Enhanced Models for Privacy and Utility in Continuous-Time Diffusion Networks

Daniele Gorla, Federica Granese, Catuscia Palamidessi

Enhanced Models for Privacy and Utility in Continuous-Time Diffusion Networks

Daniele Gorla[1], Federica Granese[1,2], and Catuscia Palamidessi[2]

[1] Dept. of Computer Science, Sapienza University of Rome
[2] INRIA Saclay and LIX

Abstract. Controlling the propagation of information in social networks is a problem of growing importance: On one hand, users wish to freely communicate and interact with their peers. On the other hand, the information they spread can bring to harmful consequences if it falls in the wrong hands. There is therefore a trade-off between utility, i.e., reaching as many intended nodes as possible, and privacy, i.e., avoiding the unintended ones. The problem has attracted the interest of the research community, and some models have already been proposed to study how information propagate and to devise policies satisfying the intended privacy and utility requirements. In this paper we adapt the basic framework of Backes et al. to include more realistic features, that in practice influence the way in which information is passed around. More specifically, we consider: (a) the topic of the shared information, and (b) the time spent by users to forward information among them. For both features, we show a way to reduce our model to the basic one, thus allowing us to extend to our scenario the methods provided in the seminal paper. Furthermore, we propose an enhanced formulation of the utility/privacy policies, to maximize the expected number of reached users among the intended ones, while minimizing this number among the unintended ones, and we show how to adapt the basic techniques to these enhanced policies.

Keywords: Diffusion Networks · Privacy/Utility · Submodular Functions.

1 Introduction

In the last decade there has been a tremendous increase in the world-wide diffusion of social networks, leading to a situation in which a large part of the population is highly connected to other people. A consequence of such high connectivity is that, once a user shares a piece of information, it may spread very quickly. The implications of this phenomenon have attracted the attention of many researchers, interested in studying their potentials and their risks. The involvement of the scientific community with this topic has already produced a large body of literature; see, for instance, [4, 6, 16, 21, 22], just to cite a few.

In general, *diffusion* [14] is a process by which information, viruses, gossips and any other behaviors spread over networks. Here, we follow a natural and

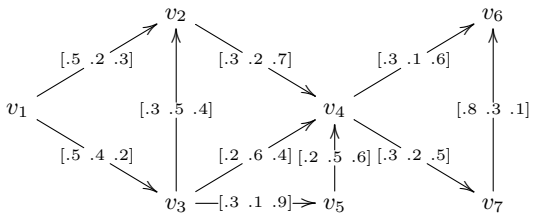

Fig. 1. A Topic vector diffusion network, in which we use topic vectors with three components (*science, movies, society*)

common approach to modeling the net as a graph where nodes represent the users and edges are labeled by the likelihood of transmission along that edge.

One of the strengths, but also the main potential hazard, of social networks relies on the speed by which information can be diffused: once a piece of information becomes viral, there is no way to control it. This means that it can reach users that it was not meant to reach. If the information is a sensitive one, users naturally have an interest in controlling this phenomenon. In [1], this problem is addressed by defining two types of propagation policies that reconcile privacy (i.e., protecting the information from those who should not receive it) and utility (i.e., sharing the information with those who should receive it). In the framework of [1], *utility-restricted privacy policies* minimize the risk, i.e., the expected number of malicious users that receive the information, while satisfying a constraint on the utility, i.e., a lower bound on the number of friends the user wants to reach. Dually, *privacy-restricted utility policies* maximize the number of friends with whom the information is shared, while respecting an upper bound on the number of malicious nodes reached by the information spread. The authors of [1] prove that both these problems are NP-hard, and propose algorithms for approximating the solution.

Being one of the first framework to study the trade-off between privacy and utility, the model proposed in [1] is quite basic. One limitation is that the likelihood that governs the transmission along an edge is a constant, fixed in time and irrespective of any other features. We argue that this is not a realistic assumption, and we propose to enrich the framework so to be able to model the situations described in the following two scenarios.

First, imagine that you are a scientific researcher spending some time on a social network. Suddenly, you see a news about the proof of the century, stating that P = NP. Whom do you wish to share such an information with? Probably with a colleague or someone interested in the subject. To support this kind of scenario, following [7], we consider social networks in which a user may choose the peers to whom to send a piece of information based on the *topic* of that information. To model such a situation, we label the edges of the net by *topic vectors*, defined as vectors in which each component represents the probability of

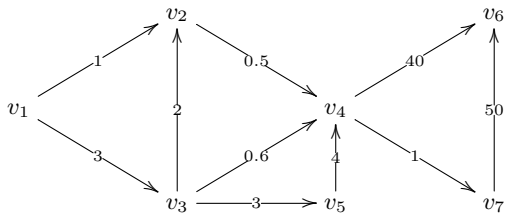

Fig. 2. A Time diffusion network with sampled times for traversing the edge

a user to send an information of the corresponding topic (or tag) to the user at the other end of the edge. Furthermore, a piece of information is usually related to several topics, not just one. To model this latter aspect, we also tag a message with a probability distribution (topic distribution) over the topics, representing the weight of each topic in the message. To obtain the probability that a node v_i sends a message to another node v_j we then consider the scalar product of the topic vector of the edge (v_i, v_j) and the topic distribution of the message.

As an example, assume that there are three topics, *science*, *movies*, and *society*. Figure 1 represents a net whose edges are labeled with instances of these kinds of topic vectors. For example, if v_3 receives a message about a new movie of a director he likes, the probability that it will forward it to v_2 (rather than not) is 0.5, while the probability of forwarding it to v_4 is 0.6 and to v_5 is 0.1, representing the fact that v_2 and v_3 are much more interested in the kind of movies that v_3 likes than v_5 is. Note that the sum of these probabilities is not 1, because these are independent events. Further, consider the P=NP message, and assume that its topic distribution is $(0.9, 0, 0.1)$. Since since the edge (v_7, v_6) has topic vector $(0.8, 0.3, 0.1)$, the probability that v_7 sends the message to v_6 is $0.9 \times 0.8 + 0 \times 0.3 + 0.1 \times 0.1 = 0.73$. Note that, being the convex combination of probabilities, the result of such scalar product is always a probability.

