This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3293233

Manipulating Opinions in Social Networks with Community Structure

Paolo Bolzern, *Member, IEEE*, Alessandro Colombo, *Senior Member, IEEE*, and Carlo Piccardi

***Abstract*—Online social media have been one of the greatest drivers of societal change of the past two decades, but are now being recognized as one of the major causes of opinion radicalization and one of the most effective tools for opinion manipulation. Starting from a class of stochastic models of opinion dynamics, and considering different structures of social networks with increasingly realistic features (including a snapshot of the Facebook friendship network), we develop a mathematical model of different forms of opinion manipulation. We then explore how network properties, and in particular degree distribution and community structure, interact with the attack to amplify or reduce its effect on the population, both globally and on specific subsets. We find, in particular, that degree heterogeneity is key to making online social media susceptible to very effective attacks, even with relatively little effort. Communities instead play a more complex role, acting both as barriers to the spread of manipulated opinions through the whole population and as amplifiers of manipulated opinions when the target of the attack is a community of the online social medium. The results of our study can help design effective strategies to prevent the manipulation of opinions through online social media.**

***Index Terms*—Opinion dynamics, opinion manipulation, social network, community structure**

I. Introduction

ONLINE social media have drastically changed the way information is disseminated and public opinion is formed in modern societies. By posting messages, every user of an online social medium can reach a very large audience with just a few clicks, potentially influencing the opinions of an immense crowd on some specific topic. As a result, malicious actors have powerful tools to target the most susceptible people and manipulate collective opinion with the intent of pursuing social, economic, or political interests.

For example, since the early 2010s, online companies have been reported to employ remunerated people for viral marketing, forming the so-called "Internet Water Army" [1] to steer consumers' choices towards specific commercial goods. During election campaigns, party activists or social bots (algorithmic-based artificial users that impersonate humans in discussions on social media platforms [2]) can exert a great influence on public opinion in online forums [3], [4]. In the 2016 US presidential election, the attack by hackers and trolls on digital media was suspected to have significantly distorted the results [5], [6] – the same problem affected the 2020 elections [7] – to the point that protecting elections from social media manipulation is becoming a serious matter of concern in several countries [8].

The authors are with Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Italy (email: alessandro.colombo@polimi.it)

Mathematical models have been widely used to obtain a quantitative assessment of the formation and evolution of opinions on social networks. These models can be grouped into two main categories: deterministic and stochastic. In deterministic models, each individual opinion on a given topic is represented by a real-valued scalar, which evolves over time depending on the opinions of its neighbors. Stemming from the pioneering work of DeGroot [9], a rich literature has flourished, revived in recent years by the advent and spread of digital social platforms. For a survey on the deterministic framework, the reader is referred to [10], [11].

To capture the complexity of the interdependence of nodes within real-world social communities, however, the adoption of a stochastic paradigm is particularly promising. Since it is impossible to account for all the microscopic sources of influence affecting the opinion of individuals, it seems reasonable to describe their opinion changes as random events governed by a probabilistic law. A classical stochastic model used to describe the opinion dynamics in a network is the so-called Voter Model, originally proposed in [12]. Each individual, at any time instant, has one of two discrete opinions and is linked to the other nodes by a weighted connection graph. The probability of flipping to the opposite opinion depends on the opinions of the neighbors, according to a simple stochastic rule. The Sznajd model [13] for opinion dynamics derives from the Ising model used in statistical mechanics to describe ferromagnetic phenomena. According to the model, the state of each agent is a binary value and the agents are interconnected by a lattice. At each step, two neighboring agents are randomly selected and, based on their agreement or disagreement, their state and the state of their neighbors is updated according to a simple logical rule. Generalizations to more complex graphs were also developed. Another important stochastic model, grounded on a sociophysics perspective, is the Galam model [14]. In its original formulation, the opinion is binary, the agents are randomly divided in groups of a given size, and the agents in each group update their opinion adopting the majority opinion in the group. Then the agents are reshuffled and the procedure is repeated until steady-state is attained. This model does not take into account the connection network of the agents.

The model considered in this paper can be seen as a refined version of the Voter Model. It was originally proposed in [15], and further developed and extended in [16]–[18]. It describes a social network as a multi-agent system, where each agent represents an individual, having a categorical opinion that varies with time within a finite set of possible options. Each agent's opinion is modeled as a stochastic Markov process which evolves in time depending on the opinions of the neighboring

agents. Precisely, the probability of moving to a different opinion is positively affected by the fraction of neighbors that share that opinion. The model parameters include the transition probabilities of each isolated agent (determining its prejudice and the volatility of its opinion), the graph topology associated with the network, the individual influenceability, and a global influence intensity parameter that measures the strength of the interaction, as modulated by the content-filtering algorithm of the online social medium. In [16], this model has been used to describe both the transient evolution and the steady-state value of the average opinion when the agents are homogeneous. The effect of a centralized tuning on emerging collective behaviors of heterogeneous agents was studied in [17], while [18] dealt with the computation of second-order moments (correlation and variance), with application to evaluating expectation and variance of the vote share in a political competition between two parties with different degree of stubbornness. In all these papers, a key role is played by the *social power* of a single agent, a centrality index that measures the weight of each agent's prejudice in determining the steady-state network average opinion. Such an index can be analytically computed from the model parameters, without the need for extensive simulation. Compared to the Voter Model, this multi-agent model is richer and more realistic, as it allows to deal with many important sociological aspects (including individual prejudice and influenceability, spontaneous change of opinion, strength of interaction) that are overlooked in other stochastic models. At the same time, it is still amenable to analytical treatment and, through the computation of the social power, provides a way to detect the agents in the network with more influence capability. It is also worth noting that the model naturally extends to more than two alternative opinions [19].

Given the documented risk of opinion manipulation through online social media, several scholars have analyzed the many facets of the problem from a mathematical modeling standpoint. A typical manipulation strategy consists of targeting some agents in the network in order to make them stubborn agents, biased in favor of one of the opinions. For example, in [20] the authors focus on a generalized voter model by introducing agents that have a fixed state and can influence others' opinions with unidirectional influence links. In particular, they study the optimal placement of such agents to have the maximum impact on the global network opinion. A similar problem is discussed in [21], highlighting the role that the network structure plays in determining the ease with which bias can be manipulated. Typically, optimal manipulation strategies tend to target nodes with high-degree. This may not be the case in different scenarios, like the one considered in [22], where two external controllers compete to steer the average network opinion in opposite directions. In that case, it is shown that the optimal strategies are strongly dependent on the opponent actions, the limitations in the manipulation effort and the level of stubbornness of attacked agents. In [23] the effect of an adversarial attack conducted to distort the voter model dynamics is considered. Interestingly, it is shown that even extremely small (and hardly detectable) perturbations in the edge weights can significant alter the vote dynamics. Manipulation strategies for majority-based models are analyzed in [24], [25]. For a summary of different opinion manipulation approaches in classical opinion dynamics models (both deterministic and stochastic) the reader is referred to [26].

An important feature of networks formed through online social media is their *community structure* [27]–[29]: when spreading or receiving information, users tend to mostly interact with a rather restricted number of users with whom they share interests, or geographical collocation, or political orientation, just to mention a few possibilities. In network terms, a community is defined as a set of nodes whose internal connectivity (i.e., edge density) is much higher than the connectivity to the rest of the network. The rigorous definitions of community, and algorithms for identifying communities in networks, are the subject of a very broad literature (see [30], [31] for surveys). Very important is the notion of *modularity* [32], a scalar indicator that quantifies to what extent the network has a strong community structure or, on the contrary, connections tend to be randomly distributed over the whole network.

It is not surprising that, when studying information diffusion and opinion dynamics on networks, the community structure plays an important role. Intuitively, the high internal density of connections favors an effective diffusion and a strong influenceability within communities, whose borders however tend to block or delay further spreading. The literature to date reports diversified results, according to the diffusion model and network properties used in the experiments (see e.g., [33]–[39]).

In this paper, we analyze the dynamics of opinions in a social network manipulated by the tampering of an online social medium characterized by a community structure. The objective of our analysis is to identify the sources of vulnerability to opinion manipulation in networks with a complex structure, with the aim of contributing to improving the resilience of future online social media (in this regard, we depart from the literature which aims to optimize the effects of the attack, e.g., [40], [41]). The set of opinions is binary (say $\{1, 2\}$) and, in the unperturbed situation (no manipulation), the social network reaches a steady-state with a certain expected number of agents in opinion 1 at each time instant. The first problem that we address is to assess the worst-case effect that an attack to a few nodes can have on the whole network. Two alternative strategies of attack will be defined: the first one shifts the average stand-alone opinion of some of the agents (*soft attack*), and the other substitutes a few individuals with stubborn agents (bots) diffusing the desired opinion (*hard attack*). To evaluate the worst-case scenario, we maximize the effect of these attacks over all possible choices of the attacked nodes, using heuristics when an exact solution is not practical to compute. The multiplicative network effect on spreading will be confirmed – few attacked agents are sufficient to yield large global variations – and, most notably, we will discuss how communities are hit inhomogeneously by the attack. Then, our second problem will be to assess how the opinion of a given subset of the social network can be influenced. In this case, we will discover that the attack strategies are much more effective when such a subset corresponds to a community in

the online social medium, proving that targeting a cohesive pool of individuals (e.g., a group that, by virtue of sharing the same interest or political orientation, forms a tight community in the online social medium) is facilitated by the strong ties among its members.

