Towards a Value-driven Explainable Agent for Collective Privacy

Extended Abstract

Francesca Mosca
King's College London
francesca.mosca@kcl.ac.uk

Jose M. Such
King's College London
jose.such@kcl.ac.uk

Peter McBurney
King's College London
peter.mcburney@kcl.ac.uk

ABSTRACT

Online social networks lack support for the collaborative management of access control. This is crucial for content that may involve multiple users such as photos, as this lack of support causes conflicts that lead to privacy violations. Previous research proposed collaborative mechanisms to support users in these cases, but most of these attempts fail to satisfy some desirable requirements, such as explainability, role-agnosticism, adaptability, and being utility- and value-driven at the same time. In this paper, we outline an agent architecture that has been designed to meet all these requirements.

KEYWORDS

Multiuser Privacy; Explainable Agents; Morally-aligned Agents

ACM Reference Format:

Francesca Mosca, Jose M. Such, and Peter McBurney. 2020. Towards a Value-driven Explainable Agent for Collective Privacy. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

In the recent years, privacy is raising increasing concern. Especially regarding online behaviour, users are becoming more aware of the consequences of privacy leaks. Online services have in general progressed in protecting the individual privacy, but disregarded the collective one. In fact, privacy is not only what we decide to share about ourselves, but also what others can disclose about us [20]. This is particularly relevant in the context of online collaborative platforms, such as social networks, which represent the most studied example, where multiuser privacy conflicts (MPCs) occur whenever the degree of online publicity/privacy that is assigned to some content by some user, namely the *uploader*, is not aligned with the privacy preferences of the other involved users, namely the *co-owners*.

In online social networks (OSNs), MPCs are highly frequent and documented [4, 10, 21, 24]. However, the large majority of conflicts involves people who are willing to collaboratively find a solution that is satisfying for all the involved users. In particular, uploaders of material wished to have known in advance the negative consequences experienced by some co-owners, to avoid the conflict beforehand [21].

Researchers have proposed a variety of solutions, ranging from game theoretical solutions [15, 17, 19, 22], to argumentation-based systems [5, 8], to learning models [6, 23], and more technical, fine-grained systems [7]. However, none of these approaches for solving MPCs fully satisfies a number of desirable requirements, which we list informed by theoretical and empirical studies in privacy, artificial intelligence and social sciences [1, 9, 14, 21]:

- *explainability*: a model should be able to justify its output, by providing explanations of its processes [11] to help the users comprehend the recommended solution [14];
- *adaptability*: a model should behave differently according to the users' subjective preferences, because different individuals manage privacy in different ways and in different contexts [1];
- *role-agnosticism*: a model should treat all the users involved in a MPC in the same way regardless of their role, because the asymmetric access control management of uploaders and co-owners is among the main causes for MPCs [24];
- *utility-driven*: a model should consider solutions to MPCs according to the personal advantage or disadvantage that the involved users can perceive [9];
- *value-driven*: a model should support the promotion of human values, because empirical evidence suggests that users go beyond their personal utility when evaluating solutions and compromises, being aware of their impact on the other involved users [4, 21].

2 THE MODEL

In this paper we outline an agent architecture that can act on behalf of the users of an OSN, both uploaders and co-owners, and support them to solve MPCs. We design the agent in such a way to satisfy all the requirements previously introduced, i.e. the agent is explainable, adaptive, role-agnostic, and both utility- and value-driven. In the remaining part of this section, we introduce the crucial features of the agent architecture which allow to satisfy these requirements.

Whenever a conflict occurs, each agent $k \in Ag$ representing a user involved in the MPC evaluates the set of possible solutions SP, which includes the initial preferences of each involved user plus an option that represents a compromise for everyone (i.e., that is different from all the initial preferences). Each agent computes for each possible candidate solution a *score* $s_{k,sp} = u_{k,sp} \cdot v_{k,sp}$, which interprets the user k's appreciation of the sharing policy sp in terms of both utility (u, defined in sec. 2.1) and moral values (v, defined in sec. 2.2). Then, the agent that acts on behalf of the uploader of the content collects from the other agents their individual scores and, by maximising the aggregated scores, identifies the optimal solution to the conflict:

$$solution = \arg\max_{sp \in SP} \sum_{k \in Ag} s_{k,sp}. \tag{1}$$

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

2.1 Utility-driven Component

We define a OSN as a graph, where the nodes represent the users and the links their connections. The policies to share content online are defined in terms of maximum distance, i.e. length of the path between users, and minimum intimacy, i.e. the weight of such path.

