 3 (Evolution of a linear threshold network)

INPUT: A linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$. A set of initial adopters $\phi_0 \subseteq \mathcal{V}$.

OUTPUT: The set of final adopters Φ^*.

1. *Let $\Phi = \phi_0$, $\bar{\Phi} = \mathcal{V} \setminus \phi_0$;*
2. *Let $k = 0$;*
3. ***while** $\phi_k \neq \emptyset$,*
   (a) *Let $k = k + 1$, $\phi_k = \emptyset$;*
   (b) ***for** $j \in \bar{\Phi}$,*
   ***if** $\frac{|\Phi_{old} \bigcap \mathcal{N}_j|}{|\mathcal{N}_j|} \geq \lambda_j$, **then** $\phi_k = \phi_k \cup \{j\}$;*
   (c) *Let $\Phi = \Phi \cup \phi_k$, $\bar{\Phi} = \bar{\Phi} \setminus \phi_k$;*

   ***end while**;*
4. *Let $\Phi^* = \Phi$.* $\diamond$

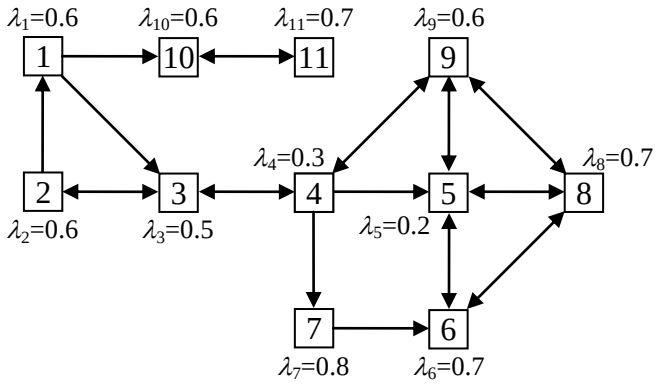

Figure 7: Network in Example 7.

Example 7 Consider the network in Figure 7 with seed set $\phi_0 = \{1, 2\}$. Next to each node i is shown the value of the threshold λ_i. An arc directed from node i to node j denotes that i influences j: note that some arcs are are bidirectional to denote that the two nodes influence each other. The evolution of the system is shown in Table 3. The set of final adopters is $\Phi^* = \{1, 2, 3, 4, 5, 7, 9\}$. $\diamond$

k	ϕ_k	Φ_k
0	$\{1, 2\}$	$\{1, 2\}$
1	$\{3\}$	$\{1, 2, 3\}$
2	$\{4\}$	$\{1, 2, 3, 4\}$
3	$\{5, 7\}$	$\{1, 2, 3, 4, 5, 7\}$
4	$\{9\}$	$\{1, 2, 3, 4, 5, 7, 9\}$
5	$\emptyset$	$\{1, 2, 3, 4, 5, 7, 9\}$

Table 3: Evolution in Example 7.

4.2 Cohesiveness

The following definition formalizes the concept of cohesive set.

Definition 2 (Cohesive set in [1]) *Given a linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$, the set $X \subseteq \mathcal{V}$ is called* cohesive *if for all $j \in X$ it holds*

$$\frac{|X \cap \mathcal{N}_j|}{|\mathcal{N}_j|} > 1 - \lambda_j, \tag{4}$$

or equivalently[3]

$$\frac{|(\mathcal{V} \setminus X) \cap \mathcal{N}_j|}{|\mathcal{N}_j|} < \lambda_j. \tag{5}$$

[3]The two conditions in eq. (4) and eq. (5) are equivalent, i.e., one holds if and only if the other one holds. Each one of them can be used to characterize a cohesive set.

According to eq. (5), a set $X \subseteq \mathcal{V}$ is cohesive if for each $j \in X$ the ratio of neighbors which do not belong to X is strictly smaller than λ_j.

Example 8 Consider the network in Figure 7 discussed in Example 9.

