Reflecting on social influence in networks

Zoé Christoff
Institute for Logic, Language and Computation
University of Amsterdam
zoe.christoff@gmail.com

Jens Ulrik Hansen
Department of Philosophy
Lund University
jensuhansen@gmail.com

Carlo Proietti
Department of Philosophy
Lund University
proietticarlo@hotmail.com

Abstract

This paper builds on the logical model of opinion dynamics under social influence in networks proposed by Liu, Seligman, and Girard (2014) as well as on the generalization proposed by Christoff and Hansen (2013). While both accounts of social influence show interesting dynamics, they both assume that agents do not reflect on how they are affected by such influence. This paper shows that, if agents are allowed to reflect upon the very rules of social influence, they may come to know (or "learn") other agents' private opinions, even though they can only observe their public behavior. To represent formally agents who are able to reason about social influence, a logic of social networks, knowledge, influence, and "learning" is introduced.

The way opinions, preferences or behaviors vary among agents situated in social networks has been studied by numerous authors within networks analysis (see for instance (Jackson 2010), (Easley and Kleinberg 2010) or (Acemoglu and Ozdaglar 2011)). Fairly recently, similar concerns have arisen within the field of logic (Seligman, Liu, and Girard 2011; Zhen and Seligman 2011). (Liu, Seligman, and Girard 2014) proposes a simple model of how agents change their opinions (or beliefs) as a result of social influence seen as social conformity pressure: agents tend to align their opinions with the opinions of agents they are related to within their social network. This framework is generalized in (Christoff and Hansen 2013) to allow for modeling of phenomena where the observable behavior of a group of agents does not reflect their private opinion. Such counter-intuitive group behavior is well-documented in social sciences, as exemplified by studies of the particular case of *pluralistic ignorance* (see for instance (O'Gorman 1986; Prentice and Miller 1993).

Consider as an example a situation where a group of agents in a network structure all enforce a norm, for instance a segregation norm. According to (Liu, Seligman, and Girard 2014), this would be interpreted as follows: all agents agree with the norm (even though they might come to agree with it only as a result of social conformity pressure). The extended framework of (Christoff and Hansen 2013) allows this situation to be analyzed in a slightly more complex way: even though agents are aligning their behavior (that is, enforcing the norm), they might still privately disagree with it, they might just face too much social pressure to dare expressing publicly a discarding opinion.

While both these frameworks allow for interesting dynamic examples of social influence, they represent fairly simpleminded agents. Indeed, according to both accounts, the agents are simple "rule-followers": they do not reflect on the fact that they, as well as the others, are such "rule-followers". However, it seems quite natural to assume that agents know the effect of social influence and can reason about it. If this is the case, even though everybody enforces a norm, somebody might very well come to know that somebody else privately disagrees with the norm, by reflecting on the dynamics of her public behavior and the one of others around her in the social network structure. This is what we want to model in this paper.

In short, while (Liu, Seligman, and Girard 2014) assumes fully "transparent" agents, whose opinions are always known to agents around them, (Christoff and Hansen 2013) assumes fully "opaque" agents, whose (private) opinions are never known to anybody else. What we want to model here is how agents might "see through" each other's behavior and come to know other's private opinions by observing their public behavior. To formally represent such less simpleminded agents, capable of deducing the private opinions of others from their public behavior, we introduce a logic of social networks, knowledge, influence, and "learning".

The paper is structured as follows. Section 1 recalls the frameworks of (Liu, Seligman, and Girard 2014) and (Christoff and Hansen 2013) and the corresponding notions of social influence. Section 2 discusses what agents can learn if they reflect on the fact that others are following the rules of social influence. Section 3 then introduces a logic of social networks, knowledge, influence, and learning to capture the issues raised in Section 2. In Section 4, these issues are formalized in the new framework. Finally, we explore possibilities for future work and briefly discuss how granting our agents with the power to reason about the dynamics of social influence might modify this very dynamics.

1 Simple frameworks of opinion dynamics under social influence

1.1 The "Facebook logic" model

(Liu, Seligman, and Girard 2014) investigates the way the beliefs, or opinions, of agents in a social network change

under social influence and is a variation of the "Facebook Logic" introduced in the programmatic paper "Logic in the Community" (Seligman, Liu, and Girard 2011). Let us briefly recall their framework.

At any given moment, each agent is in one of three (mutually exclusive) states of opinion relatively to an implicit fixed object or proposition: 1) she believes it (we will say that she has a "pro opinion"), or 2) she believes the negation of it (she has a "contra opinion"), or 3) she is without opinion about it (she has a "neutral opinion"). For instance, an agent could be pro a new law, opposing the law, or neutral towards it. While this is a fairly simplistic model of opinions it is extremely rich when combined with other factors such as social network structure and pressure of social influence. Formally, this view on opinions is captured by the fact that each agent satisfies exactly one of three atomic propositions.

The agents are situated in a social network represented by a finite set of nodes (-the agents) with an irreflexive and symmetric binary relation between them (like the relation of being friends or neighbors). The semantics is indexical in the sense that each proposition is evaluated from the perspective of some agent. In other words, propositions express properties of agents. Clearly, a modal logic is ideal to reason about such structures, where the modal operator F, quantifying over network-neighbors, reads "(at) all of *my* friends/neighbors" and its dual $\langle F \rangle$ reads "(at) some of *my* friends/neighbors". Additional hybrid logic machinery (Areces and ten Cate 2007) is included in the framework, namely *nominals* (a special kind of propositional variables used to refer to agents) and satisfaction operators $@_i$ (used to switch the evaluation point to the unique agent satisfying the nominal "i").

A social influence operator is then applied to this static model to represent how opinion repartition in the social network changes. The dynamics relies on the general social conformity pressure assumption according to which agents tend to become more similar to the agents they are directly related to within the social network, i.e they tend to align their opinions to the ones of their network-neighbors.[1] The dynamics is locally and uniformly defined by describing the possible changes induced by the situations of "strong influence" and "weak influence":

An agent a is *strongly influenced* when *all* of her network-neighbors have the same (that is, pro or contra) opinion and she has some neighbor. In this case, whatever a's initial state is, she will adopt this very opinion. An agent a is *weakly influenced* when, among her network-neighbors, *some* have an opinion while none have the opposite opinion. In this situation, if a is without opinion or if she already has the same opinion that some neighbors have, her state does not change. However, if a has the opposite opinion, she will drop it and become without opinion. Note that weak influence is the only social pressure situation which can lead to *abandon* of an opinion, while strong influence is the only situation which can lead to *adoption* of an opinion.