Second, imagine that you are a night owl; at midnight, you see a funny photo and you want to share it with one of your friends. However, he is a sleepyhead and sleeps all night; thus, he will be able to forward such a photo only the next morning. If we are tracking the diffusion process until a few hours forward, there will be no further diffusion of the photo from your friend. On the other hand, if you had sent the photo during the day, he may have seen and forwarded it soon afterwards. This scenario can be modeled by labeling each edge (v_i, v_j) with a probability density function over time δ_{ij}, representing the probability that the information takes a certain time t for traveling from v_i to v_j. For instance, if v_i is the night owl and v_j the sleepyhead, then, it is likely that δ_{ij} will be a big amount of time, but there is still some probability that the information arrives at v_i when they are both awake, in which case the transmission time will be shorter. Each edge may have a different density function: for instance, if v_i has another friend v_z who is a night owl as well, then the moment in which v_z sees the

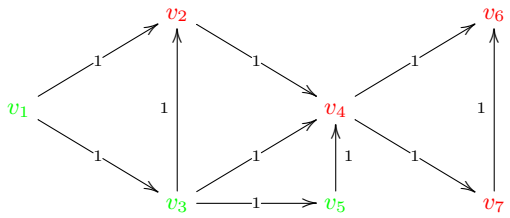

Fig. 3. A General diffusion network in which green nodes are friends and red nodes are malicious.

information sent by v_i will be likely to be closer to the one in which v_i forwards the information; hence, the amount of time for the transmission from v_i to v_z will be small. By sampling the time for each edge we obtain a snapshot of the net, which will have the same structure as a standard net. Figure 2 represents an instance of such a net.

Another limitation of the standard framework is in the way the trade-off problem is formulated: for maximizing privacy and utility, the corresponding problems try to minimize the number of malicious nodes infected *up to time* t (given a bound on the number of friends *initially* sharing the information), or to maximize the number of friends *initially* sharing the information (given a bound on the number of malicious nodes infected *up to time* t). We argue that utility would be better expressed in terms of the friends reached by the information *up to time* t, instead of the initial friends only. Furthermore, privacy and utility would be more symmetric, in that both of them would be expressed in terms of nodes reached at time t.

As an example, consider Figure 3 and suppose we want to monitor the diffusion up to time $t = 1$. Consider first the maximum utility problem under the constraint of reaching (at time $t = 1$) one malicious node at most. In the standard framework there are two solutions for the set of initial nodes: either $\{v_1\}$ or $\{v_5\}$. They are considered equivalent because we only consider further infection of the malicious nodes (and in both cases, in 1 time unit just one malicious node gets infected). In contrast, we argue that $\{v_1\}$ is a better solution, because if we start with $\{v_1\}$ then in 1 time unit the information will reach also the friend node v_3, while no further friends will be reached if we start with $\{v_5\}$.

Consider now the maximum privacy problem. Assume that we want to minimize the number of malicious nodes infected up to time $t = 1$ under the constraint of having at least two friend sharing the information. The solution of the problem in [1] is any subset formed by two friend nodes. Any such subset, in fact, leads to infect two malicious nodes at time $t = 1$. In contrast, we argue that the optimal solution would be the (smaller) initial set $\{v_1\}$. In fact this solution would respect the constraint if, *as we propose*, we did count also the friends infected at time $t = 1$, and would minimize the malicious nodes infected in the same time unit.

1.1 Related Work

There is a huge literature on information propagation in social networks, but most of the papers focus on maximizing the spread of information in the whole network. See for instance [5, 9, 11, 15, 19]. To make such works closer to real life situations, some papers revisit them on either the influence problem or the network model. For example, in [2, 3, 20], the problem is modified by considering the scenario where a company wants to use viral marketing to introduce a new product into a market when a competing product is simultaneously being introduced. Referring to A and B as the two technologies of interest, they denote with I_A (I_B) the initial set of users adopting technology A (B). Hence, they try to maximize the expected number of consumers that will adopt technology A, given I_A and I_B, under the assumption that consumers will use only one of the two products and will influence their friends on the product to use. In [2], the authors consider the problem of limiting the spread of misinformation in social networks. Considering the setting described before (with the two competitive companies), they refer to one of the two companies as the "bad" company and to the other one as the "good" company.

In the papers mentioned so far, authors always assume that all the selected top influential nodes propagate influence as expected. However, some of the selected nodes could not work well in practice, leading to influence loss. Thus, the objective of [24] is to find the set K of the most influential nodes with which initially the information should be shared, given a threshold on influence loss due to a failure of a subset of nodes $R \subseteq K$. This problem, as all the previous ones, are proven to be NP-hard; furthermore, all of [2, 3, 20, 24] assume that the diffusion process is timeless.

A different research line consists in making the underlying network model closer to reality, instead of modify the problem itself. For example, topic of information is handled in [7], where the authors infer what we call topic vector. Always considering the information item, the model in [23] endows each node with an influence vector (how authoritative they are on each topic) and a receptivity vector (how susceptible they are on each topic). While for diffusion network there exists a good amount literature about the role of users' interests [7, 23, 25, 26], the same is not true for the role of the time with respect to user habits.

An orthogonal research line is represented by works like [7, 10], aiming at inferring transmission likelihoods: given the observed infection times of nodes, they infer the edges of the global diffusion network and estimate the transmission rates of each edge that best explain the observed data. This leads to an interesting problem that can be solved with convex optimization techniques. Note that, as in [1], we are not dealing with this aspect, since we assume that the inference has already happened and we have an accurate estimate of the transmission likelihoods (whatever they are) for the whole network.