One can easily proof that $X' = \{1, 2, 3\}$ is cohesive. In fact using eq. (4) it holds:

$$\frac{|X' \cap \mathcal{N}_1|}{|\mathcal{N}_1|} = \frac{|\{2\}|}{|\{2\}|} = \frac{1}{1} = 1 > 0.4 = 1 - \lambda_1$$
$$\frac{|X' \cap \mathcal{N}_2|}{|\mathcal{N}_2|} = \frac{|\{3\}|}{|\{3\}|} = \frac{1}{1} = 1 > 0.4 = 1 - \lambda_2$$
$$\frac{|X' \cap \mathcal{N}_3|}{|\mathcal{N}_3|} = \frac{|\{1, 2\}|}{|\{1, 2, 4\}|} = \frac{2}{3} > 0.5 = 1 - \lambda_3$$

In a similar way, one can show that $X'' = \{6, 8\}$ and $X''' = \{10, 11\}$ are also cohesive sets. Finally it is possible to show that the union of cohesive sets is also a cohesive set. So sets $X' \cup X''$, $X' \cup X'''$, $X'' \cup X'''$ and $X' \cup X'' \cup X'''$ are also cohesive. $\diamond$

The following result shows that nodes in a cohesive set X cannot become active unless some node in X is initially active. In other words, a cohesive set where no node has adopted the innovation remains forever closed to the innovation spreading.

Proposition 1 *Given a linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$, let X be a cohesive set and consider a seed set ϕ_0.*

If $\phi_0 \cap X = \emptyset$ then for all $k \geq 0$ it holds $\Phi_k \cap X = \emptyset$, i.e., no individual in X will ever adopt the innovation.

Proof. We proof the result by induction.

(base step) Obviously $\Phi_0 \cap X = \phi_0 \cap X = \emptyset$ by assumption.

(inductive step) We know show that if $\Phi_k \cap X = \emptyset$ then $\Phi_{k+1} \cap X = \emptyset$. In fact in such a case using eq. (5) one can see that for all nodes $j \in X$ holds (see also the Venn diagram in Figure 8)

$$\frac{|\Phi_k \cap \mathcal{N}_j|}{|\mathcal{N}_j|} \leq \frac{|(\mathcal{V} \setminus X) \cap \mathcal{N}_j|}{|\mathcal{N}_j|} < \lambda_j.$$

This implies that no node $j \in X$ adopts the innovation at time $k + 1$ and $\Phi_{k+1} \cap X = \emptyset$. $\square$

Based on the previous property, the following result from [1] characterizes the set of final adopters in terms of the net structure and of the seed set ϕ_0.

Lemma 1 (Lemma 2 in [1]) *Given a linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$, let $\phi_0 \subseteq \mathcal{V}$ be the seed set and let $\mathcal{M} \subseteq \mathcal{V} \setminus \phi_0$ be the maximal cohesive set contained in the complement of ϕ_0. The set of final adopters Φ^* is given by:*

$$\Phi^* = \mathcal{V} \setminus \mathcal{M}, \tag{6}$$

i.e., all nodes in the network will eventually adopt the innovation except those in $\mathcal{M}$.

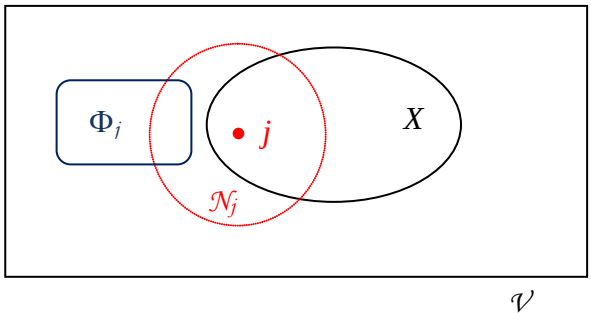

Figure 8: Diagram used in the proof or Proposition 1.