[1]The same assumption is made about *knowledge* change in (Seligman, Liu, and Girard 2011) and about *preference* change in (Zhen and Seligman 2011).

According to these model transformation rules, some configurations of opinions within a community will never stabilize. Consider this simple example: if agents a and b are each other's only friend, and a has a pro opinion while b has a contra opinion: a and b will switch opinions after updating the model. But then, by the same dynamics, they will switch back, and in fact they will keep switching back and forth. Generalizing, if for each agent (who has at least one network-neighbor), all of her neighbors have an opinion and all of her neighbors' neighbors have the opposite one, they will keep changing opinions, stuck in a "looping" dynamics. However, all other configurations will stabilize after some iterated influence changes: for instance, any model such that two neighbors have the same opinion or such that an agent has no opinion will stabilize. Therefore, an advantage of this very simple dynamics is that the sufficient and necessary conditions of *stability* can be expressed in the logic. They correspond to the negation of the preconditions of the changes induced by the situations of strong and weak influence.[2] Moreover, configurations which are *becoming stable* can be characterized too, by simply ruling out the "forever looping scenario" mentioned above.

However, this simplicity relies on the following strong assumption: the agents' opinion is influenced directly by her network-neighbors' opinions, suggesting that agents always know their neighbors' opinions. This *transparency* assumption rules out the modeling of situations where agents act in ways which do *not* reflect their actual mental states. However, as argued in (Christoff and Hansen 2013), the possibility of this very discrepancy is an essential component of some social phenomena.[3] For this reason, (Christoff and Hansen 2013) introduces a "2-layer" model of social influence, with a built-in distinction between an agent's private opinion and the opinion she expresses.

1.2 The "2-layer" model

(Christoff and Hansen 2013) considers an example of a social psychology phenomenon which relies on a discrepancy between the private opinion of agents and their public behavior: *pluralistic ignorance* (O'Gorman 1986). Briefly put, pluralistic ignorance can be seen as "[...] a psychological state characterized by the belief that one's private attitudes and judgments are different from those of others, even though one's public behavior is identical." (Prentice and Miller 1993, p. 244). This type of phenomena cannot be captured within the framework of (Liu, Seligman, and Girard 2014), precisely because the above-mentioned "transparency assumption" rules out any mistake in interpreting the behavior of others. To overcome this limitation and be able to model the dynamics of such complex social phenomena, (Christoff and Hansen 2013) enriches the framework of (Liu, Seligman, and Girard 2014) with a "two-layer" notion of opinion and social influence, distinguishing an agent's *private* opinion from what she *publicly* expresses – her visible behavior.

To capture the social network, (Christoff and Hansen

[2]For details, see (Liu, Seligman, and Girard 2014)[p. 8].

[3]See for instance (Schelling 1978).

	Inner state	$\langle F\rangle ep$	$\langle F\rangle ec$	$\langle F\rangle en$	Update
1	ip				$\rightsquigarrow ep$
2	ic	1	1	1	$\rightsquigarrow ec$
3	in				$\rightsquigarrow en$
4	ip				$\rightsquigarrow ep$
5	ic	1	1	0	$\rightsquigarrow ec$
6	in				$\rightsquigarrow en$
7	ip				$\rightsquigarrow ep$
8	ic	1	0	1	$\rightsquigarrow en$
9	in				$\rightsquigarrow ep$
10	ip				
11	ic	1	0	0	$\rightsquigarrow ep$
12	in				
13	ip				$\rightsquigarrow en$
14	ic	0	1	1	$\rightsquigarrow ec$
15	in				$\rightsquigarrow ec$
16	ip				
17	ic	0	1	0	$\rightsquigarrow ec$
18	in				
19	ip				$\rightsquigarrow ep$
20	ic	0	0	1	$\rightsquigarrow ec$
21	in				$\rightsquigarrow en$
22	ip				$\rightsquigarrow ep$
23	ic	0	0	0	$\rightsquigarrow ec$
24	in				$\rightsquigarrow en$

Figure 1: The rules of social influence $\mathcal{I}$

2013) uses the same hybrid logic tools as (Liu, Seligman, and Girard 2014). The opinion *state* of a community is then modeled in a similar way, the difference being that each agent's state is now a combination of two ingredients: her "inner opinion" and her "expressed opinion" (still on an implicit fixed object). Her private state is either "inner pro opinion" (which we will denote ip), "inner contra opinion" (ic), or "inner neutral opinion" (in), while her public state is either "expressed pro opinion" (ep), "expressed contra opinion" (ec) or "expressed neutral opinion" (en).

The social influence dynamic operator now has to determine how to transform these two-layer models. The dynamics therefore relies on a slightly different assumption: an agent's *expressed* opinion tends to become more similar to the *expressed* opinion of agents around her, but also depends on her own *inner* opinion. The output state (the opinion expressed in the next step) therefore depends on an asymmetrical local input: the expressed opinion of related agents and her own private opinion. Moreover, it is also assumed that agents tend to be "sincere" by default, i.e, they express the opinion corresponding to their actual inner opinion *whenever the social pressure does not prescribe otherwise*.[4] Finally, it is assumed that an agent's inner state of opinion never changes.