1.2 Contribution

The contribution of our paper is the following:

– We extend the basic graph diffusion model proposed in [1] by considering a more sophisticate labeling of the edges. This allows to take into account, for the propagation of information, (a) the topics and (b) the probabilistic nature of the transmission rates.
– We reformulate the optimization goals of [1] by considering a notion of utility which takes into account the friend nodes reached up a certain time t, rather than the initial set only. We argue that this notion is more natural, besides being more in line with that of privacy (the infected malicious nodes are counted up to time t as well).
– We prove that the resulting optimization problems are NP-hard.
– We modify and adapt to our framework the techniques proposed in [1] to approximate the solution in polynomial time.

1.3 Paper Organization

This paper is organized as follows. In Section 2, we recall the basic notions and results from [1]. Then, in Section 3, we present the two enhanced models, one where information transmission is ruled by the topic of conversation, the other one based on the transmission time. In Section 4, we then modify the basic definitions of utility-restricted privacy policies and privacy-restricted utility policies, and show that all the theory developed by [1] with the original definitions can be smoothly adapted to these new (and more realistic) definitions. Finally, in Section 5, we conclude the paper, by also drawing lines for future research.

2 Background

In this section we recall the basic notions from [1], which will be used in the rest of the paper.

2.1 Submodular Functions

Definition 1 (Submodular function [8]). *A function* $f\colon 2^V \to \mathbb{R}$ *is* submodular *if, for all* $S, T \subseteq V$*, it holds that* $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.

Defining $f(j|S) := f(S \cup \{j\}) - f(S)$ as the *profit* (or *cost*) of $j \in V$ in the context of $S \subseteq V$, then f is submodular iff $f(j|S) \geq f(j|T)$, for all $S \subseteq T$ and $j \notin T$. The function f is *monotone* iff $f(j|S) \geq 0$, for all $S \subseteq V$ and $j \notin S$. Moreover, f is *normalized* if $f(\emptyset) = 0$. Given a submodular function f, the *curvature* κ_f of f is

$$\kappa_f := min_{j \in V} \frac{f(j|V \setminus \{j\})}{f(\{j\})}$$

Optimizing submodular functions is a difficult task, but we can get around the problem by choosing a proper surrogate function for f and optimize it; the surrogate functions usually are upper or lower bounds. For example, the *majorization-minimization* algorithms begin with an arbitrary solution Y to the optimization

problem and then optimize a modular approximation formed via the current solution Y. Let X be the new solution (under construction); if we now let

$$m_{g_Y}(X) = f(Y) + g_Y(X) - g_Y(Y) \qquad m_{h_Y}(X) = f(Y) + h_Y(X) - h_Y(Y)$$

where g_Y and h_Y are defined as in [13], an upper bound for minimization and a lower bound for maximization can be:

$$m_{g_Y} \geq f(X) \qquad m_{h_Y} \leq f(X)$$

Both these bounds are tight at the current solution, i.e. $m_{g_Y}(Y) = m_{h_Y}(Y) = f(Y)$.

2.2 Diffusion Networks

Definition 2 (General Diffusion Network). *A general diffusion network is a tuple $N = (V, \gamma)$, where $V = \{v_i\}_{i=1\ldots n}$ is the set of nodes and $\gamma = (\gamma_{ij})_{i,j=1\ldots n}$ is the transmission matrix of the network (with $\gamma_{ij} \geq 0$, for all i, j).*

Thus, V and γ define a directed graph where each $\gamma_{ij} > 0$ represents an edge between nodes v_i and v_j along which the information can potentially flow, together with the flow likelihood. Let us now consider a general diffusion network N in which $F \subseteq V$ is the set of friendly nodes and $M \subseteq V$ is the set of malicious nodes, with $F \cap M = \emptyset$. The idea is to maximize the number of friends and minimizing the number of enemies reached by an information in a certain time window.

Definition 3 (Utility-restricted Privacy Policy). *A utility-restricted privacy policy Π is a 4-tuple $\Pi = (F, M, k, t)$ where F is the set of friend nodes, M is the set of malicious nodes, k is the number of nodes the information should be shared to, and t is the period of time in which the policy should be valid.*

Definition 4 (Privacy-restricted Utility Policy). *A privacy-restricted utility policy Υ is a 4-tuple $\Upsilon = (F, M, \tau, t)$ where F is the set of friend nodes, M is the set of malicious nodes, τ is the expected number of nodes in M receiving the information during the diffusion process, and t is the period of time in which the policy should be valid.*

Both the policies are focused on bounding the risk that a malicious node gets infected by time t, given that $F' \subseteq F$ is initially infected.

Definition 5 (Risk). *Let N be a diffusion network. The risk $\rho_N(F', M, t)$ caused by $F' \subseteq V$ with respect to $M \subseteq V$ within time t is given by*

$$\rho_N(F', M, t) = \sum_{m_i \in M} Pr[t_i \leq t | F']$$

Here, $Pr[t_i \leq t|F']$ is the likelihood that the infection time t_i of malicious node m_i is at most t, given that F' is infected at time $t = 0$.

To make notation lighter, we shall usually omit the subscript N from ρ_N, when clear from the context. To maximally satisfy a utility-restricted privacy policy and a privacy-restricted utility policy, the following two problems are defined.

Definition 6 (Maximum k-privacy – MP). *Given a utility-restricted privacy policy $\Pi = (F, M, k, t)$ and a general diffusion network N, the maximum k-privacy problem (MP, for short) is given by*

$$\begin{array}{ll} \underset{F' \subseteq F}{minimize} & \rho(F', M, t) \\ subject\ to & |F'| \geq k \end{array} \tag{1}$$

Definition 7 (Maximum τ-utility – MU). *Given a privacy-restricted utility policy $\Gamma = (F, M, \tau, t)$ and a general diffusion network N, the maximum τ-utility problem (MU, for short) is given by*

$$\begin{array}{ll} \underset{F' \subseteq F}{maximize} & |F'| \\ subject\ to & \rho(F', M, t) \leq \tau \end{array} \tag{2}$$

Both problems are NP-hard. However, they can be approximated and the approximation algorithms rely on the submodularity of the risk function: by showing that ρ is a submodular monotone function with a non-zero curvature, it is possible to derive an efficient constant factor approximation, where the approximation factor depends on the structure of the underlying network N. Recall that the curvature $\kappa_{\rho(F,M,t)}$ of $\rho(F, M, t)$ is given by $\kappa_{\rho(F,M,t)} := min_{v \in F} \frac{\rho(v|F \setminus \{v\}, M, t)}{\rho(\{v\}, M, t)}$ where $\rho(v|F \setminus \{v\}, M, t) := \rho(F, M, t) - \rho(F \setminus \{v\}, M, t)$.

