

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/130154>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# Convergence of Opinion Diffusion is PSPACE-complete

Dmitry Chistikov<sup>1</sup> and Grzegorz Lisowski<sup>1</sup> and Mike Paterson<sup>1</sup> and Paolo Turrini<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Warwick, United Kingdom  
{d.chistikov, grzegorz.lisowski, m.s.paterson, p.turrini}@warwick.ac.uk

## Abstract

We analyse opinion diffusion in social networks, where a finite set of individuals is connected in a directed graph and each simultaneously changes their opinion to that of the majority of their influencers. We study the algorithmic properties of the fixed-point behaviour of such networks, showing that the problem of establishing whether individuals converge to stable opinions is PSPACE-complete.

## Introduction

Social networks are a well established paradigm for the computational analysis of real-world phenomena such as disease spreading (Klov Dahl 1985; Jackson 2010), product adoption (Apt and Markakis 2014; Apt, Markakis, and Simon 2016) and opinion diffusion (Axelrod 1997; Grandi, Lorini, and Perrussel 2015). They are typically modelled as directed graphs over a finite set of individuals possessing certain properties, such as opinions, which spread through the network according to predefined rules. For instance, protocols for the spread can be defined as a function of the influencers' opinions.

In the plethora of social network models, threshold-based ones are certainly the best known. There, agents adopt an opinion if and only if it is shared by a given threshold of the incoming connections. While these models have a long standing tradition in the social sciences, originating from Granovetter (1978), they have received revived attention in artificial intelligence, including contributions of Ferraioli, Goldberg, and Ventre (2016), Auletta et al. (2017) or Bilò, Fanelli, and Moscardelli (2018). Notably, Auletta, Ferraioli, and Greco have received the IJCAI 2018 Distinguished Paper Award for a study of communication in threshold-based social network models.

One of the major challenges associated with these models is that convergence of the diffusion protocol is not guaranteed. Imagine you would like to have your agents make a collective decision and let them discuss first, agreeing that they would cast their vote once they have made up their mind. Depending on the chosen diffusion protocol and the initial distribution of opinions, the process might never terminate. This is the case for synchronous threshold models. Clearly, any network will converge for *some* initial input, for instance when your

agents already think the same to start with. However this is not true in general.

The typical path taken to circumvent the issue is to restrict the analysis to networks that always converge, as studied by Grandi, Lorini, and Perrussel (2015), Bredereck and Elkind (2017) and Botan, Grandi, and Perrussel (2019). Another is to consider specific protocols which guarantee termination, as done for instance by Auletta, Ferraioli, and Greco (2018): they propose an opinion-revision protocol for agents who disagree with a distinguished opinion.

Recently, Christoff and Grossi (2017) have provided a characterisation of networks in which termination of the threshold-based opinion diffusion protocol is guaranteed. However, we still do not know whether characterising convergent networks is of any advantage for their algorithmic analysis, in other words, whether we can have a characterisation that is easier to check than actually running the protocol until converging or looping in some way. Here, we settle this problem.

**Our contribution.** We study the convergence of opinion diffusion in social networks, modelled as directed graphs over a finite set of individuals, who simultaneously update their opinions. They switch their opinions if and only if the majority of their influencers disagrees with them. We look at labelled networks, where individuals start with a binary opinion, and study the problem of whether that network converges. We also look at unlabelled networks and consider the problem of whether a labelling exists for which the network does not converge — this problem concerns the *structural* aspect of opinion diffusion's convergence. Our contribution is two-fold: firstly, we present some classes of networks which are guaranteed to converge, and secondly we show that the problem of establishing whether a network converges is PSPACE-complete even for the simplest of such protocols, closing a gap in the literature. In fact, we show that any characterisation of such networks, including the one provided by Christoff and Grossi (2017) cannot result in an efficient procedure for verifying the convergence of the considered protocol (unless  $P=PSPACE$ ).

We emphasize that even though our protocol is relatively simple, the computational complexity lower bounds that we obtain extend directly to more general models. For instance,

the PSPACE-hardness of the considered problems lifts to the scenario in which each agent has its own specific update threshold. So our result implies that no complete characterisation of convergent networks can be efficiently computed in practice for a wide range of plausible diffusion protocols.

**Related literature** Our results have implications for various lines of research using opinion diffusion models.

**Social Influence Models** The graph-like structure of social networks has attracted interest in computer science, with studies of the influence weight of nodes in the network (Kempe, Kleinberg, and Tardos 2005) and the properties of the influence function (Grabisch and Rusinowska 2010). Social influence has been widely analysed in the social sciences, from the point of view of strategic behaviour (Isbell 1958) and its implications for consensus creation (de Groot 1974) and cultural evolution (Axelrod 1997).

**Opinion Manipulation Models** Issues of convergence are extremely relevant to models that deal with opinion manipulation. For instance, Brederick and Elkind (2017) study a scenario where an external agent wishes to transform the opinion of a number of members of a network to induce desired fixed-point conditions. Further, control of collective decision-making (Faliszewski and Rothe 2016) is an important topic in algorithmic mechanism design: the difficulty of establishing whether manipulation is a real threat is paramount for system security purposes.