The social pressure situations of strong and weak influence from (Liu, Seligman, and Girard 2014) are adapted accordingly: First, when all of her neighbors *express* the same (pro or contra) opinion, an agent will *express* this very opinion at the next moment, whatever her initial (private and expressed) state was. Similarly, when some of her neighbors *express* an opinion while none expresses the opposite opinion, the agent will *express* the supported opinion if she is *privately* neutral or if she *privately* has the same opinion. And if the agent *privately* has the opposite opinion, then she will *express* neutrality. In all other cases, the agent will be sincere: her expressed opinion state will simply reflect her own private opinion state. To put it another way, the type of agents defined by this dynamics is brave enough to express her actual private opinion whenever 1) she faces no expressed opposition to it from her neighbors *or* 2) she sees some expressed support for it from her neighbors. Figure 1 gives a full detailed presentation of these update rules. The "Inner state" column represents an agent's private opinion, while the three subsequent columns provide all the eight possible combinations of the expressed opinions of her friends, e.g. for the rows 4 to 6 an agent has some friend expressing a pro opinion, some friend expressing a contra opinion but no friend expressing a neutral opinion. (Thus 1 stands for true while 0 stands for false.) The "Update" column represents how the agent's expressed opinion is updated at the next step, e.g. row 1 should be interpreted as follows: if an agent has a private pro opinion and has some friends expressing a pro opinion, some expressing a contra opinion and some being neutral then she will express a pro opinion at the next step.

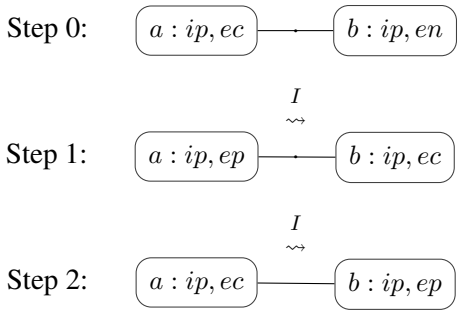

Figure 2: The two agent network of Example 1 evolved two steps beyond the original state

Now, consider a simple case of social influence illustrated by Figure 2:

Example 1 *Two agents a and b (the two nodes in Figure 2) are each other's only friend. This is represented by the edge between the two nodes in Figure 2. The initial situation is as follows: both a and b have an inner pro opinion (ip) but while a expresses a contra opinion (ec), b expresses a neutral opinion (en). This initial state of the network is represented by* Step 0 *of Figure 2. The rules of influence given by Figure 1 prescribe that, at the next moment (*Step 1*), agent a will express a (sincere) pro opinion (ep) as specified by row 19 of Figure 1 and agent b will express an (insincere) contra opinion (ec) according to row 16 of Figure 1. From*

[4]There are different ways to define such a dynamics depending on how influenciable agents are assumed to be. (Christoff and Hansen 2013) discusses three possible definitions corresponding to three different types of agents. For simplicity, we here consider only one type of agents, namely their "type 1" agents.

*then on, both agents are in a situation of maximal pressure (strong influence) and therefore, at the next moment (*Step 2*), each of them will express whatever the other expresses now (at* Step 1*). The two agents will then keep oscillating forever between* ep *and* ec*, as prescribed by rows 10 and 16 of Figure 1. Thus a non-stabilizing situation is reached.*

In this example, even though their inner opinions agree and do not change, agents oscillate between being sincere and being insincere, by trying to align their behavior (expressed opinion) with the one of their network-neighbors. According to this account of social influence, an agent's behavior is determined *only* by her own private opinion and the expressed opinion of her neighbors, without taking into account her neighbors' private opinion, which she has no access to, by the above mentioned "opacity" assumption. However, the next section will show that it is still sometimes possible to "see through" agents, that is, to infer their private opinions by observing their expressed opinion and reasoning about the social influence rule itself.

2 Reflecting on social influence

In (Christoff and Hansen 2013) the opinion expressed by an agent is a function of her private opinion and the opinion expressed by her neighbors. This type of dependency is a common way of defining rules of social influence and has been used as a key to explain many patterns of diffusion of a behavior/opinion in social networks (see (Easley and Kleinberg 2010)) including pluralistic ignorance (see (Centola, Willer, and Macy 2005)). Despite its elegance, this explanatory mechanism may sometimes be too simplistic, because it only applies to cases where agents blindly obey to social influence without considering why their peers behave in a given way. Many real life examples show that this is not always true: if agents have a clue about the actual private opinions of their peers they may respond in a different way to the very same aggregate social pressure.

An example based on everyday life is the following. Suppose that R is a reader of a financial magazine where news writer N writes and whose editor is E. Suppose that E owns some assets of a big company C and has plans for a takeover. Assume further that R knows that E is the editor of N and that E owns C's assets. One day R reads a report, authored by N, concerning C being in a critical situation. R is then likely to reason as follows: N is subject to the influence of E. Whatever N writes about C might be biased by E's interest in it. Therefore N is not an independent witness and whatever he writes about C may not reflect N's real private opinion. Summing up, by knowing how social influence works and how other agents are networked, R should reasonably disregard N's expressed opinion (especially if he himself owns assets of C). To reason like this R must know that there is a link between N and E. Moreover, he must know that N is strongly influenced by E in a way that determines his behavior (N is a big crawler indeed). When this kind of information is available, individuals are able to reason in this way. Conversely, when they are not able to do that it is very often precisely because they are ignorant about the network structure or the influence mechanisms.

Real-life dynamics, as the one just illustrated, provide a vivid idea of the problems at stake, but are often complicated to represent. For our purposes we will instead stick to our Example 1 and build upon it by modeling the uncertainty of agents as in standard epistemic logics. Indeed, while Step 0 of Figure 2 provides a compact diagrammatic representation of the initial state of our network, it does not capture the agents' initial uncertainty. As we just stressed, uncertainty in these scenarios may come in two forms. Both a and b may (i) not have access to each other's opinion or (ii) may be uncertain about the network structure, i.e. whether or not their friends have additional friends, how many of them etc. For the sake of simplicity let us assume that uncertainty of kind (ii) is settled, viz. both a and b know that they are each other's only friend. We also assume that both a and b have no uncertainty about each other's expressed opinion: a can tell apart $@_b en$-states, $@_b ec$-states and $@_b ep$-states and b can do the same w.r.t. $@_a en$-states, $@_a ec$-states and $@_a ep$-states. Still, for all a knows, b's inner state could be in, ip or ic and the same for b w.r.t. a. This state of uncertainty is represented by Step 0 of Figure 3. Here nodes represent possible states of the network (not agents!). The actual state is highlighted by a thicker frame. A dashed line with superscript a (resp. b) represents epistemic accessibility between states for agent a (resp. b). Accessibility relations are assumed to be S5, i.e. symmetric, reflexive and transitive.[5] For simplicity we omit drawing reflexive and transitive edges between nodes.