Overall, this paper contributes to the literature that studies, through mathematical modeling, the dynamics of opinion manipulation through online social media, to improve the understanding of basic mechanisms and global effects. Our hope is that, with better knowledge of the factors that make online social media more or less susceptible to manipulation, these will evolve into more robust and resilient platforms for opinion sharing.

II. A MODEL OF OPINION DYNAMICS

In this paper, we consider a stochastic model for opinion dynamics based on a network of interacting Markovian agents, first proposed in [15] and then further developed in [16]–[18]. Figure 1 reports a graphical representation of the main features of the model, both for an isolated and for a networked agent, as well as the salient features of a manipulating attack: all these notions will be introduced in this and the following sections.

We consider a social network, i.e., a group of individuals (or agents), which form their opinion based on their individual attitude towards a topic, on interactions through standard communication channels (e.g., radio, television, press, in-person social interactions), as well as through an online social medium, and we separate the effects of the online social medium from all the other effects. In its simplest formulation, the model assumes that the opinion of each agent on a given issue belongs to the binary set $\{1, 2\}$, and evolves in time according to an irreducible continuous-time Markov chain model [42].

In the absence of the online social medium, each agent r has its own transition probability rates $q_{12}^{[r]}$ and $q_{21}^{[r]}$ of passing from opinion 1 to opinion 2 and *vice versa*. Correspondingly, the steady-state probability of agent r being in opinion 1 is given by $\beta_r = q_{21}^{[r]}/(q_{12}^{[r]} + q_{21}^{[r]})$. This can be interpreted as the *stand-alone prejudice* of agent r. This prejudice corresponds to the opinion bias determined by all factors except the online social medium.

The structure of the interactions through the social medium is then described by a weighted graph $\mathcal{G}$, with N nodes corresponding to the agents and edges representing the mutual influence between agents through the online social medium. The weight $w_{rs} > 0, r \neq s$, attached to the edge directed from node s to node r measures the trust of agent r in agent s. If agents r and s do not influence each other, there is no edge linking nodes r and s and we put $w_{rs} = w_{sr} = 0$. Also, there is no self-influence, i.e., $w_{rr} = 0\ \forall r$. Notice that under these assumptions the graph $\mathcal{G}$ is directed, because $w_{rs} \neq w_{sr}$ in general, but fully reciprocated, i.e., $w_{rs} = 0$ if and only if $w_{sr} = 0$: the topological structure of $\mathcal{G}$ is thus undirected. The weights are collected in the $N \times N$ *trustiness matrix* $W = [w_{rs}]$. It is assumed, without loss of generality, that W is row-normalized, i.e., $W\mathbf{1}_N = \mathbf{1}_N$, where $\mathbf{1}_N$ is the all-one N-dimensional column vector. The weighted Laplacian of the graph is defined as $L = I_N - W$, where I_N is the $N \times N$ identity matrix.

Due to the influence through the online social medium, the transition rates of each individual are affected by an additive term depending on the current opinions of neighbors. More precisely, the transition probability rate $q_{ij}^{[r]}, i \neq j$, is increased to $q_{ij}^{[r]} + \eta_r \sum_{s \in \mathcal{N}_r} w_{rs} I_{s,j}(t)$ where $\mathcal{N}_r$ is the set of neighbors of agent r and the stochastic variable $I_{s,j}(t)$ has value 1 if agent s has opinion j at time t and 0 otherwise. The parameter $\eta_r \geq 0$ is specific to each agent and denotes her/his susceptibility to social influence. If $\eta_r = 0$, agent r is stubborn, i.e., cannot be influenced by others' opinions. Conversely, large values of η_r denote a high tendency to conform with neighboring opinions. The diagonal matrix H with entries $\eta_r, r = 1, 2, \ldots, N$, on the diagonal is called the *influenceability matrix*.

It was shown in [17] that the state of the overall system (obtained as the collection of all agents' states) behaves like a high-dimensional irreducible continuous-time Markov chain, which is ergodic and thus asymptotically converges to a unique steady-state for all initial conditions. While this property was stated in [17] under the assumption that $\mathcal{G}$ is a strongly connected graph, it can be easily shown that it holds also without such a restrictive assumption. The overall multi-agent model is denoted as the Master Markov model. The dimension of its state space is 2^N and makes it difficult to assess its theoretical properties when the number N of agents is large. However, it was proven in [17] that the Master Markov model can be dramatically simplified via marginalization.

Precisely, letting $z_r(t)$ be the probability that agent r has opinion 1 at time t and defining the vector $z(t) = \begin{bmatrix} z_1(t) & z_2(t) & \cdots & z_N(t) \end{bmatrix}^\top$, it can be proven (see [17]) that $z(t)$ converges asymptotically to the steady-state vector

$$\bar{z} = (I_N + F^{-1}HL)^{-1}\beta = M\beta, \tag{1}$$

where $\beta = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_N \end{bmatrix}^\top$ is the vector of stand-alone prejudices in the absence of the online social medium influence (i.e., when $H = 0$), F is a diagonal matrix with entries $\alpha_r = q_{12}^{[r]} + q_{21}^{[r]}, r = 1, 2, \ldots, N$, on the diagonal, and $M = (I_N + F^{-1}HL)^{-1}$. As discussed in [17], the parameter α_r is a positive time-scale parameter, associated with the rate of opinion change of agent r. Precisely, a high value of α_r corresponds to high opinion volatility. Note that the matrix $F^{-1}H$ is a diagonal matrix with nonnegative diagonal entries $\frac{\eta_r}{\alpha_r}$. As a consequence, a low influenceability or a high volatility yield the same effect.

Matrix M above has two interesting properties.

Lemma 1: Matrix $M = (I_N + F^{-1}HL)^{-1}$ is nonnegative, irrespective of the values of F, H, and L.

Proof: Matrices F and H are, by construction, diagonal with nonnegative diagonal elements, while matrix $-L$ is zero row-sum and Metzler (it has nonnegative off-diagonal elements). This ensures that the real parts of all the eigenvalues of $F^{-1}HL$ are nonnegative so that the real parts of all the eigenvalues of $-M^{-1} = -(I_N + F^{-1}HL)$ are strictly negative. Furthermore, since $-L$ is Metzler, matrix $-M^{-1}$ is itself a Metzler matrix. As shown in [43], the fact that $-M^{-1}$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3293233

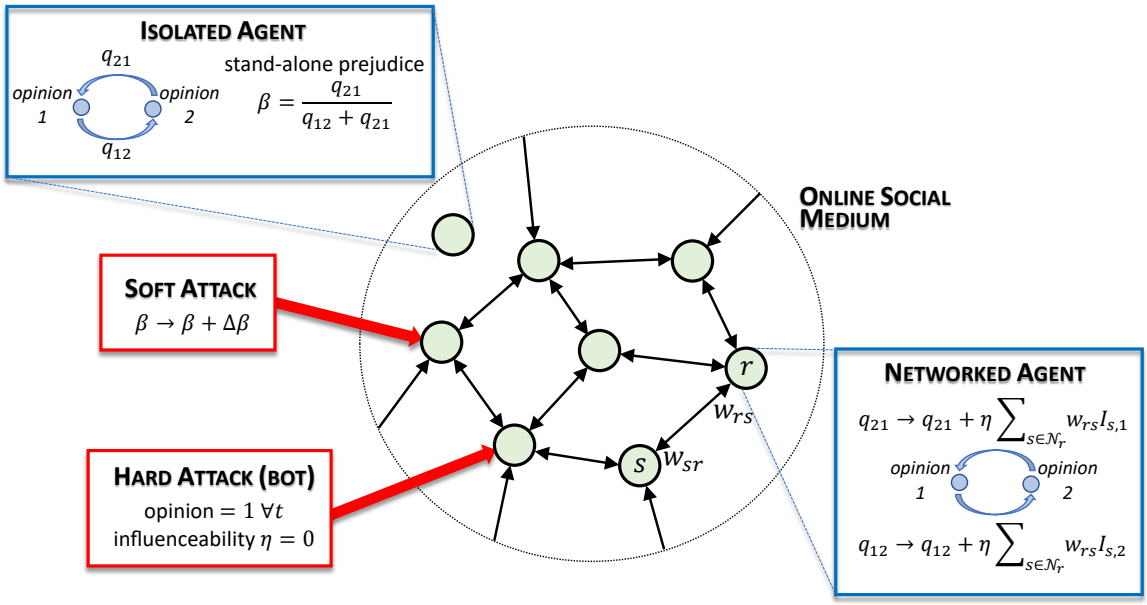


Fig. 1. A summary diagram of the opinion model and of the attack strategies. An isolated agent spontaneously switches between opinions 1 and 2, due to external factors, with transition rates q_{12}, q_{21}, having probability β of being in opinion 1 at steady-state (*stand-alone prejudice*). When the agent's opinion is influenced through the online social medium, the transition rates increase according to the number of neighbors having a given opinion ($I_{s,1}$, $I_{s,2}$), to the trust w_{rs} given to the neighbor, and to the agent's influenceability η. A *soft attack* to a given node increases the value of its stand-alone prejudice β by altering the rates q_{12}, q_{21}. A *hard attack* replaces the node with a "bot", which has opinion 1 for all t and is not influenced by neighboring agents.

is Metzler with strictly negative eigenvalues implies that M is a nonnegative matrix. ■

Lemma 2: Each row of M sums to 1, irrespective of the values of F, H, and L.