**Deliberative Democracy and Social Choice** Opinion diffusion underpins recent models of deliberative democracy, in terms of delegation (Dryzek and List 2003), representation (Endriss and Grandi 2014), and stability (Christoff and Grossi 2017). Formal models of democratic representation build on an underlying consensus-reaching protocol (de Groot 1974; Brill 2018). Social networks have also become of major interest to social choice theory, with propositional opinion diffusion (Grandi, Lorini, and Perussel 2015) emerging as a framework for social choice on social networks (Grandi 2017).

**Related computational models.** If the social networks are modelled as undirected, rather than directed, graphs, it has long been known that convergence takes at most a polynomial number of steps under majority updates (Chacc, Fogelman-Soulié, and Pellegrin 1985). In these models, PSPACE-hardness results have only been shown for more powerful *block sequential* update rules (Goles et al. 2016).

Convergence is a PSPACE-complete property in various related models, notably directed discrete Hopfield networks (Orponen 1993) and Boolean dynamical systems (see, e.g., Barrett et al. 2003 and 2007). Hardness in these results (and their strengthenings, as studied by Ogihara and Uchizawa (2017), Rosenkrantz et al. (2018) and Kawachi, Ogihara, and Uchizawa (2019)) crucially depends on the availability of functions that *identify* 0 and 1 (see the discussion of the ingredients for the hardness proofs later on). Opinion diffusion is instead based on self-dual functions, where flipping all inputs to a self-dual function always leads to flipping its output. In



Figure 1: On the left side, an unlabelled social network. On the right side, one of its labellings. Throughout the paper nodes coloured in black correspond to labelling 1, white nodes to labelling 0 and grey nodes are unlabelled.

other words, in the setting we consider the diffusion protocol is symmetric with respect to opinions held by agents.

Whilst Kosub (2008) shows the NP-completeness of deciding the existence of a fixed-point configuration if *all* self-dual functions are available, our update rule, in comparison, is monotone (i.e., has no negation). Moreover, sparse graphs of bounded fan-in — with each agent having up to six influencers — suffice for our proof of PSPACE-hardness. In the related model of cellular automata, known results show that majority is “arguably the most interesting” local update rule (Tosic 2017).

**Paper structure** We first present our basic setup and examples of networks whose convergence is easy to check. Subsequently we prove that determining convergence is PSPACE-complete. Finally, we conclude by discussing the ramifications of our results and future research directions.

## Opinion Diffusion

**Social Networks.** Let  $N = \{1, 2, \dots, n\}$  be a finite set of agents and  $E$  be a simple directed graph over  $N$ , i.e., an irreflexive relation over the set of agents. We call a tuple  $(N, E)$  a *social network*. The idea is that each agent is influenced by the incoming edges and influences the outgoing ones. For each  $i \in N$  we define the set  $E[i] = \{j \mid (i, j) \in E\}$ , i.e., the set of agents that  $i$  influences. Similarly, we define the set  $E^{-1}[i] = \{j \mid (j, i) \in E\}$ , the *influencers* of  $i$ .

We are interested in how opinions spread in a social network following the influence relation. For this we equip agents with opinions, giving *labelled social networks*.

**Definition 1** (Labelled Social Network). A labelled social network is a tuple  $SN = (N, E, f)$ , where:

- $(N, E)$  is a social network,
- $f : N \rightarrow \{0, 1\}$  is a binary labelling of each node.

Figure 1 gives examples of an unlabelled and a labelled social network.

**Opinion Diffusion Protocol.** We model opinion change as an update protocol on the network where each agent  $i$  takes the opinion of their influencers, i.e.,  $E^{-1}[i]$ , into account.

For a given labelled social network  $(N, E, f)$ , and an agent  $i$  let us call  $A(i) = \{j \in E^{-1}[i] \mid f(i) = f(j)\}$  the set of influencers who agree with  $i$ ’s opinion, and  $D(i) = E^{-1}[i] \setminus A(i)$  the ones who do not.



Figure 2: On the left, a convergent (labelled) social network. On the right, a non-convergent one.

We assume agents change their opinion if the fraction of their influencers disagreeing with them is (strictly) higher than a half. In particular, a node with  $2k$  influencers always takes the opinion of the majority of itself and these influencers.

**Definition 2** (Opinion Change). *Let  $SN = (N, E, f)$  be a labelled social network and  $i \in N$  be an agent. Then the opinion diffusion step is the function  $OD : N \rightarrow \{0, 1\}$  such that*

$$OD(SN, i) = \begin{cases} \text{flip}(f(i)) & \text{if } |D(i)| > |A(i)| \\ f(i) & \text{otherwise} \end{cases}$$

where  $\text{flip}(k) = 1 - k$  denotes the change from an original opinion to its opposite value.

We are now ready to define the protocol for the evolution of a labelled social network. Here we focus on the *synchronous update*, in which all agents modify their opinions at the same time.

**Definition 3** (Synchronous Update). *Let  $SN = (N, E, f)$  be a labelled social network. Then,  $SU(SN) = (N, E, f')$  is a social network such that for any  $i \in N$ ,  $f'(i) = OD(SN, i)$ .*

The synchronous update protocol is deterministic: given a labelled social network we can compute its state after any given number of synchronous updates. An *update sequence* of a labelled social network  $SN$  is the infinite sequence of states of  $SN$  after successive synchronous updates.

