Big Data Research 2 (2015) 53–58

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Big Data for Open Digital Innovation – A Research Roadmap ☆

Sabine Brunswicker [a], Elisa Bertino [b,*], Sorin Matei [c]

[a] Research Center for Open Digital Innovation and Department of Technology Leadership and Innovation, Purdue University, United States
[b] Cyber Center and Computer Science Department, Purdue University, United States
[c] Brian Lamb School of Communication and Cyber Center, Purdue University, United States

ARTICLE INFO

Article history:
Received 29 August 2014
Accepted 4 January 2015
Available online 11 February 2015

Keywords:
Data cyberinfrastructure
Data linkage
Digital collaboration
Digital innovation ecosystem

ABSTRACT

Digital technologies have fundamentally altered the nature of organizing innovation and production leading to open collaboration ecosystems. Individuals self-organize in open, voluntary technology-enabled collectives to share their enhancements to the data or collaborate on analyzing, disseminating, or leveraging the data for many applications, from enterprise computing to mobile, consumer oriented applications. 'Big data' is an increasingly important 'engine' to better understand the complex 'nervous system' of open collaboration. However, we need to equip open collaboration researchers with new datasets that span different contexts, as well as novel computational models and analytical techniques. In this paper, we will elaborate on research questions concerning open digital collaboration and derive the data analytical challenges that need to be addressed to answer these research questions.

1. Introduction

Digital technologies have changed the fabric of organizations, triggering novel organizational forms. Innovation and production are not confined to established organizations with clear cut boundaries. Individuals self-organize in open, voluntary technology-enabled collectives to share data and knowledge and to jointly create novel solutions for a bewildering array of applications. With these geographically dispersed groups the idea of open source has moved beyond open source software. Manifold types of these collective forms of innovation and production have emerged in which a large number of actors interact and create goods across multiple platforms, contexts, and timelines. These systems increasingly relate to socially significant domains such as health support or eScience. We refer to them as 'open collaboration' systems for innovation and production [24,7,11] and argue that they are dynamic sociotechnical systems for two reasons: (1) they are fluid and (2) work, organization, and technologies are intertwined within them [31]. Wikipedia, Amazon review systems, science network NanoHub.org, and CancerCare.org are just a few examples of open collaboration. Their emergence stimulated researchers to study their nature.

However, most contributions fail to move beyond existing theories and routine application of research tools to tackle the dynamic and sociotechnical reality of open collaboration [13,29]. We argue that 'big data' is an increasingly important 'engine' to make this turn and to better understand the complex 'nervous system' of open collaboration. The interaction in digital environments creates a gigantic stream of behavioral data that provide novel research opportunities to move beyond outdated theories. To do so, we need to equip researchers of open collaboration with new datasets about dynamic sociotechnical processes that span different contexts and users, with novel analytical techniques, and with an efficient and effective research infrastructure to support the development of novel empirically grounded theories and predictive models.

1.1. Open collaboration – an emerging research area of high significance

Open collaboration (OC) relies on a large number of goal-oriented yet loosely coordinated participants, who interact to create a product (or service) of economic value, which is made available to contributors and non-contributors alike. Indeed, OC will have significant economic as well as social impact [24,2,20]. Scholars in social and behavioral science and economics have started to address the emerging phenomena from different angles and present relevant empirical insights. For example, research drawing upon network theory tackles structural characteristics of large OC networks and also presents new insights in the role of authority, reputation, and trust in OC [28]. Contributions in organizational studies and information systems have revealed deeper insights into the factors that motivate individuals to contribute, and the social value they might have [13].

☆ This article belongs to Visions on Big Data.

* Corresponding author.
E-mail addresses: sbrunswi@purdue.edu (S. Brunswicker), bertino@cs.purdue.edu (E. Bertino), smatei@purdue.edu (S. Matei).

http://dx.doi.org/10.1016/j.bdr.2015.01.008
2214-5796/

1.2. Research gaps and challenges

While this work on OC makes important contributions, major significant gaps remain in understanding, explaining, and predicting the dynamic and sociotechnical nature of OC. Our own research and results from an expert workshops with a community of social science researchers performed in the KredibleNet project [29,26] suggest three research major areas that call for a multidisciplinary research approach to move beyond a few outdated theories, a snapshot of data without context, and routine tools:

