Utilising Provenance to Enhance Social Computation*

Milan Markovic, Peter Edwards, and David Corsar

Computing Science & dot.rural Digital Economy Hub, University of Aberdeen, Aberdeen, AB24 5UA
{m.markovic,p.edwards,dcorsar}@abdn.ac.uk

Abstract. Many online platforms employ networks of human workers to perform computational tasks that can be difficult for a machine (e.g. reporting travel disruption). Such systems have to make a range of decisions, for example, selection of suitable workers for a task. In this paper we present an approach that utilises Semantic Web technologies and provenance to support such decision-making processes.

1 Introduction

Recent years have seen the emergence of several web-based platforms[1] that use collective intelligence, i.e. "groups of individuals doing things collectively that seem intelligent" [4]. Within such platforms, workers are recruited to perform computational tasks, such as data creation or data maintenance.

Robertson and Giunchiglia define one such approach, *a social computation* (SC) as: "a computation for which an executable specification exists, but the successful implementation of this specification depends upon computer-mediated social interaction between the human actors in its implementation"[11]. Figure 1 illustrates a specification for a simple scenario that gathers reports about travel disruptions from workers (e.g. via a smartphone app). Workers can either provide the report themselves or delegate this task to other workers in their social network. When designing the specification, it can be associated with *social properties* that define: "the drivers for the adoption and spread of the computation through the social group with which it engages" [11]. For example, a social property for our scenario could be: "to secure a reward, a worker has to provide a report himself or refer the task to someone in his social network that can provide a report"[2].

The use of humans in SC can result in issues related to the reliability of workers, workforce recruitment (i.e. ensuring the most suitable people are used),

* The research described here is supported by the award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub; award reference: EP/G066051/1.

[1] Examples include: Amazon's Mechanical Turk (www.mturk.com/mturk/welcome), Zooniverse (www.zooniverse.org/); Crowdflower (www.crowdflower.com/).

[2] If the second worker further delegates the task, the first worker will not receive a reward.

H. Alani et al. (Eds.): ISWC 2013, Part II, LNCS 8219, pp. 440–447, 2013.

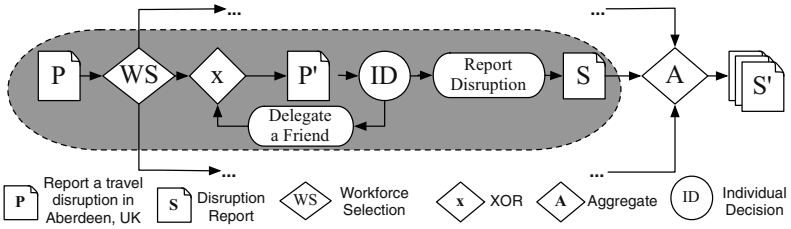

Fig. 1. An SC specification for travel disruption reporting, described using the CrowdLang [5] notation. A problem statement (P), i.e. to create a report of travel disruption in Aberdeen, UK, is distributed by a workforce selection process (WS) to a number of workers each receiving P'. Each individual worker decides (ID) whether they can contribute a solution by creating a report (S), or by delegating the task to a friend. All individual solutions (S) are aggregated into a set of solutions (S') containing reports of all the travel disruptions in Aberdeen.

and evaluating quality of generated results. To address these issues it is necessary to guide various reasoning about the operation of such systems. We argue that this reasoning can be supported by recording the provenance of SC execution (i.e. what happened during the execution) as this will increase its transparency. For example, such a provenance record can enable the system to reason about: workers' motives (e.g. to receive a reward); the steps that were taken in the process of the execution (e.g. delegation of a task to a friend); and which worker performed each part of the computation. Figure 2 shows a provenance record that could be generated during the execution of the SC outlined in the grey area of Figure 1. This describes the activities performed throughout the computation (e.g. delegating the task to a friend), and the agents (Peter and Bob) associated with those activities. Furthermore, linking to the SC executable specification in the provenance record will provide access to the associated social properties. We anticipate that parts of this provenance information (e.g. the elements in Figure 2 with single solid line) can be described using a generic provenance model such as the Prov-DM[3], the W3C recommendation. However, we argue that such a record should also include information about workers attributes, such as motivations, skills and capabilities (their means or ability to apply skills) at the time of the task execution. At this stage, it is unclear how these should be represented using existing provenance models. Linking to the social properties allows monitoring of the system to ensure the behaviours associated with the specification actually occurred during the execution. We argue that recording the provenance of a SC in this way using Semantic Web technologies will enable enhanced, automated decision-making processes that consider, for example, workers past activities, motivation and capabilities, and their compliance with social properties.