Proof: By construction, W is row-normalized and $L = I_N - W$, therefore all rows of L sum to zero. Given that both F and H are diagonal matrices, all rows of $F^{-1}HL$ must also have zero sum, and therefore all rows of $M^{-1} = I_N + F^{-1}HL$ sum to 1. This implies that $M^{-1}\mathbf{1}_N = \mathbf{1}_N$, and therefore $M\mathbf{1}_N = \mathbf{1}_N$. ■

Finally, notice that the expected fraction of agents sharing opinion 1 at steady-state (hereafter denoted as E_1 and dubbed as the *expected opinion share*) is given by

$$E_1 = \frac{1}{N}\sum_{r=1}^{N} \bar{z}_r = \frac{1}{N}\mathbf{1}_N^\top M\beta = \frac{1}{N}\|M\beta\|_1. \qquad (2)$$

The above equation states, quite simply, that the expected opinion share is the average value of the elements of $\bar{z}$ in (1). It is worthwhile observing that the r-th element of the row vector $\frac{1}{N}\mathbf{1}_N^\top M$ measures the *social power* of agent r, since it represents the weight of the agent r's prejudice in determining the expected opinion share E_1. In the context of opinion dynamics, social power is a well-established concept, dating back to the seminal work [44] on the formal theory of interpersonal relations. Roughly speaking, individual social power corresponds to the amount of influence an individual has on the overall discussion in a group.

Remark 1: The theoretical results recalled above were derived under the assumption that the Markov chain model of each agent is irreducible. In that case, the prejudice β_r takes value in the open interval $(0, 1)$. However, when discussing some attack strategies in Section IV, we will need to relax such an assumption, considering also the extreme cases when β_r, for some r, is either equal to 0 or 1. Note that eqs. (1) and (2) are still valid also in those cases. As a matter of fact, although the Master Markov model may be reducible, its graph has a unique absorbing class, since nodes with β_r equal to 0 or 1 have a deterministic asymptotic state. This ensures the uniqueness of the steady state vector $\bar{z}$.

III. Models of Network Topologies

As we will see in the next sections, the network topology, i.e., the structure of the edges that is encoded in the nonzero off-diagonal elements of matrix L, has significant effects on the response of the system to a change in some of its nodes dynamics. In this paper, we consider three classes of artificial networks with increasing complexity, and a real-world network defined by Facebook friendships (see Sec. V for details). In this section, we succinctly introduce the models used to generate the artificial networks. We refer the reader to standard textbooks (e.g., [45]) for a detailed discussion. All models yield an undirected network whose structure is coded in an *adjacency matrix* $A = [a_{rs}]$, where $a_{rs} = a_{sr} = 1$ if an edge connects nodes (r, s), and $a_{rs} = a_{sr} = 0$ otherwise (also, $a_{rr} = 0\ \forall r$). The *degree* $d_r = \sum_{s=1}^{N} a_{rs} = |\mathcal{N}_r|$ is the number of agents directly connected to r.

The network is then equipped with a trustiness matrix $W = [w_{rs}]$ which is, in general, non-symmetric as discussed in Sec. II, and which assigns a weight to each edge (notice that $w_{rs} = 0$ if and only if $a_{rs} = 0$). More specifically, we investigate three scenarios for W:

(1) *Uniform trustiness*: node r attributes the same trustiness to all its neighbors, i.e., $w_{rs} = \frac{1}{d_r}\ \forall s \in \mathcal{N}_r$.

(2) *Random trustiness*: node r attributes random trustiness to its neighbors, i.e., w_{rs} is extracted uniformly at random in $[0, 1]$, and then normalized such that $\sum_{s\in\mathcal{N}_r} w_{rs} = 1$.
(3) *Degree-dependent trustiness*: node r attributes larger trustiness to neighbors with larger degree, i.e., $w_{rs} = \frac{d_s}{n_r}$, where $n_r = \sum_{s\in\mathcal{N}_r} d_s$.

Notice that all three scenarios yield a row-normalized matrix W; the condition is explicitly enforced in the second case, while it is a consequence of the definition of w_{rs} in the other two cases.

A. Small-World (SW) network

An SW network [46] is built by initially connecting the N nodes in a circular lattice, where each node is linked to m neighbors on the right and m on the left (thus each node has initial degree $d_r = 2m$). Then a rewiring procedure is carried out: all nodes are sequentially scanned and, for each node r, all of its right-side edges are considered. With probability p, an edge is detached from the neighbor of r and connected to a node selected uniformly at random. The resulting network has an average node-to-node distance much smaller than the original one, thanks to the long-distance connections originated by rewiring ("small-world effect"), while the clustering coefficient C (the average probability that two neighbors of a node are neighbors themselves, e.g., [45]) is rather large: the coexistence of these two features is typical in social networks. On the other hand, the network is "single-scale", i.e., the degree of nodes has only small fluctuations around the average $d_{avg} = 2m$.

B. Barabási-Albert (BA) network

A BA network [47] is based on the principle of "preferential attachment": starting from a small number N_0 of arbitrarily connected nodes, a new node with m edges is added at each step $N_0 + 1, N_0 + 2, \ldots, N$. Each one of the m edges is connected to an existing node r, selected with probability proportional to its current degree d_r. For large N, the resulting network has a few nodes with disproportionately many connections, coexisting with a majority of medium/small degree nodes, giving rise to a strongly inhomogeneous ("scale-free") network with power-law degree distribution

$$Prob(d_r = d) \propto d^{-\gamma},$$

with $\gamma = 3$ and average degree $d_{avg} = 2m$. Such a functional form for the degree distribution is encountered in a large number of real-world datasets. On the other hand, BA networks have a vanishing clustering coefficient for large N, contrary to most real-world networks.

C. Lancichinetti-Fortunato-Radicchi (LFR) network

An LFR network [48] adds a further level of complexity, namely *community* or *modular structure* [30]: the network can naturally be partitioned in subgraphs, called *modules* or *communities*, with large internal edge density but loose connections to the other modules. It is a structure often encountered in real-world networks, including social networks, where a community is formed by a set of individuals who address most of their interactions with other individuals of their same community (e.g., [27], [28]).

The (maximum) *modularity* Q is the most used indicator to quantify to what extent a network is actually structured into communities [30], [32]. Given a partition $\mathcal{P} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_q\}$ of the nodes, the associated modularity $Q_\mathcal{P}$ is defined by

$$Q_\mathcal{P} = \frac{1}{2L} \sum_{\mathcal{C}_i} \sum_{(r,s)\in\mathcal{C}_i} \left(a_{rs} - \frac{d_r d_s}{2L} \right),$$

which is the (normalized) unbalance between the actual number of edges internal to communities, and the expected value of such a quantity if the edges are randomized by preserving the node degrees d_r (it can be proved that $\frac{d_r d_s}{2L}$ is indeed the expected value of a_{rs} in the randomized network [32]). Thus $Q_\mathcal{P}$ is large ($\to 1$ due to normalization) when, for the partition $\mathcal{P}$, the edges internal to communities are many more than what is expected by chance. Then the modularity $Q = \max_\mathcal{P} Q_\mathcal{P}$, obtained by maximizing over all possible partitions, is used to measure overall to what extent the network is structured in communities.

LFR networks are explicitly designed to have a built-in community structure, with tunable modularity. In LFR networks, not only the degree of nodes is power-law – and thus inhomogeneously distributed – as in BA networks, but also the size of communities is such, to mimic features typically found in real-world data.

To generate an LFR network, one has to set – besides N and d_{avg} – the values of γ and γ_c, i.e., respectively, the exponents of the power-law degree distribution and of the community-size distribution, and a "mixing parameter" μ prescribing the fraction of edges each node directs outside its community, so that the smaller μ, the stronger is the community structure.

IV. Models of attack strategies

Equation (2) establishes an algebraic relation between the system parameters (agents and network) and the steady state expected opinion share E_1. Let us now consider the objective of pushing such a quantity towards its extreme value of 1, i.e., of maximizing the expected number of agents sharing opinion 1, by acting on a small number of agents. This objective can be formulated as a maximization problem, with (2) as the benefit function. We can formulate two alternative strategies: we may attempt to influence E_1 by shifting the stand-alone prejudice of some of the agents using means external to the online social medium, for example by feeding them with manipulative advertising [49], or by offering monetary [1] or other forms of incentives (what we call a *soft attack*), and/or we may do so by substituting some of the agents in the online social medium with bots [2], [3] in charge of diffusing the chosen opinion in the network (*hard attack*). Let us see how these two kinds of attack translate in the mathematical formalism introduced above.