Every user has a preferred sharing policy for any item that could be shared online, and this can be elicited automatically [12, 18]. By comparing the preferred audience, i.e. the set of users who satisfy the conditions established by the preferred sharing policy, with a possible solution audience, i.e. the set of users who satisfy the conditions established by the sharing policy examined as solution, the agent can estimate the user's appreciation of the solution. In particular, based on empirical evidence [9, 21], we assume that (i) users gain utility when desired people are granted access to an online item; (ii) users lose utility when desired people are denied access to an online item; (iii) users lose utility when undesired people are granted access to an online item. By quantifying these intuitions and aggregating them, each agent computes the *utility function* for each user k given a sharing policy sp: $u_{k,sp}$.

2.2 Value-driven Component

In a similar way as in [13], we base the moral component of the agent on the theory of basic values by Schwartz [16], that is one of the most well-known and established theory of human values and is backed by strong empirical evidence. In this theory, values are socially desirable concepts that represent the mental goals which drive human behaviour [3, 16], with people taking daily decisions influenced by the values they believe in.

We adapt the interpretation of the Schwartz values to the MPC scenario, by considering how the main *value-dimensions*, namely self-enhancement, self-transcendence, conservation and openness-to-change, impact on the user's behaviour while interacting with other users in order to find an acceptable solution. For instance, by accommodating someone else's preference for a more private sharing policy, a user could promote self-transcendence and conservation. We can elicit the users' preferred order over the value-dimensions through some tools validated by Schwartz [16]. Once the preferred values are known, the agent can evaluate the user k's appreciation of any candidate solution sp with the *value promotion function* $v_{k,sp}$, which describes whether each value-dimension is promoted or demoted by selecting the option sp.

2.3 Explainable Components

We provide our agent architecture with the *cognitive process* [11] that is necessary for explainability. We argue that this is possible by implementing practical reasoning techniques [2] from computational argumentation.

We model the resolution of a MPC as a joint action, i.e. a complex action which comprises simple actions performed by individual agents, such as the combination of an offer, in terms of sharing policies, of the uploader and a response (either accept or reject) for each of the co-owners.

Each agent is able to instantiate an argumentation scheme that supports the choice of a particular simple action. The reasoning process that leads to the identification of the best solution is supported by an AATS+V [2], that provides the underlying semantics for the argumentation scheme and its critical questions. By going through each step of the practical reasoning process, the agent gathers all the necessary knowledge to provide an explanation for its decision, that is, the agent presents an appropriate cognitive process for the explainability requirement.

2.4 Role-agnosticism and Adaptability

The solution to the MPC is computed in such a manner that role-agnosticism is satisfied. In fact, as we show in 1, because of the commutative property of the addition, the individual scores for each candidate solution are aggregated in a way that is not sensitive to permutations of the users, i.e. all the involved users are treated the same.

Also, the solution is identified in order to satisfy as much as possible the users' preferences, which always influence explicitly the outcome. This makes the model adaptive.

3 FUTURE DIRECTIONS

We introduced an agent-based approach to solve MPCs in OSNs which, first in the related literature, satisfies a number of desirable requirements, namely explainability, adaptability, role-agnosticism, being utility-driven and value-driven. For some of these requirements and other properties such soundness and completeness, we will work on formal proofs; we will also study the computational complexity of the model.

However, we still need to validate such model. First, through software simulations, we will test the goodness of our solution concept compared with the ones provided by other models suggested in the literature. Then, user studies will inform us on the efficacy of such model, i.e. the users' appreciation for the recommended solution. However, before deploying a user study, we need to complete the design of the explainable component of the model. In fact, as it is properly pointed out in [11], the cognitive process is necessary but not sufficient for guaranteeing explainability: the agent needs to be provided with a *social process* as well, that is the social ability of satisfyingly interacting with the user. Then, the user interaction with the model will inform us on their appreciation for both the recommended output and the provided explanation.