[3] www.w3.org/TR/prov-dm/

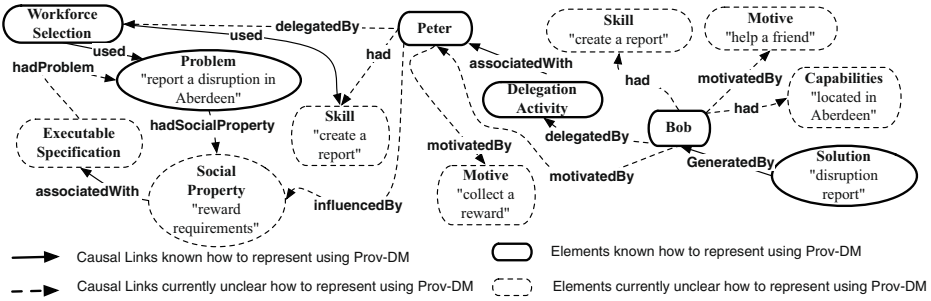

Fig. 2. An example provenance record describing the process of workforce selection and a worker's subsequent delegation to a friend. In this example Peter was selected to perform the task of creating a disruption report in Aberdeen, UK. A social property (defined earlier) influenced Peter's motive to delegate his friend Bob, who lived in Aberdeen. Peter motivated his friend to provide a solution.

The remainder of this paper outlines the research problem and its relevance to the Semantic Web community. We then discuss related work, our hypothesis, research questions, approach to answering those questions, and evaluation plans.

1.1 Problem Statement

We are exploring the suitability of current provenance models, such as Prov-DM and Open Provenance Model (OPM)[4], in the context of SC. For example, can these models capture all the elements of a SC (e.g. capabilities, motives, social properties) or are extensions required to capture them? How can provenance be used in automated decision-making processes such as workforce selection? A provenance model would be required to describe what happened during the execution, and be able to represent the executable specification to allow it to be referenced by the provenance record. The investigation of these issues will involve designing a provenance model that meets these requirements. To explore the benefits of using provenance in automated decision-making (e.g. reducing time or costs of SC), we will investigate how provenance can support two key processes, namely the assessment of worker's trustworthiness and workforce selection. For this purpose, we consider trustworthiness as an attribute associated with a worker that describes assurances such as reputation, or skill level. The results of trust assessments can be used as part of workforce selection. However, such decisions must also rank workers based on trust and match them to a specific task within a SC.

1.2 Relevance

Hendler and Berners-Lee [3] previously noted the fundamental role of Semantic Web technologies in supporting systems driven by SC. They also argue that

[4] `www.openprovenance.org/`

these systems are currently limited as their functions are largely isolated from one another as, for example, they are unable to easily share data. Semantic Web technologies provide the means to describe structured, machine-readable data, its semantics via ontologies, and to support automated reasoning (e.g. to identify suitable workers based on their skills and previous contributions). Therefore, such technology offers a potential solution when addressing those issues. We argue that the application of Linked Data Principles[5] to publishing the provenance of a SC would enable the provenance record to be integrated with both, provenance records from other systems and datasets published as Linked Data. This in turn, provides additional information that can be used to further support automated decision-making.

2 Related Work

Robertson and Giunchiglia's definition of SC (Section 1) is similar to the concept of a workflow, which can be defined as: "a collection of coordinated activities designed to carry out a well-defined complex process"[8]. Activities in such a workflow can be executed by a human or a machine. However, the majority of research on workflows has focussed on processes involving computational services, and only recently has attention been paid to human-driven processes in a workflow setting. For example, the Crowdlang programming framework [5] aims to support the design process associated with a range of SC systems by specifying their abstract workflow descriptions. The framework, however, does not feature ability to model workers and attributes associated with them.

Schall et. al. [12] propose an extension to a service-oriented execution language BPEL4People[6] to accommodate non-functional attributes of workers, such as capabilities and skill level. Further, Schall et. al. use social network analysis to identify three roles that humans can undertake. These are: workers, supervisors (leaders responsible for managing tasks allocated to a group of workers), and coordinators (who form the interface between task requesters that initiate the SC and supervisors). In their approach, Shall et. al. combine these non-functional attributes and use of social network structure to inform novel approaches to automated task allocation for a social group. Difallah et.al. [2] also describe an approach to allocation of categorized tasks to workers based on their social network activities. Difallah et.al. utilise document-based ranking mechanisms by searching for keywords relevant to a particular category topic in documents that worker tagged as of interest to them on, for example, Facebook [2].

Provenance has also been highlighted as having a role in trust assessment of agents [7,10]. These are commonly based on considering an agent's reputation and quality of service. Here, provenance provides additional contextual information that is particularly important when performing trust assessments on, for example, conflicting data or data of unknown origin [10].