Theorem 1. *There is an efficient algorithm A that approximates maximum k-privacy to a factor $\frac{1}{\kappa_\rho}$. That is, let F' be the output of A and F^* be the optimal solution; then,*

$$\rho(F', M, t) \leq \frac{1}{\kappa_\rho} \rho(F^*, M, t)$$

Algorithm 1 Maximum τ-Utility

Require: Instance F, M, τ of maximum τ-utility
Ensure: $satisfyingMU(F, M, \tau)$
  for $n \in [|F|, \ldots, 1]$ **do**
    $\tau' \leftarrow min_{F' \subseteq F} \rho(F', M, t)$ s.t. $|F'| = n$
    if $\tau' \leq \tau$ **then**
      return n
  return 0

Starting from the approximation algorithm for maximum k-privacy, maximum τ-utility can be approximated through Algorithm 1.

Theorem 2. *Let n^* be the optimal solution to an instance of maximum τ-utility, and let n be the output of Algorithm 1 for the same instance, using a $\frac{1}{\kappa_\rho}$-approximation for maximum k-privacy. Then $n \geq \kappa_\rho n^*$.*

3 Enhanced Models

In this section, we provide two different models which modify the notion of general diffusion network by using different transmission matrices. In particular, in the first model, called *topic vector diffusion network*, we bind the likelihood of transmitting an information to the topic of that information; in the second one, called *time diffusion network*, we bind the likelihood to the amount of time an information takes for been transmitted. As in [1], we are not interested in the inference of transmission likelihoods, as the aim of the following two models is the reduction to the general model for which the two kinds of policies are defined.

3.1 Topic Vector Diffusion Network

We first consider a social network where edges are labeled by *topic vectors*, that are vectors in which each component represents the probability of a user to send an information of the corresponding topic (or tag) to another user.

Definition 8 (Topic Vector Diffusion Network). *A topic vector diffusion network is a tuple $N_{TV} = (V, \mathbf{A}, k)$, where $V = \{v_i\}_{i=1\dots n}$ is the set of nodes in the network, k is the number of topics and $\mathbf{A} = (\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_n})$ is s.t. $\boldsymbol{\alpha_i}$ is the matrix of dimension $n \times k$ giving the topic vector that rules the transmission rates from node v_i to all the other nodes in the network. That is,*

$$\boldsymbol{\alpha_i} := \begin{pmatrix} \alpha_{i1}^1 & \dots & \alpha_{i1}^k \\ \vdots & & \vdots \\ \alpha_{in}^1 & \dots & \alpha_{in}^k \end{pmatrix}.$$

where every $\boldsymbol{\alpha_{ij}} = (\alpha_{ij}^1 \dots \alpha_{ij}^k)$ is called topic vector *and each α_{ij}^l (for $l = 1 \dots k$) is the probability that user i sends an information of topic l to user j.*

Notice that a topic vector is not required to be a probability distribution and that, for every i, j and l, the probability of not sending an information of topic l from i to j is $1 - \alpha_{ij}^l$. Together, V and $\mathbf{A}$ define a weighted directed graph where each $\boldsymbol{\alpha_{ij}}$ (i.e. each row of $\boldsymbol{\alpha_i}$ having non zero components) represents an edge between v_i and v_j with weight $\boldsymbol{\alpha_{ij}}$. For example, consider the network N_{TV} in Figure 4(a), with $V = \{v_1, v_2, v_3\}$, $k = 2$ and

$$\boldsymbol{\alpha_1} = \begin{pmatrix} 0 & 0 \\ 0.6 & 0.5 \\ 0.4 & 0.9 \end{pmatrix}, \ \boldsymbol{\alpha_2} = \begin{pmatrix} 0.5 & 0.5 \\ 0 & 0 \\ 0.3 & 0.8 \end{pmatrix}, \ \boldsymbol{\alpha_3} = \begin{pmatrix} 0.5 & 0.6 \\ 0.7 & 0.6 \\ 0 & 0 \end{pmatrix}$$

User v_1 will send to v_2 an information about topic 1 with probability 0.6 and an information about topic 2 with probability 0.5.

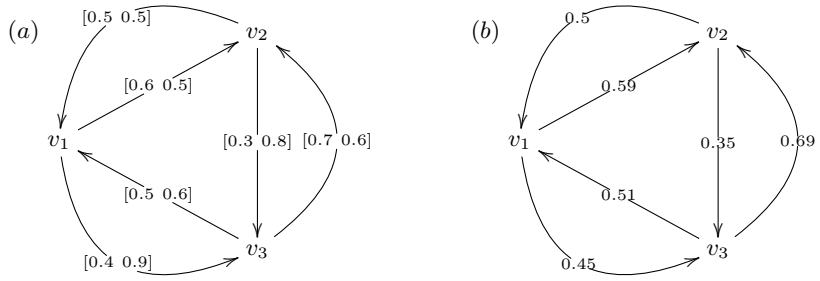

Fig. 4. From a topic vector diffusion network to the $\boldsymbol{m}$-diffusion network. (a) A topic vector diffusion network. (b) The associated (0.9 0.1)-Diffusion network

Definition 9 (Information Item). *An information item (or meme) is a k-dimensional probability vector, in which each component is the weight of a topic relating to the subject of the information. That is,* $\boldsymbol{m} := (m_1 \dots m_k)$ *such that* $m_1 + \ldots + m_k = 1$.