A. Soft attack

We begin by rewriting the vector β of stand-alone prejudices as $\beta = \beta^{\text{a.p.}} + \Delta$, where $\beta^{\text{a.p.}}$ is the vector of *a priori* prejudices,

before the attack, and the nonnegative vector Δ encodes the effect of the attack on the prejudices[1]. From (2) we obtain

$$E_1 = \frac{1}{N}\|M(\beta^{\text{a.p.}} + \Delta)\|_1 = \frac{1}{N}\|M\beta^{\text{a.p.}}\|_1 + \frac{1}{N}\|M\Delta\|_1.$$

The linearity of the 1-norm over nonnegative real vectors allows us to decouple the effects on E_1 of the *a-priori* prejudice $\beta^{\text{a.p.}}$ and of the variation of the prejudices Δ. A soft attack only affects the variation of the prejudices, therefore it only affects the last term in the above formula. Let us assume to attack k agents in the network so as to change their *a posteriori* prejudice from $\beta_r^{\text{a.p.}}$ to 1, which means that exactly k elements of Δ are nonzero. Denoting by

$$D = \text{diag}(\mathbf{1}_N - \beta^{\text{a.p.}}),$$

the diagonal matrix with entries $1 - \beta_r^{\text{a.p.}}$ on the diagonal, we can rewrite

$$\frac{1}{N}\|M\Delta\|_1 = \frac{1}{N}\|MDv\|_1, \quad v \in \{0,1\}^N,$$

where v is a binary vector with unit elements corresponding to the attacked agents. The optimal soft attack is then obtained by solving the following binary optimization problem.

Problem 1 (Soft attack with known prejudices):

$$\begin{aligned} &\max_{v\in\{0,1\}^N} \frac{1}{N}\|MDv\|_1, \\ &s.t. \quad \|v\|_1 = k, \end{aligned}$$

The solution simply amounts to selecting the k elements of vector $\mathbf{1}_N^\top MD$ of maximal value. Notice that vector $\frac{1}{N}\mathbf{1}_N^\top M$ was defined at the end of Sec. II to be the vector of the agents' social power, so that we can see $\frac{1}{N}\mathbf{1}_N^\top MD$ as the social power vector, weighted by the diagonal elements of D. The optimal solution to Problem 1 is thus given by attacking the k agents with maximum *weighted social power*, the weight being the gap between the agent's *a priori* prejudice and 1.

The above general problem slightly changes if, more realistically, we assume that the attacker has no knowledge of the *a priori* prejudices $\beta^{\text{a.p.}}$ and cannot force the *a posteriori* prejudices to 1. We may assume, rather reasonably, that $\beta^{\text{a.p.}}$ are uniformly distributed between 0 and 1 and the *a posteriori* prejudices are uniformly distributed between $\beta^{\text{a.p.}}$ and 1. In this case, the optimization of the soft attack is formulated as a maximization of the expected value

$$\mathbb{E}[E_1] = \mathbb{E}\left[\frac{1}{N}\|M\beta^{\text{a.p.}}\|_1 + \frac{1}{N}\|MDv\|_1\right],$$

where now both $\beta^{\text{a.p.}}$ and D are stochastic variables, with $\beta^{\text{a.p.}}$ uniformly distributed between $\mathbf{0}_N$ and $\mathbf{1}_N$ and the diagonal of D uniformly distributed between $\mathbf{0}_N$ and $\mathbf{1}_N - \beta^{\text{a.p.}}$. The quantity in the above expression is linear in the stochastic variables[2], so we can evaluate the expected value as

$$\mathbb{E}[E_1] = \frac{1}{2N}\|M\mathbf{1}_N\|_1 + \frac{1}{4N}\|Mv\|_1.$$

[1]The value of $\beta_r = q_{21}^{[r]}/(q_{12}^{[r]} + q_{21}^{[r]})$ is increased by either increasing $q_{21}^{[r]}$ or decreasing $q_{12}^{[r]}$.

[2]Since $\|x\|_1$ is linear in x if x is a nonnegative vector, then $\mathbb{E}[\|x\|_1] = \|\mathbb{E}[x]\|_1$ when x is a vector of nonnegative stochastic variables.

The optimal soft attack is then a solution of the following binary optimization problem.

Problem 2 (Soft attack with unknown prejudices):

$$\begin{aligned} &\max_{v\in\{0,1\}^N} \frac{1}{4N}\|Mv\|_1 \\ &s.t. \quad \|v\|_1 = k. \end{aligned} \tag{3}$$

Notice again that $\|Mv\|_1 = \mathbf{1}_N^\top Mv$, and the elements of vector $\frac{1}{N}\mathbf{1}_N^\top M$ correspond to the social power of the N agents. Therefore the above formula states that, lacking precise knowledge about the *a priori* and *a posteriori* prejudices, the soft attack that maximizes the expected number of agents sharing opinion 1 corresponds to attacking the k agents with maximum social power.

B. Hard attack

Let us now consider the modeling of a hard attack. In this case, not only the *a priori* prejudices of k agents are affected (and set strictly equal to 1 if we model bots as agents with an unchangeable opinion), but their influenceability η_r is set to 0. This means that matrix H in (1) changes with the attack, and so does M. Assuming perfect knowledge of vector $\beta^{\text{a.p.}}$, the optimal attack problem can now be formulated as follows.

Problem 3 (Hard attack with known prejudices):

$$\begin{aligned} &\max_{v\in\{0,1\}^N} \frac{1}{N}\|M_v\beta^{\text{a.p.}}\|_1 + \frac{1}{N}\|M_vDv\|_1, \\ &s.t. \quad \|v\|_1 = k, \end{aligned}$$

where M_v is matrix $M = (I_N + F^{-1}HL)^{-1}$ after setting $H_{rr} = 0$ (i.e., $\eta_r = 0$) if $v_r = 1$.

The above problem is no longer linear in the decision variable v, which now appears as an argument of matrix M_v. The exact computation of the optimal hard attack is therefore significantly more complex than the computation of the optimal soft attack.

If, as we did for the soft attack, we assume that $\beta^{\text{a.p.}}$ is unknown but uniformly distributed between $\mathbf{0}_N$ and $\mathbf{1}_N$, while the *a posteriori* prejudice of the attacked agent is set to 1 by construction, following similar reasoning as before we find that the hard attack that maximizes the expected number of agents sharing opinion 1 is a solution to the problem

$$\begin{aligned} &\max_{v\in\{0,1\}^N} \frac{1}{2N}\|M_v\mathbf{1}_N\|_1 + \frac{1}{2N}\|M_vv\|_1, \\ &s.t. \quad \|v\|_1 = k. \end{aligned} \tag{4}$$

Notice the factor $1/2$ in the second term of the benefit function, in contrast to the factor $1/4$ that we had in the stochastic soft attack, due to the fact that the hard attack guarantees *a posteriori* prejudice equal to 1 for the attacked agents.

Using Lemma 2 we have that the first term $\frac{1}{2N}\|M_v\mathbf{1}_N\|_1 = \frac{1}{2}$, so that (4) further simplifies as follows.

Problem 4 (Hard attack with unknown prejudices):

$$\begin{aligned} &\max_{v\in\{0,1\}^N} \frac{1}{2N}\|M_vv\|_1, \\ &s.t. \quad \|v\|_1 = k. \end{aligned} \tag{5}$$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3293233

Notice that the quantity $\delta E_1 = \frac{1}{2N}\|M_v v\|_1$, i.e., the benefit function of Problem 4, is the expected increase – caused by manipulation – in the fraction of agents sharing opinion 1 at steady-state. The optimal value of this quantity will be denoted by δE_1^* in the remainder of the paper.

C. Attack heuristics

Problems 1 and 3 in the previous section describe the optimization of the soft and hard attacks under the assumption of perfect knowledge of the network agents' prejudices. While in principle prejudices can be estimated, for instance through the analysis of the history of the opinions expressed by each agent on the online social medium, they are unlikely to be known exactly. Problems 2 and 4 therefore appear to be more realistic models of an attack scenario. For this reason, in the following, we focus on these two problems.

Problem 2 is solved exactly by computing the social power of each agent. This computation has complexity $O(N^3)$ due to the matrix inversion which is necessary to compute M. Problem 4, on the other hand, is nonlinear, and a computationally efficient exact solution is unknown. To tackle networks of tens of thousands of nodes, we need a scalable heuristic. A possibility is to follow the strategy which is exact for the soft attack, that is to attack the k nodes with highest social power. We call this the *Social Power Heuristic*. This heuristic stands on the assumptions that M_v is only slightly different than M, and can therefore be approximated by M, and allows to find a feasible solution to Problem 4 with complexity $O(N^3)$.