The result of applying the social influence rules of Figure 1 to each possible state (we call this operation $\mathcal{I}$) is represented by Step 1 of Figure 3. Since our rules of social influence don't take into account the agents' uncertainty, the accessibility relations are left untouched. But now the actual state is an ep-state for a, while the state on its left, which is b-accessible, is an ec-state for a. If we assume, as we did, that agents know each other's expressed opinion, b should be able to tell the two states apart. Indeed, both a and b should be able to differentiate those states where their neighbor expresses different opinions. Satisfying this constraint is what operation $\mathtt{L}$ (for "learning") does, as a transition to Step 1' represented by a simple removal of arrows (as in (Aucher et al. 2009)). As a result, in the actual state, b discards those states where a is in ic and in and therefore knows a's private opinion. Unfortunately for a, he cannot succeed in doing the same: this is witnessed by the fact that some a-arrows are kept which link states where b has different inner opinions.[6] The epistemic situation will not change for a by repeatedly applying $\mathcal{I}$ (Step 2) and $\mathtt{L}$ at successive steps and a will never know what b's private opinion is.

Why are the epistemic situations of a and b so different? The explanation is that, at every step, an agent's expressed opinion is the value of a function $f(x, l)$, where x is her pri-

[5] Assuming that knowledge is S5 for real-life agents is a strong idealization that has been widely debated. We do not want to enter this discussion here, unless for stressing that S5 can be adequate for modeling agents in simple toy scenarios like the one we present.

[6] The explanation of this fact is the following. Given a's previously expressed opinion, b would now express a contra opinion ec whatever his private opinion was at the previous state (rows 16 to 18 of Figure 1 have the same outcome).

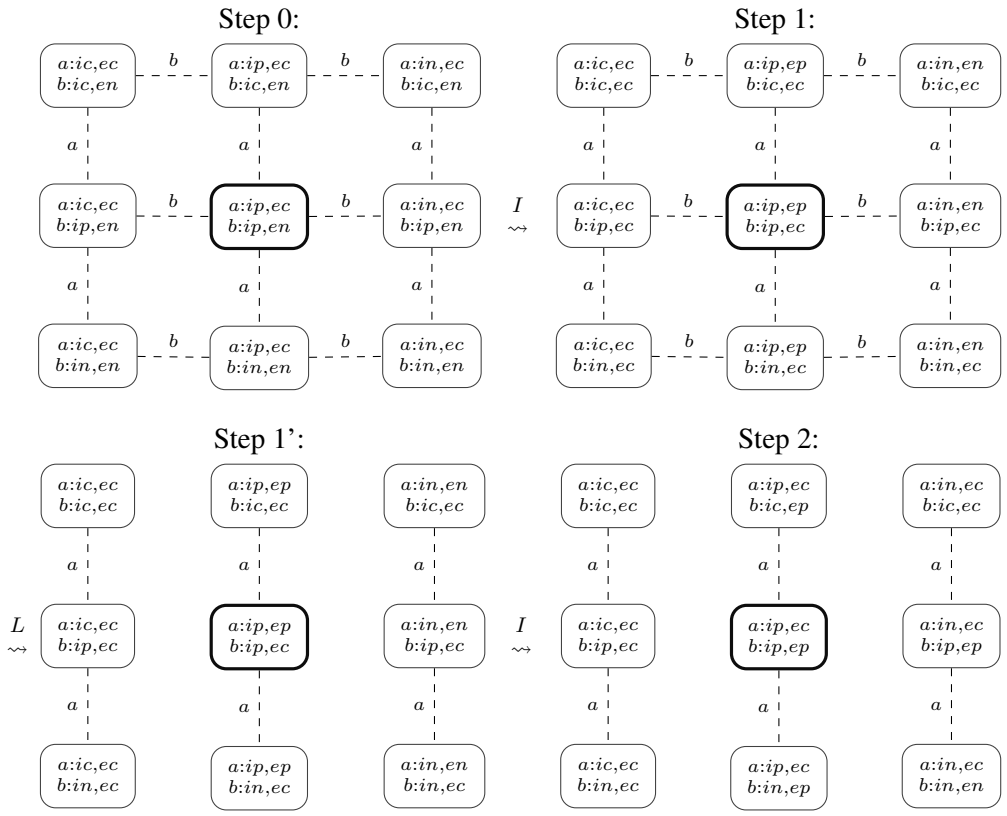

Figure 3: The two agent network of Example 1 with uncertainty represented. From Step 0 to Step 1 social influence happens, from Step 1 to Step 1' the agents learn, while from Step 1' to Step 2 social influence happens again.

vate opinion (ip, ic or in) and l is the expressed opinion of her friends (the eight possible combinations of truth values of $\langle F\rangle ep$, $\langle F\rangle ec$ and $\langle F\rangle en$). If someone knows l and the value of $f(x,l)$ (we are assuming that they know both, since they have access to each other's expressed opinions) then they can retrieve x only when $f(x,l)$ is injective for a given value of l. This is not the case for four possible values of l (corresponding to rows 7 to 9, 10 to 12,13 to 15 and 16 to 18). In case $f(x,l)$ is not injective but takes at least two possible values (rows 7 to 9 and 13 to 15) it is still possible to discard some possible values of x and specify one's uncertainty.

To summarize what we have seen this far, we can model reflective agents in a social influence framework by alternating influence updates $\mathcal{I}$ with an epistemic update $\mathtt{L}$. In the next section we shall develop a general logic for this.

3 A logic of Social networks, Knowledge, Influence, and Learning

In this section we introduce a formal logical framework to reason about how reflecting agents can infer the private states of others, as described in the previous section. We will call this *the Logic of Social networks, Knowledge, Influence, and Learning*. Our logic will mix several ideas and frameworks. First of all, we will take the two-dimensional "Facebook" logic of (Seligman, Liu, and Girard 2011), as discussed in Section 1, as our starting point. To this logic, we will then add the modality for social influence from (Christoff and Hansen 2013) – also mentioned in Section 1. Finally, we will add learning modalities inspired by the graph modifiers of (Aucher et al. 2009) and the Arrow Update Logic of (Kooi and Renne 2011).