For instance, consider vectors consisting of two components, *science* and *society*. The information item associated to a tweet on a scientific paper should be $\boldsymbol{m} = (0.9\ 0.1)$.

Remark 1. A topic vector is different from a meme since it is not a probability vector (indeed, each component of a topic vector is itself a probability).

Definition 10 (Probability of Infection Information Item). *Let* N_{TV} *be a topic vector diffusion network,* $i, j \in V$ *and* $\boldsymbol{m}$ *the input meme. Then, the probability that* i *sends* $\boldsymbol{m}$ *to* j *is given by:*

$$\beta_{ij\boldsymbol{m}} = \boldsymbol{\alpha_{ij}}\, \boldsymbol{m}^\top \tag{3}$$

Notice that, since each component of $\boldsymbol{\alpha_{ij}}$ is a probability and $\boldsymbol{m}$ is a probability vector, we obtain:

$$0 = \mathbf{0}\, \boldsymbol{m}^\top \leq \boldsymbol{\alpha_{ij}}\, \boldsymbol{m}^\top \leq \mathbf{1}\, \boldsymbol{m}^\top = 1$$

Definition 11 ($\boldsymbol{m}$-Diffusion Network). *An* $\boldsymbol{m}$*-diffusion network is a tuple* $N_m = (V, \beta_{\boldsymbol{m}})$*, where* $V = \{v_i\}_{i=1\dots n}$ *is the set of nodes and* $\beta_{\boldsymbol{m}} = (\beta_{ij\boldsymbol{m}})_{i,j=1\dots n}$ *is the transmission matrix of the network that forwards* $\boldsymbol{m}$ *(with* $\beta_{ij\boldsymbol{m}} \geq 0$*).*

Given a topic vector diffusion network and an information item, we can derive the associated $\boldsymbol{m}$-diffusion network by determining the probability of infection between each node with respect to the information item (i.e the transmission matrix $\beta_{\boldsymbol{m}}$). Resuming the example before, with $\boldsymbol{m} = (0.9\ 0.1)$ representing the information item of a scientific paper, consider the topic diffusion network in Figure 4(b), in which we suppose the topic vectors have the same tag as $\boldsymbol{m}$ (science and society). By Definition 10, we have, e.g., that $\beta_{32\boldsymbol{m}} = (0.7\ 0.6)(0.9\ 0.1)^\top = 0.69$

and $\beta_{31\boldsymbol{m}} = (0.5\ 0.6)(0.9\ 0.1)^\top = 0.51$; hence, the probability that v_3 forwards $\boldsymbol{m}$ to v_2 is greater than the probability of forwarding to v_1, since $\boldsymbol{m}$ is more focused on science than on society.

Even if the $\boldsymbol{m}$-diffusion network seems similar to the general diffusion network, it still has an important difference: it depends on the information item. Thus, consider a sample of messages $M = \{\boldsymbol{m}_1, \ldots, \boldsymbol{m}_h\}$ and their associated $\boldsymbol{m}_l$-diffusion networks derived from the same topic vector diffusion network. Let us concentrate on two nodes i, j in V and define the independent events $E_{ijl} = \{i \textit{ sends } \boldsymbol{m}_l \textit{ to } j\}$; clearly, $Pr(E_{ijl}) = \beta_{ij\boldsymbol{m}_l}$. We can define a random variable X_{ij} counting the number of information items in M sent from i to j. Thus, we can compute the probability that i sends $0, 1, \ldots, h$ information items to j as follows:

$$
\begin{aligned}
Pr(X_{ij} = 0) &= \prod_{l=1}^{h} (1 - \beta_{ij\boldsymbol{m}_l}) \\
&\vdots \\
Pr(X_{ij} = d) &= \sum_{\{l_1,\ldots,l_d\} \subseteq \{1,\ldots,h\}} \beta_{ij\boldsymbol{m}_{l_1}} \ldots \beta_{ij\boldsymbol{m}_{l_d}} \left(\prod_{l' \in \{1,\ldots,h\} \setminus \{l_1,\ldots,l_d\}} (1 - \beta_{ij\boldsymbol{m}_{l'}}) \right) \\
&\vdots \\
Pr(X_{ij} = h) &= \prod_{l=1}^{h} \beta_{ij\boldsymbol{m}_l}
\end{aligned}
$$

The derivation of the general diffusion network from a set of $\boldsymbol{m}_l$-diffusion networks (obtained from the same topic vector diffusion network) is given by first computing for each $i, j \in V$

$$E[X_{ij}] = \sum_{d=1}^{h} dPr\ (X_{ij} = d),$$

Then, by starting from these expected values, we can recover a general diffusion network, by still considering V as set of nodes and by setting $\gamma_{ij} = E[X_{ij}]$, for every i, j.

3.2 Time Diffusion Network

We now consider a diffusion network in which each edge (v_i, v_j) is equipped with a probability density function describing, for any given time interval (providing the time spent by the information in traveling along it), the probability of transmitting along that edge.

Definition 12 (Time transmission function). *A time transmission function $f(\delta)$ is a density over time.*

Definition 13 (Time diffusion network). *A time diffusion network is a tuple* $N_T = (V, \zeta)$*, where* $V = \{v_i\}_{i=1...n}$ *is the set of nodes in the network and* $\zeta = (f_{ij}(\delta_{ij}))_{i,j=1...n}$*, with* $f_{ij}(\cdot)$ *a time transmission function and* δ_{ij} *a time interval (for every* i *and* j*), is the transmission matrix of the network.*

In contrast with the discrete-time model (which associates each edge with a fixed infection probability), this model associates each edge with a probability density function. Moreover, instead of considering parametric transmission functions such as exponential distribution, Pareto distribution or Rayleigh distribution, we consider the non-parametric ones because in real word scenarios the waiting times obey to different distributions. So, for example, if two nodes are usually logged simultaneously (hence, their respective delay in transmission is small), the time function will assign high probabilities to short intervals and negligible probabilities to long ones; the situation is dual for users that are usually logged in different moments of the day.