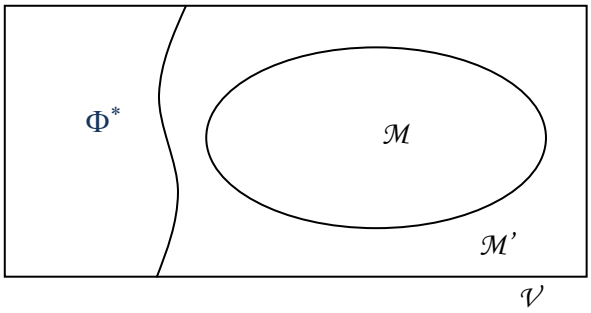

Figure 9: Diagram used in the proof or Lemma 1.

Proof. We know by Proposition 1 that no node in $\mathcal{M}$ will adopt the innovation, i.e.,

$$\Phi^* \subseteq \mathcal{V} \setminus \mathcal{M}.$$

Assume now, by contradiction, that $\Phi^* \subsetneq \mathcal{V} \setminus \mathcal{M}$ and consider the set $\mathcal{M}' = \mathcal{V} \setminus \Phi^* \supsetneq \mathcal{M}$ (see also Figure 9). Since no node in $\mathcal{M}'$ adopts the innovation when all nodes in Φ^* do, it must hold by (3)

$$\frac{|\Phi^* \cap \mathcal{N}_j|}{|\mathcal{N}_j|} < \lambda_j \qquad \text{for all } j \in \mathcal{M}'$$

which also implies

$$\frac{|(\mathcal{V} \setminus \mathcal{M}') \cap \mathcal{N}_j|}{|\mathcal{N}_j|} < \lambda_j \qquad \text{for all } j \in \mathcal{M}'$$

and thus by (5) $\mathcal{M}'$ is a cohesive set contained in $\mathcal{V} \setminus \phi_0$. However, $\mathcal{M}' \supsetneq \mathcal{M}$ contradicts the statement that $\mathcal{M}$ is the maximal such set. □

Example 9 Consider again the network in Figure 7 discussed in Example 9. Given the seed set $\phi_0 = \{1, 2\}$ it holds

$$\mathcal{V} \setminus \phi_0 = \{3, 4, 5, 6, 7, 8, 9, 10, 11\}.$$

One can easily verify that the set $\mathcal{M} = \{6, 8, 10, 11\}$ is the maximal cohesive set contained in $\mathcal{V} \setminus \phi_0$. (Note that this cohesive set if the union of the sets X' and X'' de-

fined in Example 8.) As expected, the set of final adopters computed in Example 7 is $\Phi^* = \{1, 2, 3, 4, 5, 7, 9\} = \mathcal{V} \setminus \mathcal{M}$. $\diamond$

4.3 A linear algebraic characterization of cohesive sets

In this section we want to show that the set of final adopters determined by Algorithm 3 can also be computed solving an integer programming program (IPP).

We observe that topological information about the graph describing a social network can be encoded in the *adjacency matrix* $A \in \{0, 1\}^{n \times n}$ which is defined as follows:

$$A(i, j) = \begin{cases} 1 & \text{if there is an edge from node } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

We define the *in-neighbors scaled adjacency matrix* $\hat{A} \in [0, 1]^{n \times n}$ as follows:

$$\hat{A}(i, j) = \frac{A(i, j)}{|\mathcal{N}_j|}.$$

We denote with $\Lambda = diag([\lambda_1 \ \lambda_2 \ \ldots \ \lambda_n])$ the diagonal matrix whose diagonal entries are the thresholds of the graph nodes.

We associate to each set of nodes $X \subseteq \mathcal{V}$ a characteristic vector defined as follows.

Definition 3 *Given a set $X \subseteq \mathcal{V}$, the associated characteristic vector $\boldsymbol{x} \in \{0, 1\}^n$ is such that $x_j = 1$ if $j \in X$ else $x_j = 0$.*

We now propose a linear algebraic characterization of cohesive sets.