**Definition 4** (Update Sequence). *Given a labelled social network  $SN = (N, E, f)$ , the update sequence generated by  $SN$  is the sequence of labelled social networks  $SN_{us} = (SN_0, SN_1 \dots)$  such that  $SN_0 = SN$  and for every  $n \in \mathbb{N}$ ,  $SN_{n+1} = SU(SN_n)$ .*

For a labelled social network  $SN$  and agent  $i$  we denote by  $f_k(i)$  the value given to agent  $i$  at time  $k$ , i.e., at the  $k$ -th update step. We call a social network  $SN$  *stable* if  $SU(SN) = SN$ . A social network is *convergent* if its update sequence contains a stable social network, i.e., if its update sequence reaches a fixed point, its *limit network*.

## Graph Restrictions

Some networks converge for all initial labellings, while others converge for just some labellings. The lefthand network in Figure 1, for example, converges for every labelling. However, the social networks displayed in Figure 2 behave differently. Here we look at specific instances of social networks which converge for every labelling.

Let us start with DAGs, i.e., *directed acyclic graphs*.

**Proposition 1.** *Let  $SN=(N, E)$  be a DAG. Then  $SN$  converges in at most  $k$  steps for every labelling  $f$ , where  $k$  is the length of the longest path.*

*Proof.* Given a DAG  $SN=(N, E)$ , consider an arbitrary labelled social network  $SN' = (N, E, f)$ . Let us write  $i \rightarrow j$  for  $j \in E[i]$ . Since  $SN$  is acyclic, for every  $i \in N$  there is a path to  $i$  from some source node of  $SN$ . Let  $\text{level}(i)$  be the length of the longest such path. We will show by induction on  $\text{level}(i)$  that every  $f(i)$  will stabilise after at most  $\text{level}(i)$  updates.

If  $\text{level}(i)$  is 0 then  $i$  is a source node and therefore never changes. Suppose that all  $i$  such that  $\text{level}(i) = r$  have stabilised after  $r$  updates. Take any node  $i$  with  $\text{level}(i) = r + 1$ . Since  $SN$  is acyclic, for any  $n' \in N$  such that  $n' \rightarrow i$ , we have  $\text{level}(n') \leq r$ . This means  $n'$  is already stable after  $r$  updates. Hence,  $i$  will stabilise within one step after all its influencers have stabilised, i.e., after at most  $r + 1$  updates.  $\square$

Networks that are not DAGs do not always converge, as shown in Figure 2. But some of these have interesting properties with respect to convergence. For example cliques, i.e., networks  $SN = (N, E)$  with  $E = N^2 \setminus \{(i, i) \mid i \in N\}$ .

**Proposition 2.** *Let  $SN=(N, E)$  be a clique.  $SN$  converges for every labelling if and only if  $|N|$  is odd. Moreover, if  $SN$  converges, then it does so after a single update step.*

*Proof.* It is easy to check that if  $SN$  is evenly split (and therefore of even size) then every agent flips at each update step. Otherwise, after one update every agent has the opinion of the initial majority.  $\square$

As checking whether a social network is a clique can be achieved by just counting its edges, the result above shows that for some structures establishing convergence is immediate.

Consider now the *strongly connected components* (SCCs) of a social network, i.e., subgraphs that have a path from each node to every other node and are maximal with respect to set inclusion. As is well-known (see e.g., Bollobás 1998), each network  $SN = (N, E)$  can be partitioned into SCCs, yielding a DAG  $SCC_{SN} = (SCCs, E')$  where: (i)  $SCCs$  is the set of all SCCs of  $SN$ ; (ii) for any  $SCC_u, SCC_v \in SCCs$ ,  $(SCC_u, SCC_v) \in E'$  iff for some  $i \in SCC_u, j \in SCC_v$  we have that  $j \in E[i]$ . Recall, that the set of SCCs of  $SN$  can be computed in linear time in the size of  $SN$ .

One might expect that if we knew that each SCC always converges then so would the whole network, or, put otherwise, that every network that always converges will also do so when only influenced by a network that itself always converges. Remarkably, this is not true even for very simple cases, as exemplified in Figure 3.

We now move on to the problem of checking convergence in an arbitrary social network.

## The Complexity of Checking Convergence

We consider two computational problems with respect to the protocol we are considering. The first of them is checking the convergence of a given labelled social network.

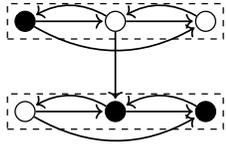


Figure 3: A labelled network that does not converge, whose two SCCs (marked by rectangles) do converge for every initial labelling. The convergence of the SCC in the lower tier is influenced by the incoming edge from the upper SCC.

**CONVERGENCE:**

*Input:* Social network  $SN = (N, E)$  and labelling  $f$ .

*Output:* Does  $SN$  converge from  $f$ ?

The second is checking for an unlabelled network whether there is a labelling which *does not* converge.

**CONVERGENCE GUARANTEE:**

*Input:* Social network  $SN = (N, E)$ .

*Output:* Is there a labelling of  $SN$  from which  $SN$  does not converge?

In the remainder of this section we will prove theorems associated with these two computational problems.

