Requirements Engineering (2020) 25:115–130
https://doi.org/10.1007/s00766-019-00310-3

ORIGINAL ARTICLE

A method for analyzing stakeholders' influence on an open source software ecosystem's requirements engineering process

Johan Linåker[1] · Björn Regnell[1] · Daniela Damian[2]

Received: 12 January 2017 / Accepted: 12 April 2019 / Published online: 19 April 2019

Abstract
For a firm in an open source software (OSS) ecosystem, the requirements engineering (RE) process is rather multifaceted. Apart from its typical RE process, there is a competing process, *external* to the firm and inherent to the firm's ecosystem. When trying to impose an agenda in competition with other firms, and aiming to align internal product planning with the ecosystem's RE process, firms need to consider who and how influential the other stakeholders are, and what their agendas are. The aim of the presented research is to help firms identify and analyze stakeholders in OSS ecosystems, in terms of their influence and interactions, to create awareness of their agendas, their collaborators, and how they invest their resources. To arrive at a solution artifact, we applied a design science research approach where we base artifact design on the literature and earlier work. A stakeholder influence analysis (SIA) method is proposed and demonstrated in terms of applicability and utility through a case study on the Apache Hadoop OSS ecosystem. SIA uses social network constructs to measure the stakeholders' influence and interactions and considers the special characteristics of OSS RE to help firms structure their stakeholder analysis processes in relation to an OSS ecosystem. SIA adds a strategic aspect to the stakeholder analysis process by addressing the concepts of influence and interactions, which are important to consider while acting in collaborative and meritocratic RE cultures of OSS ecosystems.

Keywords Open source · Software ecosystem · Requirements engineering · Stakeholder analysis

1 Introduction

Firms that use open source software (OSS), e.g., as part of their supporting infrastructure, product strategy, or business model, need to consider the requirements engineering process of the OSS itself [1]. This second, *external* to the focal firm, RE process is facilitated by the software ecosystem (cf. OSS community [2]) that surrounds the OSS [3]. Firms that are users of the OSS may also be involved in its development and maintenance and can be considered as members of the ecosystem, as well as stakeholders to the OSS. We refer to Glinz & Wieringa's definition of a stakeholder as *"...a person or organization who influences a system's requirements or who is impacted by that system"* [4]. In our context, we consider a person or an organization as the members of an OSS ecosystem, and the system being the OSS that underpins the ecosystem, using the definition by Jansen et al [3].

RE practices in OSS ecosystem may be described as informal and decentralized. There is often no central repository with requirements defined in the problem space, describing the product of need, along with heavy processes and tools for examining the requirements for completeness and consistency [5]. Instead, RE may be considered as a lightweight and evolutionary process of requirements refinement [6]. Practices such as elicitation, specification, and prioritization overlap and are done collaboratively through iterative and transparent discussions and negotiations including up-front implementations [6–8]. These discussions and implementations of requirements are spread out over a number of requirements artifacts, each with its own repository. Examples of these artifacts (cf. informalisms [7]) include

✉ Johan Linåker
Johan.Linaker@cs.lth.se

Björn Regnell
Bjorn.Regnell@cs.lth.se

Daniela Damian
Damian.Daniela@gmail.com

[1] Lund University, Box 118, SE-221 00 Lund, Sweden

[2] University of Victoria, PO Box 1700 STN CSC, Victoria, BC, Canada

reports in an issue tracker, messages in a mailing list, or commits in a version control system. Prioritization is commonly conducted by stakeholders with central positions in the ecosystem's governance structure [9, 10]. To gain such a position in OSS ecosystems with a meritocratic governance structure, a stakeholder needs to prove merit by being active, contributing back, and having a symbiotic relationship with the OSS ecosystem [11].

Hence, the focal firm is one stakeholder among many within an open and fluctuating population in the OSS ecosystem [12]. This can result in conflicting agendas and lack of control, e.g., in regard to which requirements to be implemented and prioritized, render misalignment with internal RE processes [13], and complicate contribution strategies [1]. The focal firm may, therefore, have to gain the influence necessary to affect the RE process in an OSS ecosystem according to its own agenda.

The Merriam-Webster dictionary[1] defines influence as *"the power to change or affect someone or something"*. In our context, this relates to the power of a stakeholder to change or affect the RE process in an OSS ecosystem. This notion of influence aligns naturally with what defines a stakeholder [4], and as a characteristic enables firms to, e.g., see the requirements in which stakeholders hold a certain interest, and from there be able to create an overview of their agendas in the ecosystem [14]. Further, this understanding enables the focal firm to analyze how these stakeholders invest their resources in order to satisfy their agendas [14]. By also considering other stakeholders' interactions within the ecosystem, firms may identify possible partners and competitors [15]. Moreover, this can help firms to learn how to adapt their own strategies and processes with the OSS ecosystem's and how to build their own influence and position the ecosystem's governance structure [10]. The knowledge output can then be leveraged toward other stakeholders through the politics and negotiations that take place in the ecosystem's RE process [16].