An alternative heuristic is to identify all the possible sets of k nodes with top-ranking degrees, and then perform a complete enumeration only among these sets. We call this the *Degree Heuristic*. This heuristic hinges on the assumption that the most effective targets are the agents that have the largest number of neighbors, and in highly heterogeneous networks the set of agents of top-ranking degree is much smaller than the total number of agents. Computation of the agents' degree given L has complexity $O(N^2)$, but in most networks, the Degree Heuristic requires the complete enumeration step, since agents' degree is an integer value and there are typically multiple sets of k top-ranking agents with identical degree distribution.

The two above-defined heuristics can be compared with the exact solution (i.e. the exhaustive assessment of all possible k-tuple of agents) only on very small systems. We report in Fig. 2 the comparative performance of the exact solution and the two heuristics in solving Problem 4, computed on 100 randomly generated SW networks of 40 nodes. Each of the 100 networks was weighted according to each of the 3 trustiness scenarios detailed in the previous section. We see how, irrespective of the trustiness scenario, the heuristics come very close to the optimal value, with a slight advantage for the Degree Heuristic. We cannot meaningfully test the heuristics against the exact solution on BA and LFR networks sized as in Fig. 2, because the small size prevents the generation of a reasonable node degree and community size distributions, which are the characterizing factors of BA and LFR networks. However, we report in Fig. 3 the results of applying the two

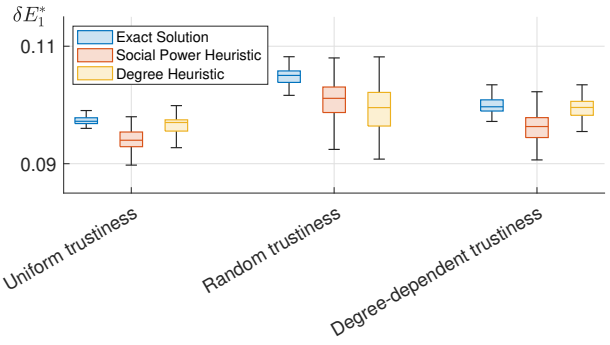


Fig. 2. Statistics of the optimal benefit δE_1^* of Problem 4, computed using the exact solution or the two heuristics when attacking 100 SW networks ($N = 40$, $m = 3$, $p = 0.2$, 4 attacked agents). All the agents have volatility $\alpha_r = 1$ and influenceability $\eta_r = 1$. Box limits are set at the lower and upper quartiles.

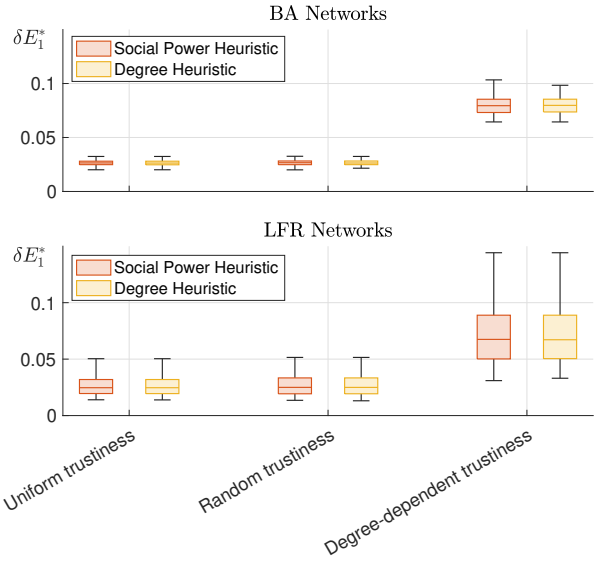


Fig. 3. Statistics of the optimal benefit δE_1^* of Problem 4 computed using the two heuristics, considering (top) 100 BA networks ($N = 1000$, $m = 3$, 4 attacked agents), or (bottom) 100 LFR networks ($N = 1000$, $d_{avg} = 6$, $\gamma = 3$, $\gamma_c = 1.35$, $\mu = 0.2$, 4 attacked agents). All the agents have volatility $\alpha_r = 1$ and influenceability $\eta_r = 1$. Box limits are set at the lower and upper quartiles.

heuristics to 100 BA and LFR networks of 1000 nodes. We can see how, in the presence of highly heterogeneous degree distribution, the Social Power Heuristic matches the Degree Heuristic for all trustiness scenarios in solving Problem 4. The Social Power Heuristic is therefore nearly optimal for hard attacks with unknown prejudices on BA and LFR networks, while it optimizes exactly the soft attacks with unknown prejudices on any network. All the results in the next section are therefore computed using this heuristic.

V. Numerical results

For our numerical experiments, we used the model of interactive Markovian agents described in Sec. II, assuming that all agents have the same volatility, $\alpha_r = 1\ \forall r$, and the same influenceability, $\eta_r = 1\ \forall r$, so that $F^{-1}H = I_N$. Note that parameters α and η always appear as the fraction α/η, so

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3293233

TABLE I
POWER-LAW EXPONENTS OF THE CURVES IN FIG. 4.

	SW	BA	LFR
Uniform	-0.98	-0.56	-0.48
Random	-0.97	-0.56	-0.47
Degree-dependent	-0.98	-0.21	-0.12

they effectively act as a single parameter in the model, related to the frequency with which agents change opinion. We created SW, BA, and LFR networks with average degree $d_{avg} = 26$. In SW networks, the rewiring probability was set to $p = 0.2$. In LFR networks, the degree distribution exponent was set to $\gamma = 3$, the same value as in BA networks, the community-size distribution exponent to $\gamma_c = 1.35$, and the mixing parameter to $\mu = 0.2$.

To validate our results on a real-world online social medium, we used the Facebook data set originally analyzed in [50], which gathers the friendship edges of a set of Facebook users in the New Orleans area in early 2009[3]. The giant component of this network (denoted by FB from now on) has $N = 63,392$ nodes and $L = 816,886$ edges (density $\rho = \frac{2L}{N(N-1)} = 4.07 \times 10^{-4}$), with clustering coefficient $C = 0.22$. The node degree d_r spans three orders of magnitude, varying from 1 to 1,098 with $d_{avg} = 25.8$. For community detection we used Louvain algorithm [51] (see [52], [53] for recent alternatives). The community analysis reveals a structure with rather large modularity $Q = 0.62$, composed of 133 communities with diversified size: 10 larger than 1,000 nodes (the largest with 16,210 nodes) and, overall, only 25 larger than 10 nodes, with a power-law community-size distribution exponent around $\gamma_c = 1.35$. Notice that the strong inhomogeneity in the node degree and in the community size are precisely the features reproduced by the LFR networks above described. Moreover, both d_{avg} and the community-size distribution exponent γ_c of the synthetic networks defined above were chosen to be consistent with the corresponding parameters of the FB network.

A. Effect of the network size on the attack

In Fig. 4 we show the value of δE_1^* after attacking 10 agents in networks of increasing size. Each point in the figure is the average result of 100 attacks with the same network size and attack strategy, on 100 randomly generated networks of the corresponding topology. In the same plots we also display the effect of attacking 10 randomly chosen agents in each network.

Since all curves appear as straight lines in log-log scale, they are well approximated by power laws with exponents reported in Table I. Notice that the slope of the lines increases with the absolute value of the exponent. Lower exponents (higher in absolute value) imply a milder effect of the attack for growing N. We observe that targeted attacks (i.e., attacks whose targets were optimized through the Social Power Heuristic) on SW networks, which have almost homogeneous degree distribution, achieve nearly the same results as a random attack, with

[3]Data available from http://socialnetworks.mpi-sws.org/data-wosn2009.html

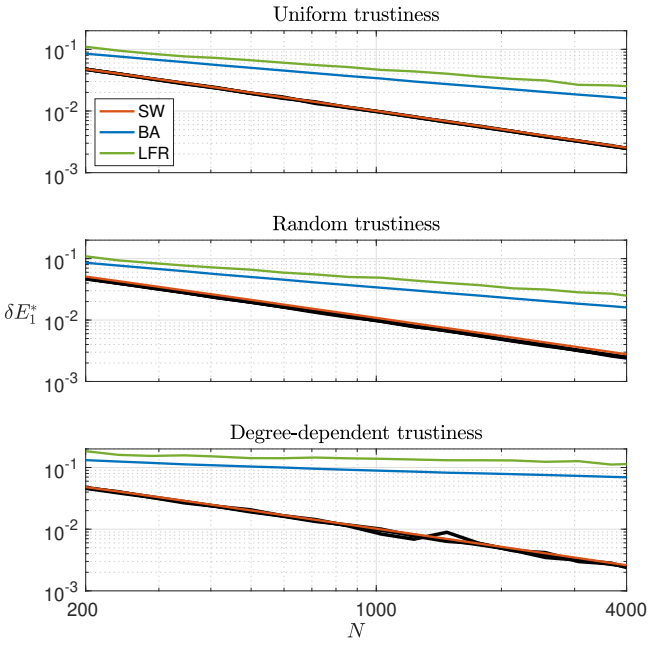


Fig. 4. Value of δE_1^* when 10 agents are attacked in networks with N ranging from 200 to 4000. Red, blue, and green curves represent the average value of δE_1^* obtained with hard attacks on 100 different networks for each value of N and for each network topology, with the Social Power Heuristic. Below each of the red SW curve appear three black curves, almost perfectly overlapping. These represent the average value of δE_1 when attacking 10 randomly chosen agents in the same SW, BA, and LFR networks.

a variation of the network expected opinion share that fits a power-law function with exponent roughly equal to -1. In other words, a random attack on any network, or the optimal attack on an SW network, achieve an effect that is roughly inversely proportional to the size of the network. On the other hand, the impact of a targeted attack with the same effort in BA and LFR networks is much larger and grows with the size of the network. In these networks, attacking a very small number of agents may have surprisingly strong effects even on very large networks, particularly with Degree-dependent trustiness. Extrapolating from the LFR curve with Degree-dependent trustiness, we infer that attacking 10 agents in a network of 10,000,000 agents we may still expect $\delta E_1^* = 0.0461$, that is, a 4.6% shift in the expected fraction of agents having opinion 1, with only a fraction of 10^{-6} attacked agents.