**Theorem 1.** CONVERGENCE is PSPACE-complete.

**Theorem 2.** CONVERGENCE GUARANTEE is PSPACE-complete.

It is important to note that the lower bounds apply to all opinion diffusion models for which our protocol is a special case. In particular, this holds in models with agent-dependent update thresholds or with weighted trust levels (i.e., with weighted majority instead of majority).

Let us notice that both problems belong to PSPACE, because each labelling of a social network  $SN = (N, E)$  takes  $|N|$  bits and the synchronous update mapping  $SU$  can be evaluated in polynomial time.

The hardness proof of Theorem 1 can be developed separately, but we choose to give a uniform presentation and derive hardness of both problems from the same construction, in order to make the proof of Theorem 2 easier to follow.

**Ingredients for the hardness proofs**

The main technical challenge for the hardness proof is that the update mapping  $SU$  applies a self-dual Boolean function (majority): if a node and all its influencers flip their opinion, then, after the update, the node will have the flipped value. This means, informally, that the nodes are indifferent to the identity of 0 and 1, which makes a direct simulation of propositional logic impossible.

Our construction below is, in hindsight, reminiscent of the observation that the *negation of the 3-input majority* is a basis for the class of all self-dual functions (Post 1941); see, e.g., (Lau 2006, Theorem 3.2.3.2). Our proof, however, does not rely on any advanced topics in the theory of Boolean functions and their clones/closed classes.

**Propositional logic and dual rail encoding.** Let us introduce the basic technical notions appearing in the proofs of

hardness of the considered problems. We will use Boolean circuits from computational complexity theory. Due to space constraints we omit the detailed introduction of Boolean circuits, which can be found, e.g., in (Papadimitriou 1994, section 4.3). Signals in these circuits are Boolean values, true and false, and we will encode them in our social networks. We need to encode logical gates (AND and NOT) and constant gates (TRUE and FALSE) too.

We use the dual rail encoding due to the monotonicity of the opinion diffusion protocol. Indeed, in the current setting opinions reinforce themselves, so logical negation cannot be directly simulated.

In the dual rail encoding, instead of considering individual nodes in a social network, we will be often considering related pairs of nodes, called *dual pairs*. The two nodes in a dual pair are ordered. Given a labelling of the network, a dual pair is *valid* if its two nodes disagree, i.e., take different values, and *invalid* otherwise. Dual pairs will be building blocks in our construction, and our network will have a mechanism to ensure their validity.

Our first step is to build constant gates. We introduce a distinguished dual pair, the *base pair*; as long as it is valid, we assume without loss of generality that its two nodes have values (1, 0). There is only one base pair in the network. Now for every valid dual pair in the network, we interpret (1, 0) as *true* and (0, 1) as *false*.

The next step is to build logical gates. All these gates in our circuits have fan-in 1 or 2, that is, each gate receives input from at most 2 other gates. The gates are depicted in Figures 4 and 5 and described in Example 1.

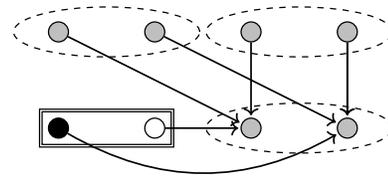


Figure 4: The AND gadget.

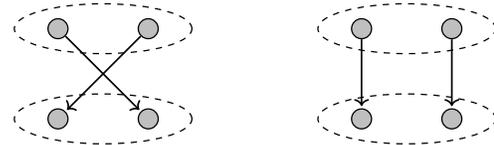


Figure 5: NOT gadget on the left, NOP gadget on the right.

**Example 1.** The gadget in Figure 4 models an AND gate, and the gadget in Figure 5 (left) models a NOT gate. The AND gadget relies on the base pair, which is depicted as a double rectangle. In more detail, if at time  $t$  the input dual pairs (the two upper ovals) in the AND gadget are valid, then at time  $t + 1$  the output dual pair is valid and represents the AND of the two input values (and similarly for the NOT gate). Finally, the gadget in Figure 5 (right) models a NOP (no-operation) gate: at time  $t + 1$  the output pair is a copy of the input pair at time  $t$ .

**Turing machines.** Further in our reduction we will need to build Boolean circuits to simulate the behaviour of Turing machines.

We will describe a restricted version of Turing machines that we use to prove Theorems 1 and 2. These Turing machines are *polynomially space-bounded*, or *PSPACE machines* (referring to the complexity class); see, e.g., (Arora and Barak 2009, section 4.2) and (Papadimitriou 1994, chapter 19) for a more detailed discussion.

We will not need a formal definition of Turing machines in this paper and will instead rely on the following properties only:

1. Any Turing machine has a finite description.
2. Any Turing machines can be *run* on arbitrary input strings of arbitrary length  $m \geq 0$  over a fixed finite alphabet.
3. At any point during a run, an instantaneous description of a Turing machine  $M$  (a *configuration*) can be encoded by a bit string of length  $c \cdot m^d$ , where the constants  $c$  and  $d$  depend only on the machine  $M$ .
4. A Turing machine may either *halt* at some point during the run, or *diverge* (run forever).
5. A run is a finite or infinite sequence of configurations; each configuration is either *halting* or has a unique *successor* configuration.

We will identify configurations of Turing machines with their encodings as  $n$ -bit strings (strings of truth values). Here  $n = c \cdot m^d$ ; when  $m$  is fixed,  $n$  is the same in all possible configurations.