First, we fix a countable set of nominals NOM, which will be used to name agents in the social networks. As atomic propositions we will use the set $\{ip, ic, in, ep, ec, en\}$. Then, the simple language of the "Facebook" logic to talk about social networks and knowledge is defined by:

Definition 2 (Syntax for $\mathcal{L}_{\mathcal{SK}}$) *The syntax of the language $\mathcal{L}_{\mathcal{SK}}$, is given by:*

$$\varphi \ ::= \ p \mid i \mid \neg\varphi \mid \varphi \wedge \varphi \mid F\varphi \mid @_i\varphi \mid K\varphi \ ,$$

where $p \in \{ip, ic, in, ep, ec, en\}$ and $i \in \mathsf{NOM}$.

We will use the standard abbreviations for $\vee$, $\rightarrow$, and $\leftrightarrow$. Moreover we will denote the dual of F by $\langle F\rangle$ and the dual of K by $\langle K\rangle$, in other words $\langle F\rangle\varphi := \neg F\neg\varphi$ and $\langle K\rangle\varphi := \neg K\neg\varphi$.

As previously mentioned, the semantics for $\mathcal{L}_{\mathcal{SK}}$ will be a two-dimensional one where the first dimension represents the agents and the network structure, while the second one represents possible worlds over which the knowledge modality K quantifies over. The nominals i and the opera-

tors $@_i$, and F, are used to talk about the network dimension. The definition of a model is the following:

Definition 3 (Model) *A* model *is a tuple* $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$*, where:*

- A *is a non-empty set of agents,*
- W *is a non-empty set of possible worlds,*
- $\succeq_w$ *is an irreflexive, symmetric binary relation on* A*, for each* $w \in W$ *(representing the network structure at the world* w*),*
- $\sim_a$ *is an equivalence relation on* W *for each* $a \in A$ *(representing the uncertainty of* a*),*
- $g : \mathsf{NOM} \to A$ *is a function assigning an agent to each nominal,*
- $V : \{ip, ic, in, ep, ec, en\} \to \mathcal{P}(A \times W)$ *is such that*
  - $V(ip)$*,* $V(ic)$*, and* $V(in)$ *are pairwise disjoint*
  - $V(ip) \cup V(ic) \cup V(in) = A \times W$
  - $V(ep)$*,* $V(ec)$*, and* $V(en)$ *are pairwise disjoint*
  - $V(ep) \cup V(ec) \cup V(en) = A \times W$

  (That is, V *assigns exactly one inner and exactly one expressed opinion to each agent at each world.)*

We can now give the semantics for the language $\mathcal{L}_{\mathcal{SK}}$:

Definition 4 (Semantics of $\mathcal{L}_{\mathcal{SK}}$) *Given a model* $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$*, an* $a \in A$*, a* $w \in W$*, and a formula* $\varphi \in \mathcal{L}_{\mathcal{KI}}$*, we define the truth of* φ *at* (w, a) *in* $\mathcal{M}$ *inductively by:*

$$
\begin{array}{lll}
\mathcal{M}, w, a \models p & \textit{iff} & (w, a) \in V(p) \\
\mathcal{M}, w, a \models i & \textit{iff} & g(i) = a \\
\mathcal{M}, w, a \models \neg\varphi & \textit{iff} & \mathcal{M}, w, a \not\models \varphi \\
\mathcal{M}, w, a \models \varphi \wedge \psi & \textit{iff} & \mathcal{M}, w, a \models \varphi \textit{ and } \mathcal{M}, w, a \models \psi \\
\mathcal{M}, w, a \models F\varphi & \textit{iff} & \forall b \in A; a \succeq_w b \Rightarrow \mathcal{M}, w, b \models \varphi \\
\mathcal{M}, w, a \models @_i\varphi & \textit{iff} & \mathcal{M}, w, g(i) \models \varphi \\
\mathcal{M}, w, a \models K\varphi & \textit{iff} & \forall v \in W; w \sim_a v \Rightarrow \mathcal{M}, v, a \models \varphi,
\end{array}
$$

where $p \in \{ip, ic, in, ep, ec, en\}$*. We say that a formula* φ *is true in a model* $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$ *if* $\mathcal{M}, w, a \models \varphi$ *for all* $w \in W$ *and all* $a \in A$*. We denote this by* $\mathcal{M} \models \varphi$*. We say that a formula* φ *is valid if* $\mathcal{M} \models \varphi$ *for all models* $\mathcal{M}$ *and denote this by* $\models \varphi$*. The resulting logic will be referred to as the logic of social networks and knowledge and will be denoted* SK*. Finally, we say that two formulas* φ_1 *and* φ_2 *are pairwise unsatisfiable if for all models* $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$*, all worlds* $w \in W$*, and all agents* $a \in A$*;* $\mathcal{M}, w, a \not\models \varphi_1 \wedge \varphi_2$*.*

Note that since we required the relations $\sim_a$ to be equivalence relations, the logic of the knowledge modality K will be standard S5 in accordance with the discussion of the previous section.

To this logic we will now add the social influence modality $[\mathcal{I}]$ from (Christoff and Hansen 2013). The language is given by:

Definition 5 (Syntax for $\mathcal{L}_{\mathcal{SKI}}$) *The syntax of the language of social networks, knowledge, and influence is as in Definition 2 with the additional clause that whenever* $\varphi \in \mathcal{L}_{\mathcal{SKI}}$*, then also* $[\mathcal{I}]\varphi \in \mathcal{L}_{\mathcal{SKI}}$*.*

To give the semantics for $\mathcal{L}_{\mathcal{SKI}}$, we need to define how the influence rules of $\mathcal{I}$ change models.