Now suppose that some external agent gives in input to some nodes of the network a certain information at time $t = 0$. Each of these nodes try to forward this information to their neighbors; clearly, this entails a certain amount of time.

Definition 14 (Transmission time). *Given two neighbor nodes* i *and* j *of a time diffusion network, the transmission time* δ_{ij} *is the amount of time the information requires for going from* i *to* j *during a diffusion process.*

Starting from a time diffusion network N_T, we can compute the random transmission times associated to each edge on the network by drawing them from the corresponding transmission functions. Consider now a diffusion process over a time diffusion network N_T and suppose that the initial set of infected nodes is F'.

Definition 15 (Infection time of a node [11]). *The infection time of* $v \in V$ *is given by:*

$$t_v = g_v(\{\delta_{ij}\}_{(i,j)\in N_T}|F') := min_{q\in Q_v(F')} \sum_{(i,j)\in q} \delta_{ij}$$

where F' *is the set of nodes infected at time* $t = 0$ *and* $Q_v(F')$ *is the set of the directed paths from* F' *to* v*.*

For preserving Theorems 1 and 2 also in this setting, we must first prove submodularity of the risk function on time diffusion networks. For this purpose, let us slightly modify Definition 5.

Definition 16 (Risk). *Let* $N_T = (V, \zeta)$ *be a time diffusion network. The risk* $\rho_{N_T}(F', M, t)$ *caused by* $F' \subseteq V$ *with respect to* $M \subseteq V$ *within time* t *is given by*

$$\rho(F', M, t) = \sum_{m_i \in M} Pr[t_i \leq t|F']$$

Here, $Pr[t_i \leq t|F'] = Pr[g_v(\{\delta_{ij}\}_{(i,j)\in N_T}|F') \leq t]$ *is the likelihood that the infection time* t_i *of malicious node* m_i *is at most* t*, given that* F' *is infected at time* $t = 0$*.*

Theorem 3. *Given a time diffusion network $N_T = (V, \zeta)$, a set of friend nodes $F \subseteq V$, a set of malicious nodes $M \subseteq V$ and a time window t, the risk function $\rho_{N_T}(F, M, t)$ is monotonically nondecreasing and submodular in F.*

Proof. By definition, all nodes in F are infected at time $t = 0$. The infection time of a given node in the network only depends on the transmission times drawn from the transmission functions. Thus, given a sample $\{\delta_{ij}\}_{(i,j)\in N_T}$, we define $r_{\{\delta_{ij}\}}(F, M, t)$ as *the number* of nodes in M that can be reached from the nodes in F at time less than or equal to t for $\{\delta_{ij}\}$; and $R_{\{\delta_{ij}\}}(f, M, t)$ as *the set* of nodes in M that can be reached from the node f at time less than or equal to t for $\{\delta_{ij}\}$.

(i) $r_{\{\delta_{ij}\}}(F, M, t)$ is monotonically nondecreasing in F, for any sample $\{\delta_{ij}\}$. Indeed, $r_{\{\delta_{ij}\}}(F, M, t) = |\cup_{f\in F} R_{\{\delta_{ij}\}}(f, M, t)|$ and so, for any $n \notin V \setminus (F \cup M)$, $r_{\{\delta_{ij}\}}(F, M, t) \leq r_{\{\delta_{ij}\}}(F \cup \{n\}, M, t)$.

(ii) $r_{\{\delta_{ij}\}}(F, M, t)$ is submodular in F for a given sample $\{\delta_{ij}\}$. Let $R_{\{\delta_{ij}\}}(f|B, M, t)$ defined as the set of nodes in M that can be reached from node f in a time shorter than t, but cannot be reached from any node in the set of nodes $B \subseteq V$ for $\{\delta_{ij}\}$. For any $B \subseteq B'$ it holds that $|R_{\{\delta_{ij}\}}(f|B, M, t)| \geq |R_{\{\delta_{ij}\}}(f|B', M, t)|$. Consider now two sets of nodes $B \subseteq B'(\subseteq V)$ and a node $b \notin B'$:

$$\begin{aligned}
&r_{\{\delta_{ij}\}}(B \cup \{b\}, M, t) - r_{\{\delta_{ij}\}}(B, M, t) \\
&= |R_{\{\delta_{ij}\}}(b|B, M, t)| \\
&\geq |R_{\{\delta_{ij}\}}(b|B', M, t)| \\
&= r_{\{\delta_{ij}\}}(B' \cup \{b\}, M, t) - r_{\{\delta_{ij}\}}(B', M, t)
\end{aligned}$$

If we average over the probability space of possible transmission times,

$$\rho_{N_T}(F, M, t) = E_{\{\delta_{ij}\}\in N_T}[r_{\{\delta_{ij}\}}(F, M, t)]$$

is also monotonically nondecreasing and submodular. □

Given a time diffusion network, if the risk function has a nonzero curvature, then the results of [1] hold also for this model. Let $S_{ij}(\delta_{ij})$ be the *survival function*, expressing the probability of v_j not being infected by node v_i. Formally, $S_{ij}(\delta_{ij}) := 1 - \int_0^{\delta_{ij}} f_{ij}(\delta')d\delta'$.

Theorem 4. *Let $N_T = (V, \zeta)$ be a time diffusion network, for which $S_{ij}(\delta_{ij}) > 0$ until time t for all $v_i, v_j \in V$. Then $\kappa_{\rho(F,M,t)} > 0$.*

Proof. The infection time of a given node in the network only depends on the transmission times drawn from the transmission functions. Thus, given a sample $\{\delta_{ij}\}_{(i,j)\in N_T}$, we first remove all $v_i \in F$ s.t. $\rho_{N_T}(\{v_i\}, M, t) = 0$, since they can be safely infected at time $t = 0$. Now pick an arbitrary $v \in F$, thus there exists a dipath P from v to some $v_m \in M$. Since by hypothesis the survival function is nonzero until time t for all pairs of nodes on the path, then $\prod_{(i,j)\in P} S_{ij}(\delta_{ij}) > 0$.