For a given  $n$ , we will assume for the sake of simplicity that all  $n$ -bit strings represent valid configurations. This assumption does not invalidate our reduction and can in fact be eliminated using the technique of the following lemma.

**Lemma 1.** *Given a Turing machine  $M$  and an integer  $n \geq 1$ , there exists an acyclic social network  $SN$  with the following properties:*

- $SN$  contains the base pair and has  $2n$  further sources and  $2n$  sinks, grouped into  $n$  and  $n$  dual pairs;
- every path from a source to a sink has the same length  $h$ , independent of  $n$ ;
- $SN$  simulates  $M$ : if at time  $t$  the base pair and input dual pairs are valid and represent a configuration  $s^{(0)} \in \{0, 1\}^n$ , then at time  $t + h$  if  $s^{(0)}$  is non-halting the output dual pairs are valid and represent  $s^{(1)}$ , the successor configuration of  $s^{(0)}$ ; otherwise at least one output dual pair at time  $t + h$  is invalid;
- $SN$  can be constructed in time polynomial in  $n$  and in the description of  $M$ .

*Proof.* The assertion relies on the observation (following the lines of (Arora and Barak 2009, Theorem 6.6), or (Papadimitriou 1994, section 8.2)) that for every polynomially space-bounded Turing machine  $M$  and every integer  $n$ , there exists a Boolean circuit which:

- has  $n$  inputs and  $n$  outputs,

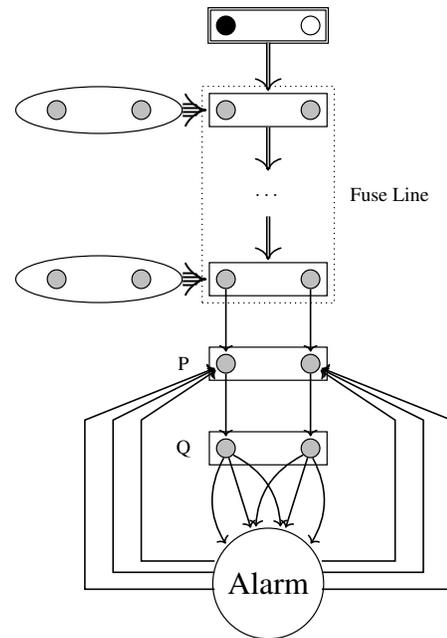


Figure 6: The fuse line.

- has equal-length paths from inputs to outputs (where this length  $h$  is independent of  $n$ ),
- transforms an arbitrary non-halting configuration of  $M$  into its successor configuration.
- can be constructed in time polynomial in  $n$  and in the description of  $M$ .

These properties map into the assertions of the lemma, using dual pairs as nodes in the circuit, and AND and NOT gadgets from Example 1 as gates. To make the network satisfy the second assertion of the lemma, we extend it using NOP gadgets where necessary.  $\square$

**Fuse line, valve, and alarm.** We will need a mechanism to check initial validity of dual pairs in our construction, as well as to detect the halting of a Turing machine, following Lemma 1. If a dual pair is or becomes invalid, this will force the convergence of the social network.

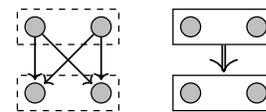


Figure 7: Two pairs in the fuse line, one feeding into the other.

Left: in detail. Right: simplified drawing (corresponding to connections between pairs in the fuse line as depicted in Figure 6), abbreviating the connections in the left picture.

The mechanism consists of a *fuse line* (sequence of pairs of nodes) leading to a *valve* and *alarm* (an even clique), as shown in Figure 6.

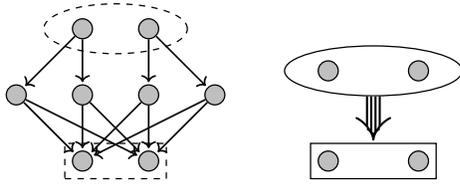


Figure 8: Dual pair connected to pair from the fuse line. Left: in detail. Note how the influence of the input pair on the output pair is stronger than in Figure 7. Right: simplified drawing (used in Figure 6), abbreviating the connections in the left picture.

Let us first discuss the fuse line itself. Pairs of nodes in the fuse line are depicted by rectangles. Each pair in the fuse line (except for the last) feeds into the succeeding pair as shown in Figure 7. In addition, all other dual pairs in the entire network (depicted for the sake of clarity as ovals) will also connect to distinct pairs in the fuse line as shown in Figure 8. We will not think of the pairs in the fuse line as dual pairs.

At the end of the fuse line shown in Figure 6, the big circle is a clique of  $2k$  nodes (an *alarm*),  $k \geq 2$ , and the *valve mechanism* is formed by the two rectangles (pairs)  $P$ ,  $Q$ , and the alarm. Both nodes of pair  $Q$  have edges to each node in the alarm, and all nodes in the alarm have edges to both nodes of pair  $P$ . In the following analysis, we say that the alarm is *evenly split* if exactly  $k$  of its nodes are labelled 0. We say that the alarm *goes off* at time  $t$  if all of its nodes agree at this time (we will usually imply that this was not the case at time  $t - 1$ ).

We show now several properties of this network which will be crucial for the PSPACE-hardness reduction.

