A Desiderata for Modeling and Reasoning with Social Knowledge

Fabio R. Gallo[1,3], Natalia Abad Santos[2],
Gerardo I. Simari[1,3], and Marcelo A. Falappa[1,3]
{fabio.gallo,gis,mfalappa}@cs.uns.edu.ar,nasantos@uns.edu.ar

[1] Institute for Computer Science and Engineering UNS-CONICET, Argentina
[2] Dept. of Mathematics, Universidad Nacional del Sur, Bahía Blanca, Argentina
[3] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

Abstract. The ongoing surge in the amount of users that engage in online activities, as well as the expansion of the type of such activities, has recently made it clear that there is a widening gap between current knowledge representation and reasoning tools and the type of knowledge that is essentially up for grabs for whoever is willing (and has the tools) to extract it from social media sites. In this position paper, we propose the concept of Social Knowledge Base (SKB, for short) as an extension of traditional KBs with representation and reasoning capabilities that arise from the singular combination of characteristics that define this setting: (i) ontological knowledge, (ii) user preferences, (iii) reasoning under uncertainty, (iv) stream reasoning, and (v) representation of complex social networks. We propose a list of desirable properties—a desiderata—that next-generation KR formalisms for modeling and reasoning with SKBs should enjoy. The treatment is non-technical, focusing on building a road map of the formidable list of problems that must be solved in this complex setting rather than proposing a concrete solution, which would be impossible in a single article. We conclude by proposing some first steps towards achieving this goal.

Keywords: Social Web, Complex Networks, Reasoning Under Uncertainty, Preferences, Ontology Languages

1 Introduction and Motivation

Recent times have seen a veritable explosion in the amount and kind of information that is available to anyone with a connection to the Internet. This explosion has its roots in the so-called World Wide Web [1], which revolutionized internet applications by allowing users to link resources with one another and easily organize the material that they wish to publish. The second revolution came with the advent of Web applications in which users produced their own material, such as in blogs or forums where users share information ranging from plain text to photos, videos, and audio—this "new version" is often referred to as *Web 2.0* to highlight that a step was taken since the implementation of the original idea.

Finally, in the last few years, the Web has once again taken an evolutionary step: in its current form, which many refer to as *Social Web* (or the *Web 3.0*), users and the relationships among them are the central participants. Another revolutionary aspect that appeared in Web 3.0 is that data is now also produced automatically by computers; examples of this are data output by the host of sensors now carried by most smart phones, or by the smart homes that are slowly becoming more and more present.

Unfortunately, research in Knowledge Representation and Reasoning formalisms has lagged behind this rapid evolution in how data is created and disseminated. The goal of this position paper is thus to explore a desiderata—a list of desirable characteristics—for the development of what we will call *social knowledge bases* (SKBs, for short). The idea behind this line of research is to derive a framework and methodology akin to Ontology Based Data Access (OBDA) [23] that is specialized for the unique social aspects discussed above. Our work is influenced by recent proposals in the complex networks literature [21, 20], which also establishes a set of criteria that is desirable for modeling cascades, a specific phenomenon that—as we will see—also plays an important role in our setting. Our desiderata are therefore inspired in this work, but necessarily go above and beyond their scope given the greater generality of the problems that need to be solved.

We now describe two settings that we will use as running examples to motivate our discussion. The first setting is an online matchmaking service.

Example 1 (Friendship/Dating site). Consider a web site where people register and complete a profile with the objective of meeting new people—the goal might be to establish a romantic relationship or simply make new friends. As a way to simplify the creation of profiles, the site offers the option to log in with the users' favorite social media site (like Facebook, Twitter, Google Plus, etc.), and optionally also link the profile with multiple such sites.

The main aim of the site is to match people who have compatible personalities; in order to have tools that can be leveraged to solve this (very difficult) problem, the site allows users to explicitly specify their preferences in different domains, such as music, literature, movies, and even relationships—these inputs are complemented by the information that is extracted (with the users' permission) from their linked social media profiles. ∎

The second example setting is a comprehensive trip organization service.