The difference between BA and LFR networks, *ceteris paribus*, is instead quite small. This means that the degree heterogeneity is much more relevant than the existence of communities in determining the effect of an attack on the expected opinion share.

B. Effect of community structure on the attack

Community structure is however far from irrelevant in determining how the effects of an attack propagate through the network, irrespective of the trustiness model. Elaborating on (2), we can assess the expected opinion share of a subset $\mathcal{S}$ of nodes in the network as

$$E_1(\mathcal{S}) = \frac{1}{|\mathcal{S}|}\theta_{\mathcal{S}}^{\top} M \beta,$$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3293233

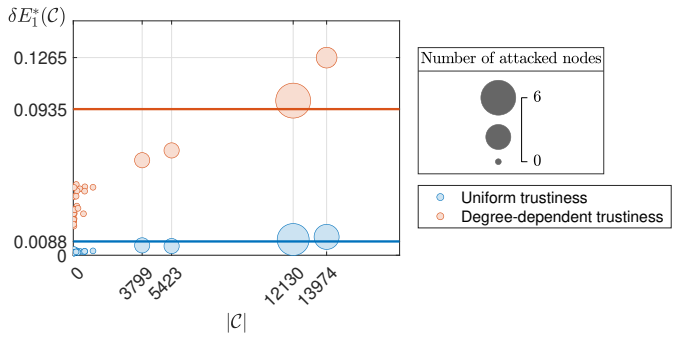


Fig. 5. LFR network, $N = 40,000$, 10 attacked agents. Variation of the expected opinion share within each community $\mathcal{C}$, as a function of the size $|\mathcal{C}|$ of the 26 communities. Each bubble corresponds to a community, and the size is proportional to the number of attacked agents in the community. The horizontal line is the expected variation of opinion share, δE_1^*, over the whole network (i.e., with uniform trustiness, the attack shifts the opinion share of the whole network by 0.88%, while with degree-dependent trustiness it shifts it by 9.35%). The results with Random trustiness overlap almost perfectly with those with Uniform trustiness and were omitted for clarity.

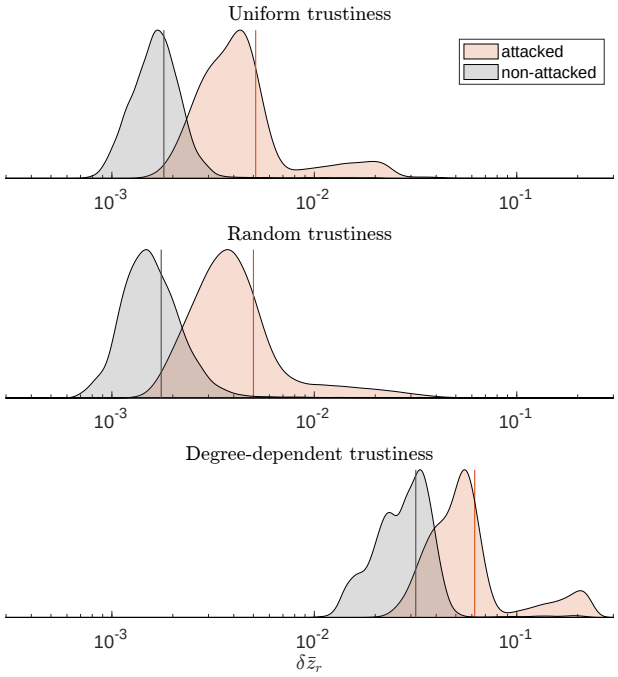


Fig. 6. Same network as in Fig. 5. Estimates of the probability densities of $\delta\bar{z}_r$ (7), for the agents in communities containing no attacked agents (gray) and at least one attacked agent (red). The solid lines mark the medians of the respective populations (notice that $\delta\bar{z}_r$ is in log scale). The maxima of the two densities were normalized to the same value for better comparability.

where the binary vector θ encodes as nonzero elements the agents in subset $\mathcal{S}$. Following the same steps that we used in Section IV to obtain (5), we can therefore assess the expected benefit of an attack encoded by binary vector v, within the set of agents $\mathcal{S}$, as

$$\delta E_1(\mathcal{S}) = \frac{1}{2|\mathcal{S}|}\theta_{\mathcal{S}}^{\top} M_v v. \tag{6}$$

In particular, we denote by $\delta E_1^*(\mathcal{C})$ the expected benefit on community $\mathcal{C}$ of the optimal attack (Problem 4, equation (5)). In Fig. 5 we report the value of $\delta E_1^*(\mathcal{C})$ for each community $\mathcal{C}$ of an LFR network with $N = 40,000$ nodes, subject to the attack of 10 agents. The effect of the attack is very uneven across communities. In the tested network, only 4 communities contained at least one attacked agent, and the expected opinion share of individuals within such communities shifts far more than that of individuals belonging to other communities. The largest community in the network has 13,974 nodes. Under network attack, for this community we obtain $\delta E_1^*(\mathcal{C}) = 0.1265$ assuming Degree-dependent trustiness, which means that the expected opinion share of its members shifts by about 12.6% or, in other words, that about 1760 members of the largest community change opinion, as a consequence of attacking just 10 nodes. The expected opinion share of members of the second-largest community shifts by little more than the global network expected opinion share, while the shift for the other two communities with attacked nodes is below the network average yet larger than any non-attacked community. With Uniform trustiness the numerical effect is smaller, but the same considerations apply. The interplay of attacks and community structure thus generates a relevant splitting in the opinion of individuals belonging to different communities. Notice, however, that the above evidence does not clarify whether the large variation in the expected opinion share of the attacked communities is the consequence of a large shift of a small group of individuals, or it is due to a more evenly distributed change throughout the whole community. Figure 6 answers this question by showing an estimate[4] of the distribution of the variation

$$\delta\bar{z} = \mathbb{E}[M(\beta^{\text{a.p}} + \Delta)] - \mathbb{E}[M\beta^{\text{a.p}}], \tag{7}$$

for agents that belong or do not belong to attacked communities. We see how the probability of having opinion 1 is significantly increased, for all trustiness models, for the majority of agents that belong to attacked communities. This means that, inside communities, opinions spread coherently via a significant mutual influence, whereas the borders of communities act, to some extent, as barriers that mitigate the contagion.

C. Validation on the FB network

Remarkably, most of the above considerations remain valid if, instead of a synthetic LFR network, we use the same heuristic to compute the optimal attack on the Facebook network FB. Figure 7 summarizes the results of the attack. The effect on communities is qualitatively similar to that on the LFR network, although the different impact on attacked vs non-attacked communities is attenuated. Interestingly, also a few very small, non-attacked communities display a variation of the expected opinion share that is above average (see the leftmost part of the plot): they are small groups of agents which, although the community detection algorithm kept technically distinct from the attacked communities, have important connections with the latter. In any case, Fig. 8 confirms that the opinion variation is significantly larger for agents in attacked communities.

[4]We used Matlab function *ksdensity*, see https://www.mathworks.com/help/stats/ksdensity.html

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2023.3293233

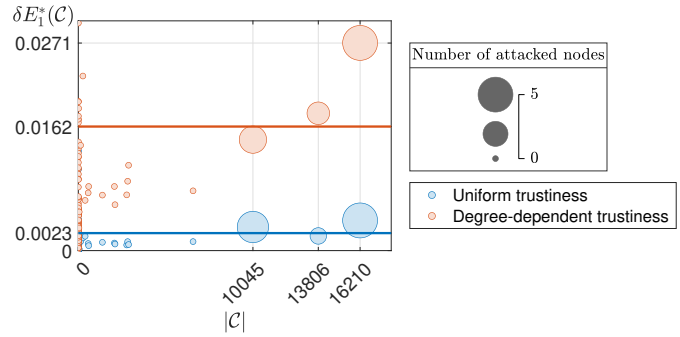


Fig. 7. Same as Fig. 5, but for the FB network, $N = 63,392$. In this case the largest community in the network has $16,210$ nodes. When this community is attacked, assuming Degree-dependent trustiness, $\delta E_1^*(\mathcal{C}) = 0.0271$, which means that the expected opinion share of members of this community shifts by about 2.7%.