Definition 6 (Influence update $\mathcal{M}^{\mathcal{I}}$) *Given a model* $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$*, the influence updated model* $\mathcal{M}^{\mathcal{I}}$ *is* $(A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V')$*, where* V' *is equal to* V *on* $\{ic, ip, in\}$ *and for* $\{ec, ep, en\}$ *is specified by the table of Figure 1.*[7]

The semantics of $\mathcal{L}_{\mathcal{SKI}}$ is given by:

Definition 7 (Semantics of $\mathcal{L}_{\mathcal{SKI}}$) *Given a model* $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$*, an* $a \in A$*, a* $w \in W$*, and a formula* $\varphi \in \mathcal{L}_{\mathcal{SKI}}$*, we define the truth of* φ *at* (w, a) *in* $\mathcal{M}$ *as in Definition 4 with the following additional clause:*

$$\mathcal{M}, w, a \models [\mathcal{I}]\varphi \quad \textit{iff} \quad \mathcal{M}^{\mathcal{I}}, w, a \models \varphi\,.$$

Notions like true in a model, valid, and pairwise unsatisfiable are defined as in Definition 4. The resulting logic of social networks, knowledge and influence will be denoted SKI*.*

While the just introduced logic allows us to reason about how agents change their expressed opinions, it still does not allow us to reason about what agents may learn about the private opinions of others from observing their expressed opinions – i.e. we cannot capture the move from Step 1 to Step 1' in Figure 3. We therefore extend our language with what we call *learning modalities* $[\mathtt{L}]$. In fact, to be able to discuss different learning abilities depending on what agents can observe in the network, we add a general class of learning modalities. Note that, by "learning" we are only referring to what agents can infer about other agents' inner opinions based on their expressed opinions. Thus, agents cannot learn any truths about the world that might make their inner opinions more accurate. In this sense, it is a restricted form of learning we introduce here.

Inspired by (Aucher et al. 2009) and (Kooi and Renne 2011), we will use 4-tuples $(i, \varphi_1, \varphi_2, j)$, where φ_1 and φ_2 are formulas and i and j are nominals, to specify how the equivalence relations $\sim_{g(i)}$ changes – i.e., how agent i can learn based on observing whether j satisfies φ_1 or φ_2. However, contrary to (Kooi and Renne 2011), we will use our 4-tuples to specify which relationships to delete instead of which to keep and differently from (Aucher et al. 2009) we will include multiple 4-tuples in each modality. Moreover, contrary to both frameworks, we will also keep track of the agent that is observed (the agent named by j).

Definition 8 (Learning modalities $\mathtt{L}$) *A learning modality* $\mathtt{L}$ *is a finite set of tuples of the form* $(i, \varphi_1, \varphi_2, j)$*, where* i *and* j *are nominals and* φ_1 *and* φ_2 *are formulas of* $\mathcal{L}_{\mathcal{SKI}}$*, such that* φ_1 *and* φ_2 *are pairwise unsatisfiable.*

The language with the new learning modalities is defined as follows:

[7] For instance, $(a, w) \in V'(ic)$ iff $(a, w) \in V(ic)$, and if $\mathcal{M}, w, a \models ip \wedge \neg\langle F\rangle ep \wedge \langle F\rangle ec \wedge \langle F\rangle en$, then $(a, w) \in V'(en)$ according to row 13 of Figure 1.

Definition 9 (Syntax for $\mathcal{L}_{\mathcal{SKIL}}$) *The syntax of the language of social networks, knowledge, influence, and learning is as in Definition 5 with the additional clause that whenever $\varphi \in \mathcal{L}_{\mathcal{SKIL}}$, then also $[\mathtt{L}]\varphi \in \mathcal{L}_{\mathcal{SKIL}}$ for any learning modality* $\mathtt{L}$.

To give the semantics of this new language we first need to specify how learning modalities change models:

Definition 10 (Learning updates $\mathcal{M}^{\mathtt{L}}$) *Given a model $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$ and a learning modality* $\mathtt{L}$*, the learning updated model $\mathcal{M}^{\mathtt{L}}$ is $(A, W, (\succeq_w)_{w\in W}, (\sim'_a)_{a\in A}, g, V)$, where $\sim'_a$ is defined by:*

$$w \sim'_a v \quad \textit{iff} \quad w \sim_a v \textit{ and, there is no } (i, \varphi_1, \varphi_2, j) \in \mathtt{L} \textit{ such that; } g(i) = a, \mathcal{M}, w, a \models @_j\varphi_1, \textit{ and } \mathcal{M}, v, a \models @_j\varphi_2 \, .$$

The intuition behind this definition is the following: If $(i, \varphi_1, \varphi_2, j) \in \mathtt{L}$ this means that the agent named by i can learn based on whether φ_1 or φ_2 is true of the agent named by j. Or put differently, the agent named by i can observe whether φ_1 or φ_2 is true of agent j.

Now we can give the semantics for $\mathcal{L}_{\mathcal{SKIL}}$:

Definition 11 (Semantics of $\mathcal{L}_{\mathcal{SKIL}}$) *Given a model $\mathcal{M} = (A, W, (\succeq_w)_{w\in W}, (\sim_a)_{a\in A}, g, V)$, an $a \in A$, a $w \in W$, and a formula $\varphi \in \mathcal{L}_{\mathcal{KIL}}$, we define the truth of φ at (w, a) in $\mathcal{M}$ as in Definition 7 with the following additional clause:*

$$\mathcal{M}, w, a \models [\mathtt{L}]\varphi \quad \textit{iff} \quad \mathcal{M}^{\mathtt{L}}, w, a \models \varphi \, .$$

Notions like true in a model, valid, and pairwise unsatisfiable are defined as in Definition 4. The resulting logic will be denoted SKIL.