Example 2 (Travel site). As a second example, suppose we have a web site (similar to TripAdvisor[1]) that is designed to help people choose a place to spend their next vacation; it includes information on destinations, transportation, hotels, tours, restaurants, best season to go, etc. Members publish reviews, including numeric scores for several different categories as well as free text where they can go into detail regarding their experience. There are rich social features available, such as tagging in posts or reviews, suggestions, and private messages.

[1] `http://www.tripadvisor.com/`

As before, we assume that users are able to sign in with their social media accounts, which gives the system the possibility to extract relevant information—for instance, to suggest a destination, the system might use the fact that a user participates in Facebook groups for learning the German language to infer that they probably would like to travel to Germany. ■

In the following, we describe a list of desiderata to achieve the goal of designing a formalism to model and reason with social knowledge bases. As we will argue below, the problem essentially requires the combination of knowledge representation machinery for areas that have up to now largely been considered in relative isolation: (i) ontology languages, (ii) preference models, (iii) reasoning under uncertainty, (iv) stream reasoning, and (v) complex social networks.

2 Desiderata for Building and Querying Social KBs

In the previous section, we argued that it is necessary to develop novel KR tools to reason with social data; we will now offer further support for this argument by proposing a series of characteristics and capabilities that SKBs should have—developing such a desiderata has the additional value of acting as a road map for guiding future research efforts in this direction. For further discussion of literature related to each point, see Section 4,

(1) Model complex networks. In social knowledge bases, it should be possible for entities to be of different types: people, products, companies, books, movies, etc. Furthermore, it should be possible for there to be different kinds of relationships among them. It is thus necessary to be able to represent networks with different kinds of nodes, as well as multiple attributes and relationships for each one—such models are often referred to as *complex networks* [2].

Consider the setting from Example 1; in this case, it is clear that it would be useful for connections between users to contain additional information about the relationship they have. For instance, kind of tie (relative, classmate, work partner, etc.), how long they have known each other, how many social media sites they frequent, etc. Another important observation is that connections do not always need to be symmetric—in the dating example, person A can consider person B to be a good match, but B may not agree. Having rich information about entities and how they are related can thus be useful to improve users' experiences.

(2) Model atomic actions. A specific set of actions (by agents or exogenous factors) that can occur in the domain need to be identified.

Considering social media sites like Facebook or Google Plus, such actions could include posting, commenting, liking/+1 a post, friending/unfriending, messaging, etc. It is these actions that will be the building blocks for inferences about preferences or regarding reasoning under uncertainty, as we will discuss below. Also in connection with a point discussed in the following is the fact that an adequate selection of atomic actions to be modeled will have an impact on computational tractability. User data in social media suffers constant

change, and a regular user could produce a large amount of data per day; depending on the way in which the SKB will be used, it may not be necessary to incorporate all of this data into the model. For instance, regarding the setting in Example 2, it may be a good idea to incorporate comments as actions since users may give information regarding preferences in their mode of travel (for example, that they are afraid to fly); on the other hand, this might be less relevant for Example 1. Hence, it is essential to characterize and prioritize atomic actions so that resources are not wasted by processing and storing unnecessary data.

(3) Model quantitative and qualitative preferences. Quantitative preferences are often useful when automatically learning from data, or in simple domains; on the other hand, qualitative preferences (defining strict partial orders) are often more naturally elicited from human beings but more difficult to extract automatically.

To illustrate this point, consider the travel setting from Example 2. Quantitative preferences could be obtained from users' explicit rankings of favorite cities, countries, museums, beaches, etc. On the other hand, reviews or polls could also provide less structured preferences, such as the fact that the user prefers beach destinations to mountain ones, or that hotels near the city center are preferred over those that are not.

(4) Reason about groups. Social knowledge is inherently related to groups of entities (where entities are not necessarily all people); groups sometimes function as higher-level entities with their own preferences, relationships, etc.