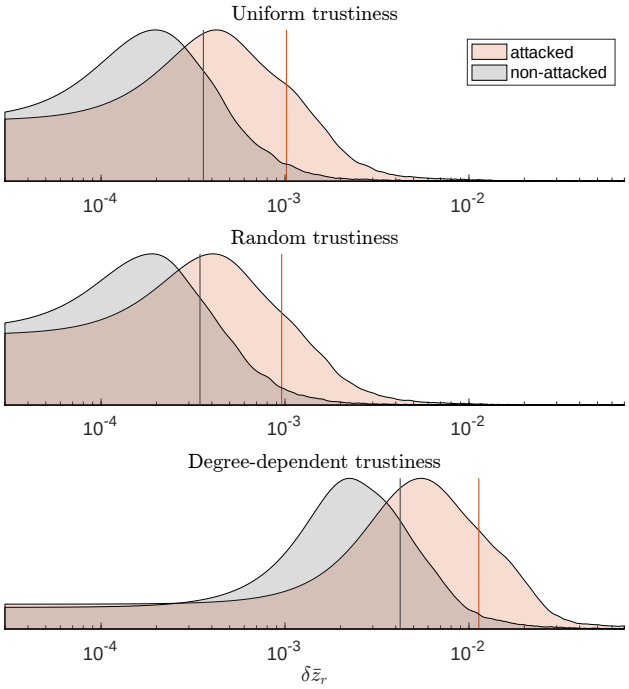


Fig. 8. Same as Fig. 6, but for the FB network, $N = 63,392$.

D. Attack targeted to a subset of the network

Now, we investigate the effects of an attack that is targeted to a subset of the network. That is, the goal is no longer to shift the opinion of the whole population, but to maximize the effect on a prescribed subpopulation, which may or may not form a community in the network. This is obtained by maximizing the benefit defined in (6), that is, by solving

$$\max_{v\in\{0,1\}^N} \delta E_1(\mathcal{S}) = \frac{1}{2|\mathcal{S}|}\theta_{\mathcal{S}}^\top M_v v,$$
$$s.t. \quad \|v\|_1 = k.$$

We see in Fig. 9 the expected variation in the opinion share of the members of each community in the same LFR network that we used in Fig. 5, with Uniform trustiness, when a hard attack is targeted to the agents in the first,

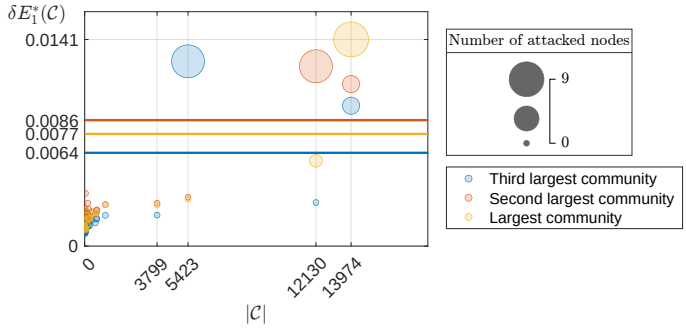


Fig. 9. Same as Fig. 5, but with an attack targeted on the third, second, or largest community of the LFR network, with Uniform trustiness.

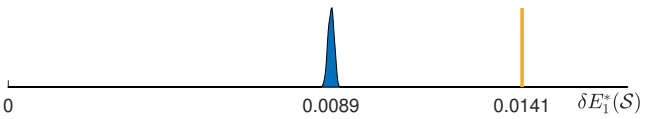


Fig. 10. Statistics of 100 attacks targeted to randomly chosen groups $\mathcal{S}$ of 13,974 agents in the network of Fig 9. The blue curve represents the estimated probability density of the values of $\delta E_1^*(\mathcal{S})$ when attacking randomly chosen groups, while the yellow line marks the value of $\delta E_1^*(\mathcal{S})$ when the target group is the largest community in the network.

second, or third largest community in the network. Attacking 10 nodes, the expected opinion share of the largest community shifts by about 1.4% if the attack is directed to the largest community. This means that about 195 nodes change opinion. The effect is reduced if the attack is directed to the second or third largest community. The whole-network expected opinion share shifts between 0.64% and 0.86% depending of which of the three largest communities is targeted. Surprisingly, the targeted attack on a community sometimes involves agents outside of the target community. The network of Fig. 9 has 14 communities of at least 100 agents. While the targeted attack on the 11 smaller communities involved only agents within the target community, the figure shows how the attack on the third or second largest community involves some agents in the largest community, and the attack on the largest community involves one agent in the second largest community: despite the strong community structure, groups of agents that belong to a community are sometimes sensitive to the opinions of some agents that are external to their community. We also see by comparing Figs. 9 and 5 (Uniform trustiness scenario) that, when the attack is targeted to a community, the variation in the opinion share of this community is larger than that achieved with an attack that is optimized to affect the whole network.

This is however not true when the attack is targeted towards a group of agents that does not form a community. Figure 10 shows the statistics of 100 attacks that were targeted towards randomly chosen groups of 13,974 agents in the network, that is, towards groups of agents of the same size as the largest community, which however do not form communities in the network. The variation in opinion share on these groups has a median of 0.0089, and in our numerical experiment never exceeds 0.00902; a significantly lower value than that obtained when the target group was the largest community (0.0141,

yellow line in Fig. 10). In other words, the fact of belonging to a community in the online social medium makes this group of agents significantly more sensitive to a targeted attack.

VI. Discussion and conclusions

In this paper, we have studied how the structure of the social network affects the attempt to manipulate, by influencing the behavior of a tiny fraction of its users, the opinions of people who communicate through an online social medium. In particular, attention has been focused on networks structured in communities, a feature that has been observed in the vast majority of social networks.

As a first result, we have established that the optimal attack strategy can be approximately implemented by attacking the nodes with the largest social power, as defined in [15]. Hence, by studying the effects of attacks directed to the nodes with largest social power, we can assess the worst-case scenario of an attack even for networks so large that the exact optimal attack is not practically computable. Then, we have seen that, in networks with low degree heterogeneity, such as in the SW network model, the effect of an attack is inversely proportional to the size of the network, and the optimal attack is not significantly more effective than one that targets random individuals in the network. Most of the real social networks however display a large degree heterogeneity, with a few nodes with very high degree (such as the influencers in online social media). When considering models with this feature (BA and LFR models, and the Facebook friendship network), we observe that the effect of an attack scales with an exponent that is much smaller (in magnitude) than 1. In other words, in networks with high degree heterogeneity the relative impact of the optimal attack grows with the size of the network, a feature that makes optimal attack strategies particularly effective on large populations [7], [23]. This gain is lost if the attack is not optimized, that is, if the targets of the attack are chosen randomly in the population. These considerations hold true independently of the trustiness model, that is, independently of whether we assume that individuals trust all their peers the same or randomly, or attribute greater trustiness to peers who are perceived as hubs (influencers) in the online social medium.

When considering social networks structured in communities, such as the LFR model or the Facebook friendship network we used for our experiments, we found that the effect of the communities on the manipulability of opinions is not trivial. On the one hand, communities act as barriers to the spreading of the manipulated opinion, so that the effect of an attack tends to be far greater in communities that contain attacked nodes, that in communities that do not contain such nodes. This confirms the existence of a blocking effect caused by community borders and, more in general, the relevant role of community structure on the breadth and speed of information diffusion [35], [38]. On the other hand, communities act as an amplification factor for the attempt of opinion manipulation: an attack to an online social medium optimized to influence the opinion of a given subset of a population is much more effective if that subset is also a community in the online social medium. This is in line with some previous results highlighting the role of communities as opinion incubators at the early stage of opinion diffusion, or as echo chambers at a later stage [7].

A potentially interesting direction of further research may regard the effect of heterogeneity in the values of the agents' volatility and influenceability parameters. With respect to the base case of homogeneous behavior, this would add behavioral diversification, and could have a significant impact on individual social power and therefore on the effectiveness of an attack. Another line of investigation regards the effect of changing the number of attacked nodes. We fixed this quantity, in an attempt to mimic a nontrivial yet reasonable attack scenario, and we explored the dependence of the result on the ratio of attacked nodes to network size by analyzing attacks on networks of increasing size. The interplay between network size, number of attacked nodes and average degree is left to further investigation.

References

[1] K. Zeng, X. Wang, Q. Zhang, X. Zhang, and F.-Y. Wang, "Behavior modeling of Internet Water Army in online forums," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 9858–9863, 2014.

[2] T. Khaund, B. Kirdemir, N. Agarwal, H. Liu, and F. Morstatter, "Social bots and their coordination during online campaigns: A survey," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 530–545, 2022.