[5] `http://www.w3.org/DesignIssues/LinkedData.html`
[6] `http://docs.oasis-open.org/bpel4people/bpel4people-1.1-spec-cd-06.pdf`

While provenance has received significant interest from the workflow community, however, some limitations of existing provenance models when applied to workflows have been identified. Missier et al. [6] addressed a limitation of Prov-DM in modelling a process structure (i.e. the exact steps that were performed during an activity[7] execution) in scientific workflows. They proposed the D-Prov extension to Prov-DM, enabling the description of sub-activities of an activity. Pignotti et al. [9] highlighted the importance of capturing constraints and goals (e.g. intent) of an agent controlling a process in a scientific workflow. They have extended OPM and developed their extension's ontological realisation.

Our approach builds on existing work and proposes the use of provenance to support automated trust assessments of workers and workforce selection. A provenance record provides a novel way of obtaining information that can be used to inform decision-making in SC. In addition, our approach provides the ability to assess workers on new, previously unrecorded information, such as social properties. Recording provenance as linked data improves the reusability of such information and enables the integration of provenance records from other systems. This increases the amount of information available to decision-making processes in SC.

3 Hypothesis

In Section 1 we introduced our plan to investigate provenance in the SC context. To frame our research we have formulated following hypothesis:

Recording the provenance of a social computation can inform decision-making processes about aspects of that and future social computations. Evaluation of this hypothesis will involve realising a provenance model that is capable of capturing the provenance of a SC and its use in a SC system. Additionally, the provenance will be used to inform automated workforce selection and trust assessments of workers to demonstrate the use of provenance in decision-making associated with SC.

3.1 Research Questions

We have identified five research questions related to our hypothesis. **Q1: What are the requirements for capturing the provenance of a SC?** SC integrates humans as part of a wider computational process and relies on various social properties, which affect this computation. We will investigate if the requirements for provenance of a SC execution differ from those of existing provenance models. **Q2: Can a provenance model be developed in order to satisfy these requirements?** Satisfying the requirements identified in Q1 using an abstract formal model will provide means for its ontological realisation. **Q3: Can a framework be realised to demonstrate the practical utilisation of this model?** This investigates the implementation of a computational

[7] Prov-DM notation for a high level description of a process.

framework to capture the provenance of a SC using the provenance model from Q2. **Q4: How can the provenance of a SC be used to support trust assessments of workers?** We will explore how the provenance record of a SC can inform reasoning about a worker's trustworthiness to perform a task. This will include information such as, past interactions with the system (worker's reputation) and associated contextual information (e.g. to determine worker's motivation). **Q5: How can the provenance of a SC be used to guide workforce selection?** We will investigate how the attributes of a worker extracted from provenance records of previous SCs can be used to determine a suitable worker for a particular task within an executable specification.

4 Approach

To address our research questions, we will first produce and analyse a series of use cases[8] to determine the requirements for provenance in SC. The structure of use cases follows the template format of W3C Provenance Incubator Group[9] and includes: background and current practices, description of related provenance dimensions[10], goals, use case scenario, and problems and limitations in achieving goals defined by the use case. The aim is to identify a set of requirements for provenance in SC that can be compared to the requirements satisfied by existing provenance models. In cases where existing models do not satisfy our requirements, we will determine and define the necessary extensions. The new provenance model will be realised in an ontology that will be used to record provenance in a computational framework we will develop to evaluate our hypothesis. This framework will operate on top of existing SC platforms, such as CrowdFlower, or custom built applications (e.g. a travel disruption app), capturing and facilitating the use of provenance in these environments. To date, we have built a simple framework that consists of a number of RESTful services providing functionality for data retrieval, creation, and maintenance for the travel disruption app. Data is stored in a triple store and accessed via a SPARQL[11] endpoint.

Further, we will design models for trust assessment and workforce selection, based on existing work with the incorporation of additional attributes captured in a provenance record. The trust model, will utilise the agent's motivation (e.g. to receive a reward) captured during previous task to predict their future behaviours in similar tasks. For example, trusting a worker that previously provided a solution or delegated a friend who provided a solution. In addition, we plan to explore the capabilities of D-Prov (discussed in Section 2) to better document

[8] To the date we have identified the following use cases: compliance with SC executable specification, assessing social properties, generation of worker profiles, capturing worker capabilities, capturing worker motivation, capturing task execution, and identifying workers roles