These aspects highlight the importance of stakeholder identification and analysis as input to the continuous and complex decision-making process which RE constitutes [17] by helping to answer questions as which other stakeholders exist in the ecosystem, what are their agendas, and how do they aim to achieve them [14]. However, current practices [18] are not adapted to consider these strategic aspects [19] in the context of OSS ecosystem [1] and its informal and collaborative RE process [6, 7], specifically the importance of understanding stakeholders' influence and interactions. Involved firms are no longer the vantage point, and instead, form part of a larger set of interdependent stakeholders [15]. We address this gap with a design science research approach [20, 21] and define it as a design problem [20]:

DP *How to characterize the influence of stakeholders on the OSS ecosystem's RE process, so that a focal firm can understand other stakeholders' agendas, collaborations, and resource investments in pursuing these agendas?*

The contribution of our work is the proposal of the stakeholder influence analysis (SIA) method. Its aim is to help firms to analyze an OSS ecosystem to identify its stakeholders' influence by the impact they have with respect to the requirements that get implemented in the OSS. We base SIA on social network analysis constructs [22–24] that have proven to be useful in characterizing the influence of stakeholders [15, 25], but also effective when analyzing a firm's participation in OSS ecosystems [25, 26] and requirement-centric stakeholder collaborations [27–29]. An analysis approach used in an earlier reported case study of the Apache Hadoop OSS ecosystem [30] is formalized to consider how requirements may be informally represented in multiple artifacts in decentralized repositories present in OSS ecosystems [6, 7]. The influence analysis is then operationalized with a stakeholder mapping approach based on earlier work [31–33]. To demonstrate SIA's applicability and utility, we present a case study of the Apache Hadoop OSS ecosystem.

The rest of this paper is structured as follows: In Sect. 2, we describe the research approach used in the development of SIA. In Sect. 3, we give a detailed presentation of SIA, while in Sect. 4, we demonstrate its applicability and utility with a case study. In Sect. 5, we discuss alternative approaches to characterize influence and threats to validity. Finally, we conclude the paper in Sect. 6.

2 Research approach

To develop SIA, we used a design science research approach [20, 21] where research is conducted iteratively through design cycles. A design cycle consists of three phases: problem investigation, artifact design, and artifact validation [20]. Below we describe these steps in detail.

Problem Investigation phase: Here, the research goal and the problem context are (re-)analyzed before any artifact is designed, or any improvements implemented [20]. In previous work [30], we explored how centrality measures could be used to characterize the influence of stakeholders within an OSS ecosystem, and how this evolved over time. Findings helped to create an understanding of the problem context and helped define the design problem (**DP**) as stated in Sect. 1.

[1] http://www.merriam-webster.com/dictionary/influence.

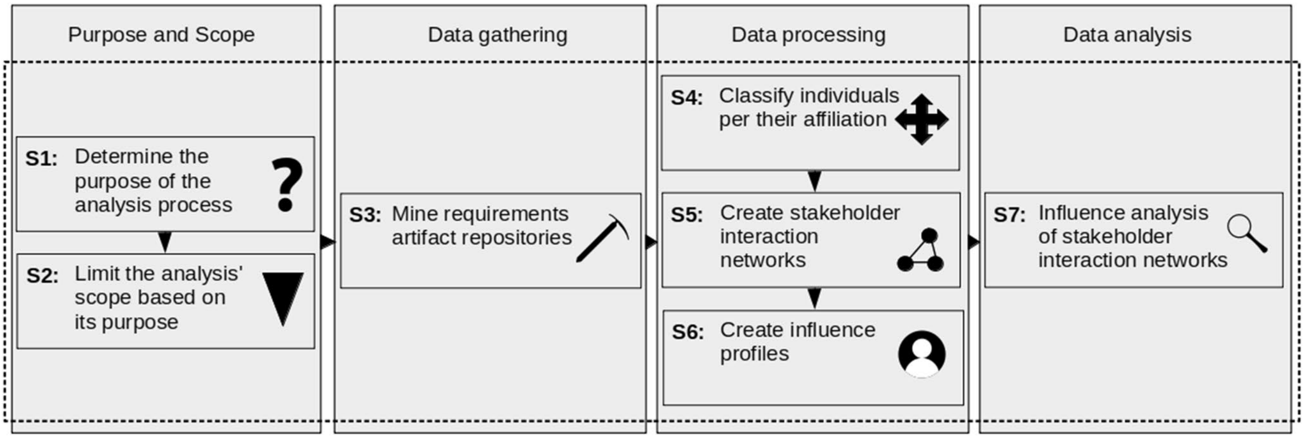

Purpose and Scope
Data gathering
Data processing
Data analysis
S1: Determine the purpose of the analysis process
S2: Limit the analysis' scope based on its purpose
S3: Mine requirements artifact repositories
S4: Classify individuals per their affiliation
S5: Create stakeholder interaction networks
S6: Create influence profiles
S7: Influence analysis of stakeholder interaction networks