**Lemma 2.** *If at time  $t \geq 1$  a pair in the fuse line is invalid, then:*

- (a) *it remains invalid forever and*
- (b) *the succeeding pair is invalid from time  $t + 1$  on.*

*Proof.* Assertion (a) follows the fact that the two nodes in any single pair in the fuse line have the same set of influencers.

In order for assertion (b) to fail, the succeeding pair must be valid at times  $t$  and  $t + 1$ . Since, again, the two nodes in this succeeding pair have the same set of (six) influencers, this set should be evenly split at time  $t$ . But this is impossible, because two of these influencers agree by the assumption of the lemma, and the remaining four cannot be split into 1 and 3 for every  $t \geq 1$  by the construction of the connection in Figure 8.  $\square$

**Lemma 3.** *If some dual pair in the network is invalid at some time, then the last pair in the fuse line becomes invalid at some time and remains invalid forever.*

*Proof.* Follows directly from Lemma 2.  $\square$

The final part of our construction of the network is that every node in the alarm has edges to every node in the network, except for the fuse line, nodes connecting dual pairs to pairs in the fuse line and pair  $Q$  of the valve (in other words, to all

dual pairs and to pair  $P$ ). This includes the two nodes of the base pair (depicted, as previously, as a double rectangle).

**Lemma 4.** *Suppose at time  $t$  at least one of the following conditions holds:*

- (a) *the last pair in the fuse line is invalid,*
- (b) *the two nodes of  $P$  agree,*
- (c) *the two nodes of  $Q$  agree, or*
- (d) *the alarm is not evenly split.*

*Then by time  $t + 3$  the alarm goes off and by time  $t + 6$  all nodes in the network agree.*

*Proof idea.* If the alarm does *not* go off, then either (i) it remains evenly split (and the nodes in each of the pairs  $P$  and  $Q$  disagree), or (ii) it is split into sets of size  $k - 1$  and  $k + 1$ , flipping on each step, and the nodes of  $Q$  keep agreeing with each other and alternating between (0, 0) and (1, 1). In scenario (ii), the alternation between these 2 states is stopped by the valve mechanism, which is essentially a cycle of length 3.  $\square$

**Auxiliary labelling  $a(s)$ .** Let  $SN$  be the social network from Lemma 1 and  $s \in \{0, 1\}^n$  a configuration. Recall that  $SN$  is acyclic and all paths from source to sink in  $SN$  have equal length,  $h$ ; this means that the set of all nodes of  $SN$  can be partitioned into  $h + 1$  layers  $0, \dots, h$ , where layer 0 is the source layer and layer  $h$  the sink layer. Denote by  $SN'$  the social network obtained from  $SN$  by removing the sink layer; we now define the labelling  $a(s)$  of  $SN'$  as follows. Notice that every dual pair is contained in one layer. Consider first any labelling of  $SN'$  where the  $n$  dual pairs in layer 0 are assigned the values that represent  $s$ . The network  $SN$  converges after  $h - 1$  updates by Proposition 1. We then pick as  $a(s)$  the labelling of the limit network.

**Construction of network  $MN$  and labelling  $f$ .** We construct a social network from the components described above. Given a Turing machine  $M$ , we take the network  $SN$  from Lemma 1 and combine it with the fuse line, valve, and alarm as follows:

- For each  $i = 1, \dots, n$ , the  $i^{\text{th}}$  source dual pair of  $SN$  is identified with the  $i^{\text{th}}$  sink dual pair of  $SN$ . (This transforms  $SN$  into a cyclic network, where all cycles have length divisible by  $h$ .)
- Every dual pair in  $SN$  connects to a distinct pair in the fuse line. (As described above.)
- Every node in the alarm has edges to all dual pairs in  $SN$ , except for the fuse line and pair  $Q$  of the valve (in other words, to all dual pairs and to pair  $P$ ). (As described above.)

The fuse line needs as many pairs as there are dual pairs in  $SN$ , and  $k$  can be chosen as 2 (based on the proof of Lemma 4). This completes the construction of the network  $MN$  in our reductions.

Given a configuration  $s \in \{0, 1\}^n$  of the Turing machine  $M$ , consider any labelling of  $MN$  that satisfies the following conditions: (i) nodes in  $SN$  are labelled according to the

auxiliary labelling  $a(s)$  defined above; (ii) each pair in the fuse line and the valve is valid (i.e., its nodes disagree); (iii) the alarm is evenly split; and (iv) in every connection of the form shown in Figure 8, exactly 2 out of 4 intermediate nodes have value 0. Denote this labelling by  $f$ .

### Hardness proofs

Let us proceed to proving the computational hardness of the problems for Theorems 1 and 2.

**Proof of Theorem 1.** We already argued membership in PSPACE above and will prove hardness here. We rely on the fact that there exists a universal, polynomial-space bound Turing machine  $U$  for which the following problem is PSPACE-complete:

*Input:* an integer  $n \geq 1$  and a configuration  $s^{(0)} \in \{0, 1\}^n$  of  $U$ .

*Output:* does  $U$  diverge when started from configuration  $s^{(0)}$ ?

The complexity of this problem is shown similarly to Exercise 4.1 in (Arora and Barak 2009). See also Theorem 19.9 in (Papadimitriou 1994).