4 Modeling influence and learning

In this section we will formalize the reasoning that reflecting agents can perform, as discussed in Section 2. Let us first formalize the learning that agent b does in Step 1' of the example of Figure 3. Here b learns because he can see a's expressed opinion. In other words, we want to cut all accessibility links between any two worlds where a satisfies two different propositions from the set $\{ep, ec, en\}$. This is easy to capture by a learning modality $\mathtt{L}_1$, defined by

$$\mathtt{L}_1 := \{(i, ep, ec, j), (i, ep, en, j), (i, ec, en, j) \mid i, j \in N\},$$

where N is a finite set of nominals. Assume that N contains the nominals i and j and let $\mathcal{M}$ be the model of Step 1 of Figure 3 with $a \succeq_w b$ for all worlds w, $g(i) = a$, and $g(j) = b$. Then it is not hard to see that the model of Step 1' of Figure 3 is $\mathcal{M}^{\mathtt{L}_1}$. For any world w satisfying ep of a, i.e. $\mathcal{M}, w, b \models @_i ep$, and any world w' satisfying ec of a, i.e. $\mathcal{M}, w, b \models @_i ec$, since (j, ep, ec, i) is in $\mathtt{L}_1$ (and $g(j) = b$ and $g(i) = a$), we can infer from Definition 10 that $w \not\sim'_b w'$.

Now, consider another example with an extra agent c, where for all worlds w, $a \succeq_w c$, $a \succeq_w b$, but $c \not\succeq_w b$. With the learning modality $\mathtt{L}_1$, b can potentially also learn about c's expressed opinion (if there is a nominal $k \in N$ such that $g(k) = c$). This might be the type of learning we want in some examples, but often the entire point of having an explicit network structure is that what agents can observe is limited, i.e., they can only observe the expressed opinions of their network neighbors. This can easily be fixed by replacing the learning modality $\mathtt{L}_1$ by the learning modality $\mathtt{L}_2$ defined by:

$$\begin{aligned}\mathtt{L}_2 := \{&(i, ep \wedge \langle F\rangle i, ec \wedge \langle F\rangle i, j),\\ &(i, ep \wedge \langle F\rangle i, en \wedge \langle F\rangle i, j),\\ &(i, ec \wedge \langle F\rangle i, en \wedge \langle F\rangle i, j) \mid i, j \in N\} \, .\end{aligned}$$

With this learning modality, neither $ep \wedge \langle F\rangle j$, $ec \wedge \langle F\rangle j$, nor $en \wedge \langle F\rangle j$ is satisfied of c and thus, b will not be able to learn based on c's expressed opinion. In general, any formula like $ep \wedge \langle F\rangle i$ can only potentially be true of an agent that is friends with the agent named by i, hence, in this way i can only observe the expressed opinions of her friends.

Note that in this example b might indirectly learn about c. Even if he has no clue about c's expressed opinion b may still infer that, at Step 0, c expressed either ep or en because, had c expressed ec then a would not have expressed ep whatever his inner state was (according to rows 13 to 15 of Figure 1).

We could also consider agents with a larger radius of observability, for instance we might assume that i can observe the expressed behavior of her friends as well as the friends of her friends. This can be captured by extending the learning modality $\mathtt{L}_2$ to include the following set of tuples:

$$\begin{aligned}\{\ &(i, ep \wedge \langle F\rangle\langle F\rangle i, ec \wedge \langle F\rangle\langle F\rangle i, j),\\ &(i, ep \wedge \langle F\rangle\langle F\rangle i, en \wedge \langle F\rangle\langle F\rangle i, j),\\ &(i, ec \wedge \langle F\rangle\langle F\rangle i, en \wedge \langle F\rangle\langle F\rangle i, j)\ \}.\end{aligned}$$

So far we have assumed that all agents obey the influence rules specified by $[\mathcal{I}]$ and discussed how they can infer other agents' inner opinions based on observation of their expressed opinions. However, one could easily imagine that the pressure to express a pro opinion, when all of one's friends express a pro opinion, will be much less or even nonexistent if one knows that one's friends' inner opinions are actually contra. For example, after Step 1' of Figure 3, b may want to consider the situation further. According to what is prescribed by the influence rules, he will continue oscillating between expressing a pro and a contra opinion although being aware that a has an inner pro opinion and (half of the times) is expressing a contra opinion just in virtue of b's influence. This situation may be strategically convenient to b and he may want to keep his vantage point over a by oscillating according to the rules. However, b may also decide not to follow social influence anymore and to adapt his expressed opinion to his inner opinion: playing a social role has a cost and b may want to give up the masquerade.[8] If he does so a will then be faced with a strange situation: b is now breaking the social influence rules. At this point a may simply adapt to b's new trend and he will also stabilize on expressing a

[8] There there several theories within social psychology that will predict that agents will try and avoid any inconsistency between their private opinions and their behavior, for instance the theory of cognitive dissonance (Festinger 1957). Thus, it seems natural in many real life cases to assume that agent b will indeed give up the masquerade.

	Inner state	Knowledge state	$\langle F\rangle ep$	$\langle F\rangle ec$	$\langle F\rangle en$	Update
1	ip	$K\langle F\rangle ip \vee K\neg\langle F\rangle ic$	—	—	—	$\rightsquigarrow ep$
2	ic	$K\langle F\rangle ic \vee K\neg\langle F\rangle ip$	—	—	—	$\rightsquigarrow ec$
3	ip	$\neg(K\langle F\rangle ip \vee K\neg\langle F\rangle ic)$	see Fig. 1	see Fig. 1	see Fig. 1	see Fig. 1
4	ic	$\neg(K\langle F\rangle ic \vee K\neg\langle F\rangle ip)$	see Fig. 1	see Fig. 1	see Fig. 1	see Fig. 1
5	in	—	see Fig. 1	see Fig. 1	see Fig. 1	see Fig. 1

Figure 4: An alternative definition of social influence $\mathcal{I}'$

pro opinion (which conforms to his private one). Or else he may ask himself why b is breaking the rules.

All this suggests that we should change the notion of social influence specified by Figure 1 to include potential knowledge of the inner state of other agents. We therefore define a new notion of social influence $\mathcal{I}'$ for reflective agents by the update rules shown in Figure 4. Here agents express their private pro or contra opinion if they know that they have some sincere support or they know that they have no sincere opposition. "—" means that any value may be inserted here.