There are many ways in which groups can be important when leveraging social knowledge. In Example 1, a group may be defined with respect to people's age group and interests, and the general preferences of such groups can be used in order to supplement the preferences of the individual. On the other hand, in Example 2 one can take the users' closest friends as a source of suggestions for travel destinations or activities—in this case, the group of friends is used as the basis of a kind of crowdsourcing. Challenges in this respect involve identifying the best possible composition of groups (for instance, determining who the users' closest friends are by considering how long they have known each other, share interests, etc.), and what to do about group members with conflicting preferences. The latter has been recently addressed in [16].

(5) Reason about cascading processes. One of the main characteristics of social networks is that information "flows" through them—this kind of dynamic is often referred to as a "cascading process" [11].

A clear example of this kind of process can occur in the travel domain (Example 2), where a user might travel to a new destination and post a series of pictures with very positive comments about their experience. This might cause several of the user's connections to "like" that destination and even plan trips there—the process can of course continue, with the new converts' activities causing some of their connections to do the same. It is thus important to model how influence propagates; there is extensive work in this area, and the logic programming proposal of [20] is perhaps the closest in spirit to the general approach that is required for SKBs.

(6) Flexible characterization of consistency/inconsistency. Classical conceptions of consistency are not adequate for modeling the kind of information that occurs in social settings—a more flexible approach is required for handling conflicts.

In our example settings, simple inconsistency cases might occur, for instance, when users have accounts in several social media sites but focus more on one than the others. Since data is usually not shared between accounts, it can occur that a user who lives in city C_1 later moves to city C_2 and only updates their profile for one of the accounts. An SKB taking information from these profiles would thus encounter an inconsistency. A more challenging case of inconsistency, much more difficult to characterize, is the case of a user of the system in Example 2 who strongly prefers beach destinations but suddenly starts paying attention to mountain-related places and activities (such as with +1s, posts, comments). The classical way to deal with the above situations is to try to modify the information contained in the knowledge base as little as possible in order to reach a consistent state without losing unnecessary information [7]; this is closely related to the following point.

(7) Social network-based belief revision operators. In close connection to the previous point, belief revision operations need to be applied in response to different kinds of events that signal changes in the SKB. The difference with respect to the classical setting is in relation to other points on this list—in particular, consistency, cascades, and uncertainty.

Among these, the relationship between cascades and belief revision operators is, to the best of our knowledge, never been studied. As an example, consider our travel setting and suppose an influential individual changes their opinion with respect to a certain destination (for instance, they start to express negative opinions about it and "unlike" the relevant pages), causing others to follow suit; how should this cascading belief revision process evolve?

(8) Reason about uncertainty. Conflicting information and inherently uncertain data makes it necessary to have an explicit representation of uncertainty.

There are many examples of the need to reason with uncertain knowledge. In our example dating application, some user information is private, and so cannot be directly used and perhaps only approximations can be obtained. For instance, user location can be approximated by content-based methods leveraging features of posts, such as mentions of place names and use of local dialect—since these are prone to error, a measure of probability must be assigned that depends on the kind and amount of information that supports each inference. Approaches to reasoning with ontological knowledge and user preferences have recently been proposed in [17].

(9) Rich query answering. Social knowledge is rich, and access to such knowledge often requires queries that combine the basic relational database-style queries with the graph-based queries often used in linked data [4].

Consider the travel setting from Example 2; queries to an SKB in this case might involve complex requests such as "*hotels with free wi-fi connection that have been positively reviewed by people who share my views and that at least one*

connection recommends, in order of preference". This involves reasoning under uncertainty (it is not always possible to determine if free wi-fi is available), reasoning about groups, and network structure, and preferences. Formalizing novel types of queries for SKBs, and obtaining effective algorithms to answer them, is therefore one of the main challenges ahead. Recent work [10] that can be leveraged towards this goal has proposed efficient algorithms for social networks under uncertainty.

(10) Time and space constraints: scalability and stream reasoning. Successful SKB formalisms must be able to cope with very large knowledge bases that are updated often with information that must be processed on the fly (or nearly so).