[9] `http://www.w3.org/2005/Incubator/prov/wiki/Use_Case_Template`

[10] `http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Dimensions`

[11] `http://www.w3.org/TR/rdf-sparql-query/`

how a worker performed a task. Such information can be used, for example, to calculate workers reputation (e.g. their compliance with the expected behaviours of the executable specification). In terms of workforce selection, we will design mechanisms to classify the different roles (such as those identified by Shall et. al. [12]) workers can undertake during the computation based on their previous behaviours. The workforce selection model will also feature a ranking mechanism of workers based on previous research in the expert finding domain [1]. We will start by exploring two common approaches: candidate-based and document-based. In the first approach, profiles of experts (e.g. workers) are constructed and used to rank the suitability of workers for a given task. The second approach ranks the documents (e.g. tasks described in relevant provenance records) given the query (e.g. a request containing keywords describing the topic a desired expert is sought for) and then infers the relationships between experts and those documents [2,1]. We will explore the possibility of searching provenance records (e.g. matching a current task to provenance records featuring agents performing similar tasks). We will investigate how information about worker attributes captured in these provenance records can be used to determine relationships between those workers and their level of expertise for a particular task. Finally, we will implement trust assessments and workforce selection processes within our framework.

4.1 Reflections

The approach described in this paper addresses the issues relating to the capture and use of provenance of a SC in such a way that would be beneficial to administrators, data consumers, and workers in SC systems. If the approach is successful, our framework can be used to develop new techniques for managing data and workers within such systems. We have outlined our intentions to build on previous research in areas of provenance, workforce selection, and trust assessment. If unsuccessful, our research will identify challenges (practical or theoretical) that need to be addressed before such an approach becomes possible.

5 Evaluation

To evaluate our hypothesis we will first use our provenance model to determine if provenance of SC execution can be captured theoretically (e.g. by applying to a use case scenario). The provenance framework (Section 4) will be used to evaluate the practical implementation of this model. The framework will be used to capture provenance for a number of experimental SC executions. If successful, this will show that provenance can be captured for the execution of SCs. Further, analysis of the generated provenance records will provide an understanding of the characteristics of SC provenance (e.g. record contains lots of activities but very few motivations).

We will then design experiments to evaluate the support provided by the trust assessment and workforce selection processes. Initially, SCs will be executed

repeatedly, without any decision-making support. The SCs will then be repeated with the decision-making support. Evaluating the utility of the support will include comparison of quantitative figures (e.g. number of participants, time required to complete the task) and qualitative analyses of the generated results (i.e. a result generated without the support is similar or equivalent to a result generated with the support). We expect to demonstrate that computations with the decision-making support will show differences in the amount of workers, time, and reward required to motivate workers, while preserving or improving the standard of the results.

References

1. Balog, K., Fang, Y., De Rijke, M., Serdyukov, P., Si, L.: Expertise retrieval. Foundations and Trends in Information Retrieval 6, 127–256 (2012)
2. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Pick-a-crowd: tell me what you like, and i'll tell you what to do. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 367–374 (2013)
3. Hendler, J., Berners-Lee, T.: From the semantic web to social machines: A research challenge for AI on the world wide web. Artificial Intelligence 174(2), 156–161 (2009)
4. Malone, T.W., Laubacher, R., Dellarocas, C.N.: Harnessing crowds: Mapping the genome of collective intelligence. MIT Sloan Research Paper No. 4732-09. SSRN, `http://ssrn.com/abstract=1381502`
5. Minder, P., Bernstein, A.: Crowdlang: a programming language for the systematic exploration of human computation systems. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 124–137. Springer, Heidelberg (2012)
6. Missier, P., Dey, S., Belhajjame, K., Cuevas-Vicenttin, V., Ludaescher, B.: D-prov: extending the prov provenance model with workflow structure. Technical report, School of Computing Science, Newcastle University (2013)
7. Moreau, L., Missier, P.: Prov-dm: The prov data model. W3C Recommendation (April 2012), `http://www.w3.org/TR/prov-dm/`
8. Mukherjee, S., Davulcu, H., Kifer, M., Senkul, P., Yang, G.: Logic-based approaches to workflow modeling and verification. In: Logics for Emerging Applications of Databases, pp. 167–202. Springer (2004)
9. Pignotti, E., Edwards, P., Gotts, N., Polhill, P.: Enhancing workflow with a semantic description of scientific intent. Journal of Web Semantics: Science, Services and Agents on the World Wide Web 9, 222–244 (2010)
10. Rajbhandari, S., Contes, A., Rana, O.F., Deora, V., Wootten, I.: Trust assessment using provenance in service oriented applications. In: 10th IEEE International Conference on Enterprise Distributed Object Computing Workshops, EDOCW 2006, p. 65. IEEE (2006)
11. Robertson, D., Giunchiglia, F.: Programming the social computer. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371(1987) (2013)
12. Schall, D., Satzger, B., Psaier, H.: Crowdsourcing tasks to social networks in BPELl4people. World Wide Web, 1–32 (2012)