Apply the construction above to the Turing machine  $U$ . Take the network  $MN$  and the labelling  $f$  defined above. First note that dual pairs in  $SN$  have inputs from inside  $SN$  and  $2k$  inputs from the alarm. This means that  $SN$  will function “autonomously” as long as the alarm remains evenly split. By Lemma 1,  $SN$  will in this case compute consecutive configurations of the Turing machine  $U$ . There is a “pipelining” effect involved: the labelling of the source level of  $SN$  will be set to  $s^{(0)}$  at time 0, then to  $s^{(1)}$ , the successor of  $s^{(0)}$ , at times  $1, \dots, h$ , then to  $s^{(2)}$  for the next  $h$  steps, then to  $s^{(3)}$ , etc.

Observe that if the Turing machine  $U$  diverges when started from the configuration  $s^{(0)}$ , then, by the above, the alarm will always remain evenly split, flipping forever. This means that  $MN$  does not converge. On the other hand, if  $U$  terminates, then some dual pair will become invalid (Lemma 1), the alarm will go off (Lemmas 3 and 4), and the network will converge. Theorem 1 follows.

**Proof of Theorem 2.** Again, we already argued membership in PSPACE above and will prove hardness here. We will now rely on PSPACE-completeness of the following problem:

*Input:* an integer  $n \geq 1$  and (a description of) a Turing machine  $M$ .

*Output:* is there a configuration  $s \in \{0, 1\}^n$  such that  $M$  diverges when started from  $s$ ?

The hardness of this problem is a straightforward variation of the Corollary of Theorem 19.9 in (Papadimitriou 1994).

The proof of Theorem 2 extends the proof of Theorem 1. Instead of  $U$ , we now have a Turing machine  $M$ . Recall from the previous proof that if there is a configuration  $s \in \{0, 1\}^n$  from which  $M$  diverges, then there is an initial labelling from which  $MN$  fails to converge. So we will now consider the case

where  $M$  terminates started from every configuration. Can there now be a labelling from which  $MN$  fails to converge?

To answer this question, let us look into various initial labellings of  $MN$ . Let  $g$  such a labelling. By Lemma 4, if there exists a time  $t \in \{0, 1, \dots, h-1\}$  for which the network  $MN$  has an invalid dual pair, then  $MN$  converges. The same holds if  $MN$  has an invalid pair in the fuse line or valve, or if the alarm is not evenly split.

Suppose none of the above applies; then consider configurations  $s_0, \dots, s_{h-1} \in \{0, 1\}^n$  formed by the values of the source-layer dual pairs of  $SN$  at times  $0, 1, \dots, h-1$ . By the arguments above, the network  $MN$  simulates the Turing machine  $M$  in the following way. For each  $i \in \{0, 1, \dots, h-1\}$ , at times  $t \in \{i, i+h, i+2h, \dots\}$  the source-layer dual pairs of  $SN$  form consecutive configurations of  $M$  started from  $s_i$ . If  $M$  terminates when started from some  $s' \in \{s_0, \dots, s_{h-1}\}$ , then  $MN$  converges when started from the labelling  $g$ . This means that a necessary condition for  $MN$  to fail to converge (starting from  $g$ ) is that  $M$  diverges when started from every  $s_i, i \in \{0, 1, \dots, h-1\}$ . In this case, there certainly exists a configuration  $s_i$  from which the Turing machine  $M$  diverges. This completes the proof of Theorem 2.

## Conclusions

We have shown that checking convergence of opinion diffusion in social networks is PSPACE-complete. Our results extend to majority-based multi-issue opinion diffusion (Grandi, Lorini, and Perrussel 2015), also in presence of integrity constraints (Botan, Grandi, and Perrussel 2019), and to all update rules that admit suitable modification of our gadgets, such as quota rules (in which an agent switches an opinion if a specified fraction of their influencers disagrees with them).

There are many possible directions for further research. First, we have noted how some classes of networks, e.g., DAGs, are convergent and this can be verified efficiently. Our results imply that there is no efficiently computable characterisation of convergent networks, however we can ask whether a meaningful characterisation exists for networks that converge fast. Second, an interesting question is whether the existence of a *non-trivial* (i.e., different from all-0 and all-1) fixed-point configuration in our model is an NP-complete property. Third, we have limited ourselves to the study of synchronous opinion diffusion protocols. This is possibly the simplest social network update model, widely adopted in the literature. It is also of interest what happens in asynchronous networks. We note that losing synchronicity makes the system nondeterministic, so the question of convergence changes significantly. We would for example need to study different forms of convergence, e.g., for all possible update orderings, for some, and the like. Finally, our results are based on a worst-case complexity analysis and an important question remains regarding the complexity of verifying convergence in random networks.