Micro-blogging is a clear example of how often new data is created: Twitter has about 100M active users who post over 230M tweets a day [3]. Processing such a high volume of data—much of which may not even be valuable [13] and that has a short life span—is a formidable challenge. An even greater challenge is to make the tools and processes that we propose in the previous points work adequately in such a setting. Considering the travel application from Example 2, a site with many active users must deal with a large volume of new comments, reviews, multimedia posts, and connections between users; an SKB that models even a portion of this activity must therefore be able to keep up with updates that, as we have seen, involve complex reasoning tasks.

3 Outlining a Framework for Social Knowledge Bases

Using the list of features discussed in Section 2 as a guide, we now briefly outline what a framework that integrates all of them might look like. A social knowledge base can be modeled as a 5-tuple of the form $SKB = (O, N, P, M, B)$, where:

- O is an *ontology* modeling the general knowledge about the domain. For instance, in the travel domain O would contain the database of hotels, flight routes, etc., as well as intensional knowledge such as *hostels are a kind of lodging*, or *wi-fi is a kind of internet connection*. This component could be modeled with the Datalog+/– family of ontology languages [5], which contains many different fragments focused on tractable query answering that generalize other well-known ontology languages such as the DL-Lite family of description logics.
- N is a model of the underlying *social network structure*. Since this is a kind of ontological knowledge, it could also be modeled using Datalog+/–; however, we propose to model them as separate components so that other approaches that are more specific can be used, such as the MANCaLog language [20].
- P is a *preference model* over the consequences of ontology O. This kind of integration has already been proposed in [14] and later extended to preferences under uncertainty [17] and preferences over groups [16].

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
O – Ontology	×	×	×	×	×	×	×	×	×	×
N – Network	×			×	×	×	×	×	×	×
P – Preferences			×	×		×		×	×	×
M – Probabilistic model	×				×	×		×	×	×
B – BR operators		×					×	×		×

Fig. 1. Assuming an SKB of the form (O, N, P, M, B), this table shows an example of the involvement of each component in satisfying the desiderata from Section 2. Different application settings may require different setups.

- M is a *probabilistic model* for ontology O. There are different ways in which probabilistic uncertainty can be integrated into ontological knowledge. For instance, in [8] annotations are added to both extensional and intensional knowledge, and the probabilistic model provides a probabilistic distribution over the annotations—this is an elegant way to allow for a separation of interests between the two models. Of course, other possibilities may be more appropriate depending on the domain of application.
- B is a set of *belief revision operators*. As was motivated before, revision operators that are informed by all the other components are needed in order to modify the knowledge base when new information needs to be incorporated. One approach in the logic-based probabilistic belief revision literature is the recent work of [22], which studies quantitative approaches to belief revision in a probabilistic structured argumentation language.

These components are coupled differently depending on the modeling or reasoning task that they are required to perform—the table in Figure 1 shows how each component might be typically involved in addressing each desideratum proposed above. For instance, desideratum D6 regarding consistency might involve components O, N, P, and M, since determining consistency may require ontological knowledge, access to network connections, user preferences, and probabilities. Of course, different applications may require different setups; returning to our example, in this case perhaps social connections are not considered for assessing consistency.

4 Discussion: Related Work and Challenges

Extracting, representing, and reasoning about the kind of information described above is a complex problem; although the examples may look simple, many issues arise when trying to combine all available data. There are some recent developments in the literature on ontological languages that are related to our present efforts in that they have already begun to investigate how some subsets of these areas can be adequately combined. The Datalog+/– family of ontology languages [5] has recently received a lot of attention given its flexibility and variety of available fragments that ensure tractable query answering. In [14],

the authors explore an extension of Datalog+/– with preference models that allows to rank the answers to queries with respect to users' preferences; a further extension to this approach was proposed in [16], where group preferences are considered as well. A related approach, considering the problem from the somehow dual perspective of extending the general model of CP-theories for preference representation with ontological constraints, was recently proposed in [19]. Another recent approach is the Prob-$\mathcal{EL}$ formalism [9], which extends the $\mathcal{EL}$ description logic with probabilistic uncertainty over both assertional and terminological knowledge.