Lemma 2 *Given a linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$, a set $X \subseteq \mathcal{V}$ is cohesive if and only if its characteristic vector $\boldsymbol{x}$ for all $j \in X$ satisfies*

$$\boldsymbol{x}^T \hat{A}(\cdot, j) \geq 1 - \bar{\lambda}_j$$

where

$$\bar{\lambda}_j = \begin{cases} \lambda_j - \dfrac{1}{|\mathcal{N}_j|} & \text{if } \lambda_j \cdot |\mathcal{N}_j| \in \mathbb{N} \\ \lambda_j & \text{if } \lambda_j \cdot |\mathcal{N}_j| \notin \mathbb{N} \end{cases}$$

Proof. Firstly we make the following obvious remark:

$$\boldsymbol{x}^T \hat{A}(\cdot, j) = \frac{\boldsymbol{x}^T A(\cdot, j)}{\mathbf{1}^T A(\cdot, j)} = \frac{|X \cap \mathcal{N}_j|}{|\mathcal{N}_j|}.$$

Then we observe that equation (4) can be rewritten as follows:

$$\frac{|X \cap \mathcal{N}_j|}{|\mathcal{N}_j|} > 1 - \lambda_j \quad \Leftrightarrow \quad |X \cap \mathcal{N}_j| > |\mathcal{N}_j| - \lambda_j \cdot |\mathcal{N}_j|. \tag{7}$$

Since the LHS of the last inequality of (7) is an integer, we consider two cases:

- if $\lambda_j \cdot |\mathcal{N}_j| \in \mathbb{N}$ the inequality can be rewritten as:
$$|X \cap \mathcal{N}_j| \geq |\mathcal{N}_j| - \lambda_j \cdot |\mathcal{N}_j| + 1;$$
- if $\lambda_j \cdot |\mathcal{N}_j| \notin \mathbb{N}$ the inequality can be rewritten as:
$$|X \cap \mathcal{N}_j| \geq |\mathcal{N}_j| - \lambda_j \cdot |\mathcal{N}_j|.$$

Dividing these inequalities by $|\mathcal{N}_i|$ the result follows immediately. □

According to the definition of $\bar{\lambda}_j$ introduced in Lemma 2 we define the diagonal matrix $\bar{\Lambda} = diag([\bar{\lambda}_1\ \bar{\lambda}_2\ \ldots\ \bar{\lambda}_n])$.

We can finally show how the maximal cohesive subset of $\mathcal{V} \setminus \phi_0$ can be computed solving an IPP whose optimal solution is the characteristic vector of $\mathcal{M}$.

Proposition 2 *Given a linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$, let $\phi_0 \subseteq \mathcal{V}$ be a seed set with characteristic vector $\boldsymbol{y}$.*

The maximal cohesive set $\mathcal{M}$ contained in $V \setminus \phi_0$ has a characteristic vector $\boldsymbol{x}^*$ that is the optimal solution of the following IPP:

$$\begin{cases} \max \quad \mathbf{1}^T \cdot \boldsymbol{x} & \\ \quad\ \boldsymbol{x} & \leq\ \mathbf{1} - \boldsymbol{y} \\ \left[I - \bar{\Lambda} - \hat{A}^T\right] \cdot \boldsymbol{x} & \leq\ \mathbf{0} \\ \quad \boldsymbol{x} \in \{0,1\}^n & \end{cases} \tag{8}$$

Proof. Firstly, we observe that if $\boldsymbol{x}$ is the characteristic vector of a set $X \subseteq \mathcal{V}$

$$X \cap \phi_0 = \emptyset \qquad \Longleftrightarrow \qquad \boldsymbol{x} + \boldsymbol{y} \leq \mathbf{1},$$

which can be rewritten as the first constraint in (8).