Finally, we would like to extend our model in order to consider also adversarial behaviour. This is because, even if the vast majority of MPCs are created without malicious intent and most MPC situations can be considered non-adversarial and collaborative [21], there are some much less frequent but severe cases of MPCs where adversarial behaviour may be present, such as revenge porn and cyber-bullism.

REFERENCES

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.

[2] Katie Atkinson and Trevor Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *AIJ* 171, 10-15 (2007), 855–874.

[3] Anat Bardi and Shalom H Schwartz. 2003. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin* 29, 10 (2003), 1207–1220.

[4] Andrew Besmer and Heather Richter Lipford. 2010. Moving beyond untagging: photo privacy in a tagged world. In *Proc. of CHI*. ACM, 1563–1572.

[5] Ricard L Fogues, Pradeep K Murukannaiah, Jose M Such, and Munindar P Singh. 2017. Sharing policies in multiuser privacy scenarios: Incorporating context, preferences, and arguments in decision making. *ACM TOCHI* 24, 1 (2017), 5.

[6] Ricard L Fogues, Pradeep K Murukannaiah, Jose M Such, and Munindar P Singh. 2017. SoSharP: Recommending Sharing Policies in Multiuser Privacy Scenarios. *IEEE Internet Computing* 21, 6 (2017), 28–36.

[7] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/Off: preventing privacy leakage from photos in social networks. In *Proc. of CCS*. ACM Press, 781–792.

[8] Nadin Kökciyan, Nefise Yaglikci, and Pınar Yolum. 2017. An argumentation approach for resolving privacy disputes in online social networks. *ACM TOIT* 17, 3 (2017), 27.

[9] Hanna Krasnova, Sarah Spiekermann, Ksenia Koroleva, and Thomas Hildebrand. 2010. Online social networks: Why we disclose. *Journal of information technology* 25, 2 (2010), 109–125.

[10] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. 2011. We're in it together: interpersonal management of disclosure in social network services. In *Proc. of CHI*. ACM, 3217–3226.

[11] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *AIJ* (2018).

[12] Gaurav Misra and Jose M Such. 2017. PACMAN: Personal Agent for Access Control in Social Media. *IEEE Internet Computing* 21, 6 (2017), 18–26.

[13] Francesca Mosca, Jose M Such, and Peter McBurney. 2019. Value-driven Collaborative Privacy Decision Making. In *Proc. of AAAI PAL Symposium*.

[14] F. Paci, A. Squicciarini, and N. Zannone. 2018. Survey on access control for community-centered collaborative systems. *ACM CSUR* 51, 1 (2018).

[15] Sarah Rajtmajer, Anna Squicciarini, Jose M Such, Justin Semonsen, and Andrew Belmonte. 2017. An Ultimatum Game Model for the Evolution of Privacy in Jointly Managed Content. In *Proc. of GAMESEC*. Springer, 112–130.

[16] Shalom H Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online readings in Psychology and Culture* 2, 1 (2012), 11.

[17] Anna Cinzia Squicciarini, Mohamed Shehab, and Federica Paci. 2009. Collective privacy management in social networks. In *Proc. of WWW*. ACM, 521–530.

[18] Anna Cinzia Squicciarini, Smitha Sundareswaran, Dan Lin, and Josh Wede. 2011. A3p: adaptive policy prediction for shared images over popular content sharing sites. In *Proc. of HT*. ACM, 261–270.

[19] Jose M. Such and Natalia Criado. 2016. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE TKDE* 28, 7 (2016), 1851–1863.

[20] Jose M. Such and Natalia Criado. 2018. Multiparty Privacy in Social Media. *Commun. ACM* 61, 8 (2018), 74–81.

[21] Jose M Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo privacy conflicts in social media: a large-scale empirical study. In *Proc. of CHI*. ACM, 3821–3832.

[22] Jose M. Such and Michael Rovatsos. 2016. Privacy Policy Negotiation in Social Media. *ACM TAAS* 11, 1 (2016), 1–29.

[23] Onuralp Ulusoy and Pınar Yolum. 2019. Emergent Privacy Norms for Collaborative Systems. In *Proc. of PRIMA*. 514–522.

[24] Pamela Wisniewski, Heather Lipford, and David Wilson. 2012. Fighting for my space: Coping mechanisms for SNS boundary regulation. In *Proc. of CHI*. ACM, 609–618.