While $\mathcal{I}'$ is an improvement of $\mathcal{I}$ for reflecting agents, it might not be the end of the story. For instance, if an agent knows that she has some sincere support from a friend, she might only want to express her true opinion if she knows that this friend will also start being sincere, which in turn might be reduced to her sincerely supporting friend knowing that she has sincere support as well. In other words, the knowledge state described in row 1 of Figure 4 should maybe be replaced by something along the following lines:

$$K\langle F\rangle(ip \wedge K\langle F\rangle ip) \vee K\neg\langle F\rangle ic.$$

Of course this supporting friend might be contemplating the same issue and we might have to add another "$K\langle F\rangle ip$". However, this could go on forever and we will leave further discussion of this for future research.

5 Conclusion and future research

This paper builds on (Christoff and Hansen 2013), which introduces a logical framework for reasoning about social influence that improves the work of (Liu, Seligman, and Girard 2014). In Section 2, we illustrated with some examples how agents may learn about others' private opinions by observing their expressed behavior – under the assumption that every agent obeys the rules of social influence $\mathcal{I}$ and that this is common knowledge. Moreover, we discussed how reflecting on the mechanisms of social influence and on why others act as they do may improve the agents' decision making. In Section 3 we defined a logic of social networks, knowledge, influence, and learning, which we used in Section 4 to formalize the issues raised in Section 2.

We have argued that reasoning by reflective agents in social networks is relevant to adequately capture the complex dynamics of social influence. As we have shown, some of this complexity can be captured by our logic of social networks, knowledge, influence, and learning (SKIL). Still, we have not settled on one true notion of social influence. It is unlikely that any such exists. In any case, there is still plenty of further work to be done in logical formalizations of social influence. As our new proposed notion $\mathcal{I}'$ has become even more complex than $\mathcal{I}$, it may be hard to find analytic results concerning it and it might be better to implement the full machinery of social influence $\mathcal{I}'$ in a multi-agent simulation environment instead. This we will leave for future work.

While our framework does not settle the issue concerning which notion of social influence is the right one for reflecting agents, our formal treatment of learning is actually much more general than the examples discussed so far. First of all, we have assumed, in all the examples mentioned, that the network structure among the agents was common knowledge. However, our logic SKIL can represent learning and reasoning about social influence even in cases where this assumption is dropped. This can be achieved by having different network structures at different worlds (i.e. by having different $\asymp_w$ for different worlds w).

Another direction in which our framework can be easily generalized concerns the influence modality $[\mathcal{I}]$. In (Christoff and Hansen 2014), the influence operator $\mathcal{I}$ is generalized to capture more complex dynamic phenomena on social networks. Firstly, instead of our two properties (inner and expressed opinion) each attaining one of three values (pro, contra, or neutral), (Christoff and Hansen 2014) allows for any finite number of properties each attaining a value from some fixed finite set. Moreover, the dynamics can be specified by dynamic transformations that allow any formula of the language as precondition for a change of values to properties instead of the 28 preconditions of Figure1. In this way, the alternative notion of social influence $\mathcal{I}'$ can also be captured by the framework of (Christoff and Hansen 2014). Now the learning modalities introduced in this paper can also be added to the framework of (Christoff and Hansen 2014) and thereby provide a model of how to reason about learning in other dynamic phenomena on social networks. We leave the details of this for future work though.

Acknowledgments The research of Zoé Christoff leading to these results has received funding from the European Research Council under the European Communitys Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 283963.

Jens Ulrik Hansen is sponsored by the Swedish Research Council (VR) through the project "Collective Competence in Deliberative Groups: On the Epistemological Foundation of Democracy".

Carlo Proietti is sponsored by the Swedish Research Council (VR) through the project "Logical modelling of collective attitudes and their dynamics".

References

Acemoglu, D., and Ozdaglar, A. 2011. Opinion dynamics and learning in social networks. *Dynamic Games and Applications* 1(1):3–49.

Areces, C., and ten Cate, B. 2007. Hybrid logics. In Blackburn, P.; van Benthem, J.; and Wolter, F., eds., *Handbook of Modal Logic*. Amsterdam: Elsevier. 821–868.

Aucher, G.; Balbiani, P.; del Cerro, L. F.; and Herzig, A. 2009. Global and local graph modifiers. *Electronic Notes in Theoretical Computer Science* 231(0):293–307. Proceedings of the 5th Workshop on Methods for Modalities (M4M5 2007).

Centola, D.; Willer, R.; and Macy, M. 2005. The emperor's dilemma. a computational model of self-enforcing norms. *American Journal of Sociology* 110(4):1009–1040.

Christoff, Z., and Hansen, J. U. 2013. A two-tiered formalization of social influence. In Huang, H.; Grossi, D.; and Roy, O., eds., *Logic, Rationality and Interaction, Proceedings of the Fourth International Workshop (LORI 2013)*, volume 8196 of *Lecture Notes in Computer Science*. Springer. 68–81.

Christoff, Z., and Hansen, J. U. 2014. Dynamic social networks logic. ILLC Prepublication Series Report PP-2014-09.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. New York, USA: Cambridge University Press.

Festinger, L. 1957. *A theory of cognitive dissonance*. Stanford University Press.

Jackson, M. O. 2010. *Social and Economic Networks*. Princeton University Press.

Kooi, B., and Renne, B. 2011. Arrow update logic. *The Review of Symbolic Logic* 4:536–559.

Liu, F.; Seligman, J.; and Girard, P. 2014. Logical dynamics of belief change in the community. *Synthese* 1–29.

O'Gorman, H. J. 1986. The discovery of pluralistic ignorance: An ironic lesson. *Journal of the History of the Behavioral Sciences* 22(October):333–347.

Prentice, D. A., and Miller, D. T. 1993. Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology* 64(2):243–256.

Schelling, T. C. 1978. *Micromotives and Macrobehavior*. W. W. Norton and Company.

Seligman, J.; Liu, F.; and Girard, P. 2011. Logic in the community. In Banerjee, M., and Seth, A., eds., *Logic and Its Applications*, volume 6521 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 178–188.

Zhen, L., and Seligman, J. 2011. A logical model of the dynamics of peer pressure. *Electronic Notes in Theoretical Computer Science* 278(0):275–288. Proceedings of the 7th Workshop on Methods for Modalities (M4M 2011) and the 4th Workshop on Logical Aspects of Multi-Agent Systems (LAMAS 2011).