Secondly, if X is a cohesive set, by Lemma 2 it holds

$$\begin{array}{c} \forall j \in \mathcal{M}, \quad \boldsymbol{x}^T \hat{A}(\cdot, j) \geq 1 - \bar{\lambda}_j \\ \Updownarrow \\ \forall j \in \mathcal{V}, \quad \boldsymbol{x}^T \hat{A}(\cdot, j) \geq (1 - \bar{\lambda}_j) x_j \\ \Updownarrow \\ \boldsymbol{x}^T \hat{A} \geq \boldsymbol{x}^T [I - \bar{\Lambda}] \end{array}$$

and this can be immediately rewritten as the second constraint in (8).

Finally, the cohesive set computed by IPP (8) is maximal because of the chosen objective function. □

Note that, as shown in [1] a such a maximal cohesive set always exists – but may be the empty set – and is unique.

Lemma 1 shows that, given a network with seed set ϕ_0, the knowledge of the maximal cohesive set $\mathcal{M} \subseteq \mathcal{V} \setminus \phi_0$ permits an immediate computation of the set of final adopters Φ^*. Thus the linear algebraic characterization of cohesive sets given in Proposition 2 could be used to compute the set final adopters for given seed set as stated in the following corollary.

Corollary 1 *Given a linear threshold network $\mathcal{G}_{LT} = (\mathcal{V}, \mathcal{E}, \lambda)$ and a seed set $\phi_0 \subseteq \mathcal{V}$ with characteristic vector $\boldsymbol{y}$, let $\boldsymbol{x}^* \in \{0,1\}^n$ be the optimal solution of IPP* (8)*. The set of final adopters Φ^* has characteristic vector $\boldsymbol{w}^* = \mathbf{1} - \boldsymbol{x}^*$.*

Proof. Follows from Lemma 1 and Proposition 2. In fact $\boldsymbol{w}^* = \mathbf{1} - \boldsymbol{x}^*$ implies $\Phi^* = \mathcal{V} \setminus \mathcal{M}$. □

Thus we have two different approaches to compute the set of final adopters Φ^*. On one hand we can compute by a single simulation using Algorithm 3. On the other hand, we could solve the Binary Programm Problem in eq. (8).

Computing an IPP is a computationally hard problem, especially for large graphs. However, the main advantage of the characterization in IPP (8) is that it can be used to model and solve other problems, such as the *influence maximization problem* presented in [6] where the objective is to find a seed set of bounded cardinality that ensure the maximal diffusion of the innovation. In problems of this type the set of initial adopters is not known and thus the influence maximization problem cannot be solved by simulation unless one tests all possible seed sets, which is impractical.

References

[1] Acemoglu, D., A. Ozdaglar, and E. Yildiz, "Diffusion of innovations in social networks," *50th IEEE Conf. on Decision and Control and European Control Conference* (Orlando, FL, USA), Dec. 2011.

[2] C. C. Aggarwal, A. Khan, and X. Yan, "On flow authority discovery in social networks," in *Proc. SIAM Int. Conf. on Data Mining* (Mesa, AZ, USA), pp. 522-–533, 2011.

[3] M. Granovetter, "Threshold models of collective behavior." *American Journal of Sociology*, 83:1420–1443, 1978.

[4] D. Kempe, J. Kleinberg, E. Tardos, "Maximizing the spread of influence in a social network." *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[5] Rosa, D., A. Giua, "A non progressive model of innovation diffusion in social networks," *52nd IEEE Conf. on Decision and Control* (Florence, Italy), Dec. 10-13, 2013.

[6] Rosa, D., A. Giua, "On the Spread of Innovation in Social Networks," *NecSys'13: 4th IFAC Workshop on Distributed Estimation and Control in Networked Systems* (Koblenz, Germany), Sep. 25–26, 2013.

[7] T. Schelling, *Micromotives and Macrobehavior*, Norton, 1978.

[8] W.J. Yang, L. Brenner, A. Giua, "Influence Maximization in Independent Cascade Networks Based on Activation Probability Computation," *IEEE Access*, Vol. 7, pp. 13745–13757, 2019.