1. *Capturing the multilevel, multimodal and the dynamic nature of roles in OC through multi-modal and pattern data*. OC often implies the interaction of a very diverse set of actors embedded in different contexts, technological platforms and subpopulations. OC is a multimodal and pluri-actor process connecting people through the products and services they create [3,10,31]. Such multi-modal networks form around ties between individuals, their resources, and multiple goods, rather than social relationships [29]. The phenomenon of OC is not confined to one context as individuals are embedded in different contexts and subgroups. Thus, it is crucial to consider linked datasets covering different perspectives and communities hosted on different technological platforms or connected through different goods. In addition, roles are dynamic, fluid, and often enacted in the moment. Thus, a simple structural perspective is not enough. Novel analytical techniques that capture dynamics and consider the interrelation between individual behavior, goods, and group activities offers create opportunities to tackle this gap and understand the inner working of OC.
2. *Acknowledging and unwrapping the constituting role of technology in OC through novel data and analytics*. Researchers regularly 'blackbox' technology even though it holds a constituting role [6,27]. A deeper understanding is thus required of how technological features act as "shapers" of behavior, and enable them to solve complex social problems or create novel services. Indeed, existing OC environments make use of sophisticated features such as recommendation systems (e.g. Amazon review), or visualization techniques (e.g. tag clouds, network visualization) but little is known about how different descriptive or predictive features shape collaboration. Neither do we have sufficient data and analysis about user and technology interaction nor do we sufficiently understand their effect. We require novel data-driven research that captures micro-level data about technologies in different contexts and environments in order to develop novel theories, explanations of collaboration and innovation in OC.
3. *Unpacking the dynamic drivers of performance and sustainability of OC through novel computational models and analytics*. OC has significant implications on established assumptions about organizational forms and their performance implications for firms, society, nations, and beyond. With the increasingly digital economy, the open and collaborative models become economically more viable [2]. At the same time, there is insufficient understanding on what ensures sustainability of OC and the dynamics that drive innovation and performance within it. The availability of longitudinal data and novel computational models that unpack the complex dynamic processes within OC would provide novel insights about how such systems evolve and sustain themselves. Today, we lack a deep understanding of the dynamics at multiple levels.

It is critical to address these challenges and opportunities. However, we need to encourage researchers to take turn and move beyond a few theories, a snapshot of data, and standard research tools. Thus, we need new datasets about dynamic sociotechnical behaviors and novel tools and an efficient and effective research infrastructure to break new grounds in explaining and predicting OC and its impact on innovation.

1.3. Data, infrastructure and novel computational models and analytical tools

Addressing the gaps and data opportunities mentioned above – and the manifold research questions that may arise within them – requires novel datasets on OC embedded in a collaborative infrastructure, and a portfolio of research tools and computational models to analyze this data. There are three major building blocks:

1. *Datasets*: There are today many datasets that can be leveraged for research in the area of OC, including existing and processed data on Wikipedia, open source software development platforms, OC systems forming around platforms like OpenData.gov, as well as data on virtual science infrastructures like NanoHub.org, and open innovation networks like Ninesights. It is crucial however that new datasets be made available, possibly through derivation from existing datasets by mining, experiments, novel processing, and existing and new datasets be linked to create context-rich datasets.
2. *Computational and analytical tools*: There is today a large number of such tools, including metanetwork models, network discovery, dynamic and predictive statistical network analysis, genetic computation, network analysis algorithms, agent-based simulations, sequencing analysis and statistical prediction, event study tools, and collaboration and visualization tools. It is critical however that such tools be easily integrated and made available on unified digital platforms.
3. *Collaborative cyberinfrastructure*: Their goal is to serve as a virtual living lab for experiments, offering the community to build capacity, to share data and results, and communicate findings seamlessly across different media.

1.4. Paper organization

In what follows we first discuss the major research themes in OC. For each such theme we identify guiding research questions and data analytic challenges. We then discuss the relevant features of toolkits needed to address such data analytic challenges. We will then outline a few concluding remarks.

2. Research themes in open collaboration

Fig. 1 presents our framework that proposes a dynamic sociotechnical system perspective to a roadmap towards a multidisciplinary research on OC and the identification of data analytic challenges. Our framework conceptualizes OC systems as a sociotechnical system which subsume loosely individuals which freely contribute to develop novel goods, with particular resources and technological features [24]. It addresses three major research dimensions, namely (1) multi-modal and dynamic roles, (2) technological affordances, and (3) the performance and sustainability of the overall dynamic system (as well as the actors and goods within them). These three dimensions will guide our research roadmap.

2.1. Thematic area 1: multilevel, multi-modal and dynamic roles

In OC, individuals 'take' or 'make' different roles which shape the products and goods that are created within OC as well as the adoption and usage of them. Due to loose coordination of OC, such roles are usually not assigned through formal governance but are achieved, emerge, or are even enacted in the moment in