[3] C. Cheng, Y. Luo, and C. Yu, "Dynamic mechanism of social bots interfering with public opinion in network," *Physica A: Statistical Mechanics and its Applications*, vol. 551, p. 124163, 2020.

[4] T. Mihaylov, T. Mihaylova, P. Nakov, L. Màrquez, G. D. Georgiev, and I. K. Koychev, "The dark side of news community forums: opinion manipulation trolls," *Internet Research*, vol. 28, no. 5, pp. 1292–1312, oct 2018.

[5] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 258–265.

[6] K. H. Jamieson, *Cyberwar: How Russian hackers and trolls helped elect a President: What we don't, can't, and do know*. Oxford University Press, 2020.

[7] E. Ferrara, H. Chang, E. Chen, G. Muric, and J. Patel, "Characterizing social media manipulation in the 2020 U.S. Presidential election," *First Monday*, vol. 25, no. 11, Oct. 2020.

[8] S. Aral and D. Eckles, "Protecting elections from social media manipulation," *Science*, vol. 365, no. 6456, pp. 858–861, 2019.

[9] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[10] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. part I," *Annual Reviews in Control*, vol. 43, pp. 65–79, 2017.

[11] ——, "A tutorial on modeling and analysis of dynamic social networks. part II," *Annual Reviews in Control*, vol. 45, pp. 166–190, 2018.

[12] R. A. Holley and T. M. Liggett, "Ergodic theorems for weakly interacting infinite systems and the voter model," *The Annals of Probability*, pp. 643–663, 1975.

[13] K. Sznajd-Weron, J. Sznajd, and T. Weron, "A review on the Sznajd model—20 years after," *Physica A: Statistical Mechanics and its Applications*, vol. 565, p. 125537, 2021.

[14] S. Galam, "Sociophysics: A review of Galam models," *International Journal of Modern Physics C*, vol. 19, no. 03, pp. 409–440, 2008.

[15] P. Bolzern, P. Colaneri, and G. De Nicolao, "Opinion dynamics in social networks with heterogeneous Markovian agents," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 6180–6185.

[16] ——, "Opinion influence and evolution in social networks: A Markovian agents model," *Automatica*, vol. 100, no. 2, pp. 219–230, 2019.

[17] ——, "Opinion dynamics in social networks: The effect of centralized interaction tuning on emerging behaviors," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 362–372, 2020.

[18] ——, "Effect of social influence on a two-party election: A Markovian multiagent model," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 3, pp. 1056–1067, 2022.

[19] ——, "Multi-opinion Markovian agent networks: Parametrization, second order moment and social power," *Automatica*, vol. 153, p. 111026, 2023.

[20] E. Yildiz, A. Ozdaglar, D. Acemoglu, A. Saberi, and A. Scaglione, "Binary opinion dynamics with stubborn agents," *ACM Transactions on Economics and Computation*, vol. 1, no. 4, 2013.

[21] C. J. Kuhlman, V. A. Kumar, and S. Ravi, "Controlling opinion propagation in online networks," *Computer Networks*, vol. 57, no. 10, pp. 2121–2132, 2013.

[22] G. Romero Moreno, E. Manino, L. Tran-Thanh, and M. Brede, "Zealotry and influence maximization in the voter model: when to target partial zealots?" in *Complex Networks XI: Proceedings of the 11th Conference on Complex Networks CompleNet 2020*, H. Barbosa, J. Gomez-Gardenes, B. Gonçalves, G. Mangioni, R. Menezes, and M. Oliveira, Eds. Springer, 2020, pp. 107–118.

[23] K. Chiyomaru and K. Takemoto, "Adversarial attacks on voter model dynamics in complex networks," *Physical Review E*, vol. 106, no. 1, p. 014301, 2022.

[24] R. Bredereck and E. Elkind, "Manipulating opinion diffusion in social networks," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 894–900.

[25] A. N. Zehmakan, "Majority opinion diffusion in social networks: An adversarial approach," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 6, pp. 5611–5619, May 2021.

[26] H. Noorazar, "Recent advances in opinion propagation dynamics: A 2020 survey," *The European Physical Journal Plus*, vol. 135, pp. 1–20, 2020.

[27] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, ser. LinkKDD '05. New York, NY, USA: ACM, 2005, pp. 36–43.

[28] V. Red, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM Review*, vol. 53, no. 3, pp. 526–543, 2011.

[29] S. Das and A. Biswas, "Deployment of information diffusion for community detection in online social networks: A comprehensive review," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 5, pp. 1083–1107, 2021.

[30] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[31] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016.

[32] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 23, pp. 8577–8582, 2006.

[33] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Scientific Reports*, vol. 3, AUG 28 2013.

[34] D. M. Romero, C. Tan, and J. Ugander, "On the interplay between social and topical structure," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013, pp. 516–525.

[35] A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn, "Optimal network modularity for information diffusion," *Physical Review Letters*, vol. 113, p. 088701, Aug 2014.

[36] Z.-L. Hu, Z.-M. Ren, G.-Y. Yang, and J.-G. Liu, "Effects of multiple spreaders in community networks," *International Journal of Modern Physics C*, vol. 25, no. 05, p. 1440013, mar 2014.

[37] Noorazar, Hossein, "Recent advances in opinion propagation dynamics: a 2020 survey," *Eur. Phys. J. Plus*, vol. 135, no. 6, p. 521, 2020.

[38] H. Peng, A. Nematzadeh, D. M. Romero, and E. Ferrara, "Network modularity controls the speed of information diffusion," *Physical Review E*, vol. 102, p. 052316, Nov 2020.

[39] W. Tang, L. Tian, X. Zheng, G. Luo, and Z. He, "Susceptible user search for defending opinion manipulation," *Future Generation Computer Systems*, vol. 115, pp. 531–541, 2021.

[40] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," *Theory of Computing*, vol. 11, pp. 105–147, 2015.

[41] S. Banerjee, M. Jenamani, and D. K. Pratihar, "A survey on influence maximization in a social network," *Knowledge and Information Systems*, vol. 62, no. 9, pp. 3417–3455, 2020.

[42] C. G. Cassandras and S. Lafortune, *Introduction to discrete event systems*. Springer, 2008.

[43] C. Briat, "Sign properties of Metzler matrices with applications," *Linear Algebra and its Applications*, vol. 515, pp. 53–86, 2017.

[44] J. R. French Jr, "A formal theory of social power." *Psychological review*, vol. 63, no. 3, p. 181, 1956.

[45] A. L. Barabási, *Network Science*. Cambridge University Press, 2016.

[46] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[47] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[48] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical Review E*, vol. 78, no. 4, Part 2, p. 046110, 2008.

[49] M. Crain and A. Nadler, "Political Manipulation and Internet Advertising Infrastructure," *Journal of Information Policy*, vol. 9, pp. 370–410, 12 2019.

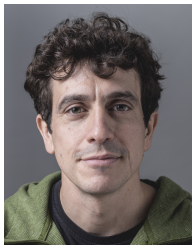
[50] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

[51] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics - Theory and Experiment*, p. P10008, 2008.

[52] H.-J. Li, Z. Bu, Z. Wang, and J. Cao, "Dynamical clustering in electronic commerce systems via optimization and leadership expansion," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5327–5334, 2020.

[53] H. Li, W. Xu, C. Qiu, and J. Pei, "Fast Markov clustering algorithm based on belief dynamics," *IEEE Transactions on Cybernetics*, vol. 53, no. 6, pp. 3716–3725, 2023.

Paolo Bolzern received his Doctor's degree (Laurea) in Electronic Engineering from Politecnico di Milano in 1978. From 1983 to 1987 he was a research fellow with the Centro di Teoria dei Sistemi of the National Research Council (CNR). In 1987 he joined Politecnico di Milano, where he is currently a Full Professor of Automatic Control. He is a member of IEEE, Associate Editor of IEEE Control Systems Letters, and past Associate Editor of IEEE Transactions on Automatic Control (2015-2018). He was appointed as Electronic Media Editor for IFAC (2015-2020) and served in the Organizing Committee of the IFAC World Congresses of Milano (2011) and Toulouse (2017). His current research interests include coordination of multi-agent systems, positive systems, control of deterministic and stochastic switched systems, and opinion dynamics modelling.



Alessandro Colombo received the Diplôme d'Ingénieur from ENSTA in 2005, and the Ph.D. from Politecnico di Milano in 2009. He was Postdoctoral Associate at the Massachusetts Institute of Technology in 2010-2012, and is currently Associate Professor in the Department of Electronics, Information and Bioengineering at Politecnico di Milano. His research interests are in the analysis and control of hybrid systems and in applications of control theory to the modelling and understanding of different human activities.

Carlo Piccardi is Full Professor of Systems and Control at Politecnico di Milano. His activity is in the area of complex systems and networks, dynamical systems and automatic control. His recent areas of research include the analysis of the structural properties of networks, with applications in economics, finance, and social sciences; spreading processes on complex networks; and synchronization of networks of dynamical systems.