In a separate but closely related vein, Datalog+/– was also extended with probabilistic models in [8], where the authors study both algorithms for ranking answers with respect to their associated probabilities and query answering under inconsistency. These two lines were considered together in [17], where the authors explore the problem of ranking answers to queries with respect to both probabilistic uncertainty and user preferences. Several other ontology languages have been extended with probabilistic uncertainty—see [18] for a survey of earlier approaches. Also related to this line of research is the study of probabilistic databases [12], where the ontological aspect is missing but the focus is rather on computational tractability. Another quite recent formalism for expressing preferences under uncertainty—also not ontology-based—was introduced in [15].

Stream reasoning [6] refers to the problem of processing information that continuously becomes available and cannot all be stored (a fixed window is generally assumed). From the point of view of making sense of data in social media, the recent work of [3] analyzes key research questions for mining data with semantic content from social media streams. Their work is perhaps the closest in spirit to our goal, though the main difference is that they are focused primarily on extracting information while we are focusing on the problems of adequately organizing and accessing the information that is already extracted.

Towards a general framework

We have thus far proposed a set of desirable properties and sketched the organization of a framework for modeling and reasoning with SKBs; however, there are many challenges towards materializing the general vision. Figure 2 shows a high-level outline of this vision—SKBs are populated by three general sets of sources: social media and general Web-based resources, users themselves, and users' interactions with others. A mix between learning, scraping, and elicitation techniques, as well as knowledge engineering in general, will help obtain not only the information necessary for the individual components of the SKB but also the relationships between them. These components will be built by leveraging as much as possible existing tools (such as Bayesian networks, Datalog+/–, etc.). Even if we assume that all necessary information is available to populate these components, there are many challenges associated with bringing them together: scalability issues arising from the combination of individually tractable components, semantic issues arising from the combination of open-world and closed-world assumptions, alignment issues arising from different schemas used

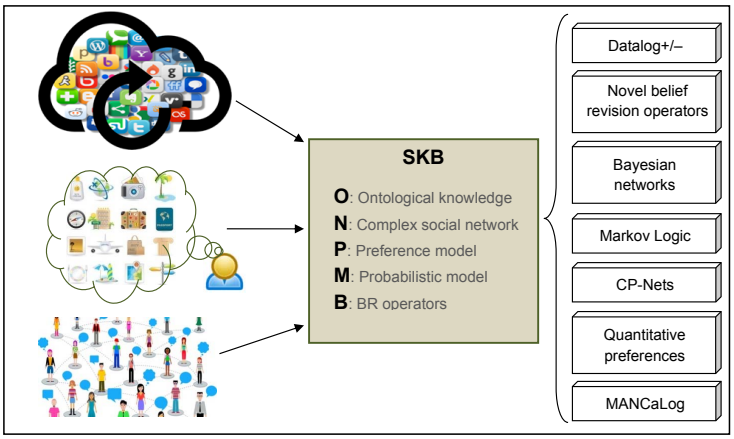

Fig. 2. A high-level overview of the proposed process of modeling and reasoning with social knowledge. SKBs are built with information from the Web, individual users interacting with online services, and social media. Individual components of the SKB are modeled using different kinds of formalisms proposed in the literature for solving more specific subproblems.

in different components, normalization issues arising from combining different quantitative preferences, and so on.

5 Conclusions

In this position paper, we have discussed the need to develop novel knowledge representation and reasoning tools and techniques that are adequate for tackling the challenges that come with modeling social knowledge. We proposed a set of desiderata to guide the development of such formalisms, and briefly outlined how a unifying model can be built by leveraging existing research and novel developments. The main contribution of such a discussion is the proposal of a road map to guide research efforts towards this goal, as well as the novel proposal of combining several research lines that up to now have been considered largely in isolation: ontologies, preferences, uncertainty, stream reasoning, and complex social networks.