This fact, together with Equations (2) and (6) of [11], entails that the likelihood of infection of every node on this path is decreased if this path is removed. Moreover, this implies $\rho_{N_T}(F, M, t) - \rho_{N_T}(F \setminus \{v\}, M, t) > 0$. Thus, by definition of curvature, we obtain $\rho_{N_T}(v|F \setminus \{v\}, M, t) > 0$ and therefore $\kappa_{\rho_{N_T}(F,M,t)} > 0$. □

4 Policy Enhancements

Let us consider a general diffusion network $N = (V, \zeta)$, with fixed and disjoint sets of friend nodes F and of malicious nodes M. Starting from the propagation policies given for the basic framework in Section 2, we give a new definition for when an initial infection $F' \subseteq F$ within a network satisfies a utility-restricted privacy policy or a privacy-restricted utility policy. To this aim, we first introduce the notion of *gain*.

Definition 17 (Gain). *The gain $\pi(F', F, t)$ caused by $F' \subseteq F$ within time t is given by*

$$\pi(F', F, t) = \sum_{f_i \in F} Pr[t_i \leq t|F']$$

Here, $Pr[t_i \leq t|F']$ is the likelihood that the infection time t_i of a friend node f_i is at most t, given that F' is infected at time $t = 0$.

Hence, the gain function is similar to the risk function but, instead of determining the expected number of infected nodes in M, it gives us the expected number of infected nodes in F. Clearly, since our gain function $\pi(F', M, t)$ derives from the risk function $\rho(F', M, t)$, computing $\pi(F', M, t)$ is also $\#P$-hard. We follow the approach in [1] for the risk function, assuming to have an oracle that exactly computes the gain function for a given initial infection F'.

Definition 18 (Satisfy a Utility-restricted Privacy Policy). *An initial infection F' satisfies a utility-restricted privacy policy $\Pi = (F, M, k, t)$ in a general diffusion network N if $F' \subseteq F$ and $\pi(F', M, t) \geq k$. A set F' maximally satisfies Π in N if there is no other set $F'' \subseteq F$ with $\pi(F'', F, t) \geq k$ and $\rho(F'', M, t) < \rho(F', M, t)$.*

Definition 19 (Satisfy a Privacy-restricted Utility Policy). *An initial infection F' satisfies an extended privacy-restricted utility policy $\Upsilon = (F, M, \tau, t)$ in a general diffusion network N if $F' \subseteq F$ and $\rho(F', M, t) \leq \tau$. A set F' maximally satisfies Υ in N if there is no other set $F'' \subseteq F$ with $\rho(F'', M, t) \leq \tau$ and $\pi(F'', F, t) > \pi(F', F, t)$.*

For finding an initial infection meeting Definitions 18 and 19, we define the following problems.

Definition 20 (Extended Maximum k-Privacy - EMP). *Given a utility-restricted privacy policy $\Pi = (F, M, k, t)$ and a general diffusion network N, the extended maximum k-privacy problem (EMP, for short) is given by*

$$\begin{aligned} \underset{F' \subseteq F}{minimize} \quad & \rho(F', M, t) \\ subject\ to \quad & \pi(F', F, t) \geq k \end{aligned}$$

Definition 21 (Extended Maximum τ-Utility – EMU). *Given a privacy-restricted utility policy $\Upsilon = (F, M, \tau, t)$ and a general diffusion network N, the extended maximum τ-utility problem (EMU, for short) is given by*

$$\begin{aligned} \underset{F' \subseteq F}{maximize} \quad & \pi(F', F, t) \\ subject\ to \quad & \rho(F', M, t) \leq \tau \end{aligned}$$

Clearly, if F' is an optimal solution to the EMP problem with respect to Υ, then F' maximally satisfies Υ and if F' is an optimal solution to the EMU problem with respect to Π, then F' maximally satisfies Π.

Unfortunately, EMP and EMU problems are NP-hard; this can be proved by reducing MP and MU to them.

Theorem 5. *Extended maximum k-privacy and extended maximum τ-utility are NP-hard.*

Proof. We just show the reduction of MU to EMU since the other one is symmetric. Let ϕ be an instance of the MU problem, we can construct an instance of the EMU problem ω by setting the time parameter of the gain function to $t = 0$. Hence, F' is the seed set of ϕ, respecting the risk constraint, iff F' is the maximum set of initially infected nodes always respecting the risk constraint. As the EM problem is NP-hard [1], also EMU is NP-hard. □

Remark 2. As the gain function is different from the risk function only for the set in which the propagation of the information is monitored, the same proof for the submodularity of ρ in [1] can be adopted to show the submodularity of π.

Following the work in [1, 12], we can solve EMP and EMU problems by choosing surrogate functions for both π and ρ. In particular, EMP and EMU problems can be solved by slightly modifying the algorithms in [1]. Recalling Section 2.1, a strategy for optimizing submodular function is based on choosing a surrogate function and optimize it. In Algorithm 2, the surrogate function of the risk function is defined as in [1]. Thus, given a candidate solution $Y \subseteq F$, the modular approximation of the risk function ρ is given by

$$m_{g_Y}(X) = \rho(Y) + g_Y(X) - g_Y(Y)$$

where

$$g_Y(X) = \sum_{v \in X} g_Y(v) \qquad \text{and} \qquad g_Y(v) = \begin{cases} \rho(v|F \setminus \{v\}), & \text{if } v \in Y \\ \rho(v|Y), & \text{otherwise.} \end{cases}$$

At each iteration, Algorithm 2 finds the new set that minimizes the upper bound of the risk function. Clearly, since this set minimizes the upper bound of the risk function, it also minimizes the risk function.[1]

Algorithm 2 Extended Maximum k-Privacy

Require: Instance F, M, k of extended maximum k-privacy
Ensure: *satisfyingEMP*(F, M, k)
  $C \leftarrow \{X \subseteq F \mid \pi(X, F, t) = k\}$
  Select a random candidate solution $X^1 \in C$
  $t \leftarrow 0$
  repeat
    $t \leftarrow t + 1$
    $X^{t+1} \leftarrow argmin_{X \in C} m_{g_{X^t}}(X)$
  until $X^{t+1} = X^t$
  return X^t

To conclude, notice that Algorithm 1 can be easily adapted for handling the new definition of Maximum τ-utility: it suffices to replace "$|F'| = n$" with "$\pi(F', F, t) = n$" in the calculation of τ'. For completeness, we report it as Algorithm 3.