## References

- Apt, K. R., and Markakis, E. 2014. Social networks with competing products. *Fundam. Inform.* 129(3):225–250.
- Apt, K. R.; Markakis, E.; and Simon, S. 2016. Paradoxes in social networks with multiple products. *Synthese* 193(3):663–687.
- Arora, S., and Barak, B. 2009. *Computational Complexity: A Modern Approach*. Cambridge University Press.
- Auletta, V.; Caragiannis, I.; Ferraioli, D.; Galdi, C.; and Persiano, G. 2017. Robustness in discrete preference games. In *AAMAS*, 1314–1322.
- Auletta, V.; Ferraioli, D.; and Greco, G. 2018. Reasoning about consensus when opinions diffuse through majority dynamics. In *IJCAI*, 49–55.
- Axelrod, R. M. 1997. The dissemination of culture: a model with local convergence and global polarization. *The Journal of Conflict Resolution* 203–226.
- Barrett, C. L.; Hunt, H. B.; Marathe, M. V.; Ravi, S. S.; Rosenkrantz, D. J.; and Stearns, R. E. 2003. Reachability problems for sequential dynamical systems with threshold functions. *Theor. Comput. Sci.* 295:41–64.
- Barrett, C. L.; Hunt, H. B.; Marathe, M. V.; Ravi, S. S.; Rosenkrantz, D. J.; Stearns, R. E.; and Thakur, M. 2007. Predecessor existence problems for finite discrete dynamical systems. *Theor. Comput. Sci.* 386(1-2):3–37.
- Bilò, V.; Fanelli, A.; and Moscardelli, L. 2018. Opinion formation games with dynamic social influences. *Theoretical Computer Science* 746:73–87.
- Bollobás, B. 1998. *Modern Graph Theory*. Graduate texts in mathematics. Springer.
- Botan, S.; Grandi, U.; and Perrussel, L. 2019. Multi-issue opinion diffusion under constraints. In *AAMAS*, 828–836.
- Bredereck, R., and Elkind, E. 2017. Manipulating opinion diffusion in social networks. In *IJCAI*, 894–900.
- Brill, M. 2018. Interactive democracy. In *AAMAS*, 1183–1187.
- Chacc, E. G.; Fogelman-Soulié, F.; and Pellegrin, D. 1985. Decreasing energy functions as a tool for studying threshold networks. *Discrete Applied Mathematics* 12(3):261–277.
- Christoff, Z., and Grossi, D. 2017. Stability in binary opinion diffusion. In *International Workshop on Logic, Rationality and Interaction*, 166–180.
- de Groot, M. H. 1974. Reaching a consensus. *Jour. of the Am. Stat. Assoc.* 69:118–121.
- Dryzek, J., and List, C. 2003. Social choice theory and deliberative democracy: a reconciliation. *Br. Jour. of Pol. Science* 33(1):1–28.
- Endriss, U., and Grandi, U. 2014. Binary aggregation by selection of the most representative voters. In *AAAI*, 668–674.
- Faliszewski, P., and Rothe, J. 2016. Control and bribery in voting. In *Handbook of Computational Social Choice*. Cambridge University Press. 146–168.
- Ferraioli, D.; Goldberg, P. W.; and Ventre, C. 2016. Decentralized dynamics for finite opinion games. *Theoretical Computer Science* 648:96–115.
- Goles, E.; Montealegre, P.; Salo, V.; and Törmä, I. 2016. PSPACE-completeness of majority automata networks. *Theor. Comput. Sci.* 609:118–128.
- Grabisch, M., and Rusinowska, A. 2010. A model of influence in a social network. *Theory and Decision* 69(1):69–96.
- Grandi, U.; Lorini, E.; and Perrussel, L. 2015. Propositional opinion diffusion. In *AAMAS*, 989–997.
- Grandi, U. 2017. Social choice and social networks. *Trends in Computational Social Choice*. *AI Access* 169–184.
- Granovetter, M. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83(6):1420–1443.
- Isbell, J. R. 1958. A class of simple games. *Duke Mathematical Journal* 25(3):423–439.
- Jackson, M. O. 2010. *Social and economic networks*. Princeton university press.
- Kawachi, A.; Ogihara, M.; and Uchizawa, K. 2019. Generalized predecessor existence problems for boolean finite dynamical systems on directed graphs. *Theor. Comput. Sci.* 762:25–40.
- Kempe, D.; Kleinberg, J. M.; and Tardos, É. 2005. Influential nodes in a diffusion model for social networks. In *ICALP*, volume 3580 of *LNCS*, 1127–1138.
- Klov Dahl, A. S. 1985. Social networks and the spread of infectious diseases: The AIDS example. *Social science and medicine* 21:1203–16.
- Kosub, S. 2008. Dichotomy results for fixed-point existence problems for boolean dynamical systems. *Mathematics in Computer Science* 1(3):487–505.
- Lau, D. 2006. *Function Algebras on Finite Sets*. Springer.
- Ogihara, M., and Uchizawa, K. 2017. Computational complexity studies of synchronous boolean finite dynamical systems on directed graphs. *Inf. Comput.* 256:226–236.
- Orponen, P. 1993. On the computational power of discrete Hopfield nets. In *ICALP*, volume 700 of *Lecture Notes in Computer Science*, 215–226. Springer.
- Papadimitriou, C. H. 1994. *Computational Complexity*. Addison-Wesley.
- Post, E. L. 1941. *The two-valued iterative systems of mathematical logic*. Annals of Mathematics Studies. Princeton University Press.
- Rosenkrantz, D. J.; Marathe, M. V.; Ravi, S. S.; and Stearns, R. E. 2018. Testing phase space properties of synchronous dynamical systems with nested canalizing local functions. In *AAMAS*, 1585–1594.
- Tosic, P. T. 2017. Phase transitions in possible dynamics of cellular and graph automata models of sparsely interconnected multi-agent systems. In *AAMAS*, 474–483.