Acknowledgments. This work was supported by funds provided by CONICET and Universidad Nacional del Sur, Argentina. Some of the authors of this work were also supported by the U.S. Department of the Navy, Office of Naval Research, grant N00014-15-1-2742. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

References

1. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american 284(5), 28–37 (2001)
2. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports 424(4), 175–308 (2006)
3. Bontcheva, K., Rout, D.: Making sense of social media streams through semantics: a survey. Semantic Web 1, 1–31 (2012)
4. Broecheler, M., Pugliese, A., Subrahmanian, V.: DOGMA: A disk-oriented graph matching algorithm for RDF databases. In: Proc. of ISWC 2009, LNCS, vol. 5823, pp. 97–113. Springer (2009)
5. Calì, A., Gottlob, G., Lukasiewicz, T.: A general Datalog-based framework for tractable query answering over ontologies. J. Web Sem. 14, 57–83 (2012)
6. Della Valle, E., Ceri, S., Van Harmelen, F., Fensel, D.: It's a streaming world! reasoning upon rapidly changing information. IEEE Intell. Sys. (6), 83–89 (2009)
7. Fermé, E., Hansson, S.O.: Agm 25 years. J. of Philos. Log. 40(2), 295–331 (2011)
8. Gottlob, G., Lukasiewicz, T., Martinez, M.V., Simari, G.I.: Query answering under probabilistic uncertainty in Datalog+/– ontologies. AMAI 69(1), 37–72 (2013)
9. Gutiérrez-Basulto, V., Jung, J.C., Lutz, C., Schröder, L.: A closer look at the probabilistic description logic Prob-EL. In: Proc. of AAAI (2011)
10. Kang, C., Pugliese, A., Grant, J., Subrahmanian, V.: STUN: querying spatio-temporal uncertain (social) networks. Soc. Netw. Anal. Min. 4(1), 1–19 (2014)
11. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proc. of KDD '03. pp. 137–146. ACM (2003)
12. Koch, C., Olteanu, D., Re, C., Suciu, D.: Probabilistic Databases. Morgan-Claypool (2011)
13. Liu, Y., Kliman-Silver, C., Mislove, A.: The tweets they are a-changin: Evolution of twitter users and behavior. In: Proc. of ICWSM. vol. 13, p. 55 (2014)
14. Lukasiewicz, T., Martinez, M.V., Simari, G.I.: Preference-based query answering in Datalog+/– ontologies. In: Proc. of IJCAI. pp. 1017–1023. IJCAI/AAAI (2013)
15. Lukasiewicz, T., Martinez, M.V., Simari, G.I.: Probabilistic preference logic networks. In: Proc. of ECAI 2014. pp. 561–566 (2014)
16. Lukasiewicz, T., Martinez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Ontology-based query answering with group preferences. ACM Trans. Internet Techn. 14(4), 25:1–25:24 (2014)
17. Lukasiewicz, T., Martinez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Preference-based query answering in probabilistic Datalog+/– ontologies. J. Data Semantics 4(2), 81–101 (2015)
18. Lukasiewicz, T., Straccia, U.: Managing uncertainty and vagueness in description logics for the Semantic Web. J. Web Sem. 6(4), 291–308 (2008)
19. Noia, T.D., Lukasiewicz, T., Martínez, M.V., Simari, G.I., Tifrea-Marciuska, O.: Combining existential rules with the power of CP-theories. In: Proc. of IJCAI 2015. pp. 2918–2925 (2015)
20. Shakarian, P., Simari, G.I., Callahan, D.: Reasoning about complex networks: A logic programming approach. TPLP 13(4-5-Online-Supplement) (2013)
21. Shakarian, P., Simari, G.I., Schroeder, R.: MANCaLog: A logic for multi-attribute network cascades. In: Proc. of AAMAS-2013 (2013)
22. Simari, G.I., Shakarian, P., Falappa, M.A.: A quantitative approach to belief revision in structured probabilistic argumentation. AMAI (2015, *In Press*)
23. Spanos, D.E., Stavrou, P., Mitrou, N.: Bringing relational databases into the semantic web: A survey. Semantic Web 3(2), 169–209 (2012)