Algorithm 3 Extended Maximum τ-Utility

Require: Instance F, M, τ of extended maximum τ-utility
Ensure: *satisfyingEMU*(F, M, τ)
  for $n \in [|F|, \ldots, 1]$ **do**
    $\tau' \leftarrow min_{F' \subseteq F} \rho(F', M, t)$ s.t. $\pi(F', F, t) = n$
    if $\tau' \leq \tau$ **then**
      return n
  return 0

5 Conclusion

In this paper, we proposed some enhancements of the basic model in [1] for controlling utility and privacy in social networks. In particular, we added topics of conversation and time of the infection within the transmission likelihood. Furthermore, we modified the basic definitions of policy satisfaction, to make them closer to the intuitive meaning of such policies. Then, we extended the methods and results of [1] to our setting. We have demonstrated the applicability of our

[1] This methodology can be seen as the gradient descent method for minimizing continuous differentiable functions: we start from a random point y and we iteratively move in the direction of the steepest descent, as defined by the negative of the gradient.

enhanced framework on various situations. Arguably these are toy examples, but reflecting, nonetheless, aspects of real-life social networks.

In the future, we are planning to extend this work and try to cope with the problems in Definitions 20 and 21, e.g. by finding a trade-off between the risk and the gain functions through *multiobjctive optimization* [17, 18]. Clearly, one of the main problems could be the submodular nature of our objective functions. Orthogonally, we would like to set up a few experiments on real-life data, in order to empirically validate our results.

References

1. Backes, M., Gomez-Rodriguez, M., Manoharan, P., Surma, B.: Reconciling privacy and utility in continuous-time diffusion networks. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF). pp. 292–304 (2017)
2. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web. pp. 665–674. ACM (2011)
3. Carnes, T., Nagarajan, C., Wild, S.M., van Zuylen, A.: Maximizing influence in a competitive social network: A follower's perspective. In: Proceedings of the Ninth International Conference on Electronic Commerce. pp. 351–360 (2007)
4. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1029–1038. ACM (2010)
5. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 199–208. ACM (2009)
6. De Choudhury, M., Mason, W., M. Hofman, J., Watts, D.: Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10. pp. 301–310 (2010)
7. Du, N., Song, L., Woo, H., Zha, H.: Uncover topic-sensitive information diffusion networks. In: AISTATS (2013)
8. Fujishige, S.: Submodular Functions and Optimization. Annals of Discrete Mathematics, Elsevier Science (2005)
9. Gomez Rodriguez, M., Schölkopf, B.: Influence maximization in continuous time diffusion networks. In: Proceedings of the 29th International Conference on Machine Learning. pp. 313–320. Omnipress (2012)
10. Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: ICML (2011)
11. Gomez-Rodriguez, M., Song, L., Du, N., Zha, H., Schölkopf, B.: Influence estimation and maximization in continuous-time diffusion networks. ACM Trans. Inf. Syst. **34**(2), 9:1–9:33 (2016)
12. Iyer, R., Bilmes, J.: Submodular optimization with submodular cover and submodular knapsack constraints. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 2436–2444. Curran Associates Inc. (2013)
13. Iyer, R., Jegelka, S., Bilmes, J.: Fast semidifferential-based submodular function optimization. 30th International Conference on Machine Learning, ICML 2013 (2013)

14. Kasprzak, R.: Diffusion in networks. Journal of Telecommunications and Information Technology **nr 2**, 99–106 (2012)
15. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 137–146. ACM (2003)
16. Lappas, T., Terzi, E., Gunopulos, D., Mannila, H.: Finding effectors in social networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1059–1068. ACM (2010)
17. Papadimitriou, C.H., Yannakakis, M.: On the approximability of trade-offs and optimal access of web sources. In: 41st Annual Symposium on Foundations of Computer Science,. pp. 86–92. IEEE (2000)
18. Papadimitriou, C.H., Yannakakis, M.: Multiobjective query optimization. In: Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,. ACM (2001)
19. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 61–70. ACM (2002)
20. Tzoumas, V., Amanatidis, C., Markakis, E.: A game-theoretic analysis of a competitive diffusion process over social networks. In: Proceedings of the 8th International Conference on Internet and Network Economics. pp. 1–14. WINE'12, Springer-Verlag, Berlin, Heidelberg (2012)
21. Watts, D., Dodds, P.: Influentials, networks, and public opinion formation. Journal of Consumer Research **34**, 441–458 (2007)
22. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (1998)
23. Yu, M., Gupta, V., Kolar, M.: An influence-receptivity model for topic based information cascades. In: International Conference on Data Mining (ICDM). pp. 1141–1146. IEEE (2017)
24. Zeng, Y., Chen, X., Cong, G., Qin, S., Tang, J., Xiang, Y.: Maximizing influence under influence loss constraint in social networks. Expert Syst. Appl. **55**(C), 255–267 (2016)
25. Zhou, D., Wenbao, H., Wang, Y.: Identifying topic-sensitive influential spreaders in social networks. International Journal of Hybrid Information Technology **8**, 409–422 (2015)
26. Zhou, J., Zhang, Y., Cheng, J.: Preference-based mining of top-k influential nodes in social networks. Future Generation Computer Systems **31**, 40 – 47 (2014)