Fig. 1 Overview of SIA's seven steps (S1–S7) divided in purpose and scope, data gathering, processing, and analysis

In order to further understand the problem context, a literature survey was conducted to identify related work on:

- the informal and collaborative RE processes within OSS ecosystems (e.g., [2, 6, 7, 9, 12]),
- how awareness of the dynamics behind stakeholder interactions and interrelationships may be used to analyze their agendas (e.g., [1, 10, 14, 15, 18, 34, 35]), and
- how social network constructs may be used to characterize the stakeholders' interactions and influence on the RE process of the OSS ecosystem (e.g., [15, 22–26, 28, 29, 36, 37]).

Surveyed literature provided conceptual foundations, which together with findings from previous work [30] constituted a knowledge base for the artifact design process.

Artifact Design phase: Here, knowledge gained from the previous phase is used as input to the design of an artifact with the hypothesis that it may act as a treatment for the design problem [20]. The stakeholder influence analysis (SIA) method was formalized and structured as seven steps, as presented in Sect. 3 (S1–S7) and in Fig. 1. S1–S2 involves setting the purpose and scope of the analysis. S3 concerns data gathering, while S4–S6 concerns data processing. Finally, S7 regards the analysis of the processed data.

Artifact Validation phase: Here, the previously designed artifact is tested in the problem context in order to evaluate its treatment of the design problem [20]. To test SIA, we apply it in a proof of concept demonstration that it is functional and practical, through a case study on the Apache Hadoop OSS ecosystem (see Sect. 4). It can be seen as an early form of descriptive validation where information from the knowledge base, and detailed scenarios can be used to demonstrate an artifact's applicability and utility [21]. The Apache Hadoop OSS ecosystem was chosen due to the high concentration of firms in the ecosystem, and because it is the Apache project with the highest number of committers.[2] The case study further helped to evolve and refine SIA and its seven steps as can be expected by an iterative design process.

3 The stakeholder influence analysis (SIA) method

SIA aims to help firms involved in OSS ecosystems to structure their stakeholder identification and analysis process systematically when bridging their internal RE process with that of the ecosystem's (see Fig. 1). The focus is specifically on identifying and characterizing stakeholders' interactions and influence on the RE process in the OSS ecosystem. As proposed by Glinz and Wieringa [4], SIA considers both individuals and organizations as stakeholders but primarily from an organizational level, meaning that the individuals in an OSS ecosystem should be aggregated to their organizational affiliation as far as possible. Below, we give a detailed overview of SIA and its seven steps, as outlined in Fig. 1 and Table 1.

Determine the purpose of the analysis process (S1): The first step is to determine what questions are of interest to answer based on the stakeholder analysis, e.g., to identify potential partnerships or competitors, to identify and learn from stakeholders in a certain position, or to identify conflicting agendas in regard to certain requirements.

Limit the analysis' scope based on its purpose (S2): Based on the purpose of the analysis process, limitations may be implied that can affect how the analysis should be narrowed down in terms of what requirements artifacts should be included in the analysis, e.g., is the analysis limited to:

[2] https://projects.apache.org/projects.html?number.

Table 1 Overview of SIA and its seven sequential steps (S1–S7) along with related descriptions and examples

	Step	Description
S1	Determine the purpose of the analysis process	Purpose could include: – Understand how an ecosystem is set up in terms of power structure and general collaboration patterns – Identify potential partners or competitors as input to contribution strategies or collaborations – Identify stakeholders with aligning or conflicting agendas in regard to RE-related activities and negotiations – Identify influential stakeholders to learn from in order to raise one's own influence in the OSS ecosystem
S2	Limit the analysis' scope based on its purpose	Regards boundaries for what data that should be collected and is determined by the purpose of the analysis process. For example, is the interest limited to: – a certain component or set of features of the OSS? – a certain individual or set of stakeholders? – a certain time period or set of releases?
S3	Mine requirements artifact repositories	Refers to the main repositories through which stakeholders interact in regard to the RE process. For example, – IRC or other chat-based communication – Issue trackers – Code review – Software code repository – Discussion boards
S4	Classify individuals per their affiliation	Concerns identification of organizations to which individual developers are affiliated. For example, by – Interacting and studying the communication within an OSS ecosystem – E-mail domain analysis – Heuristically through social media and public electronic sources – Identity pattern matching If no affiliation can be found or exists, the individuals can either be considered either as individual stakeholders or as an aggregated group
S5	Create stakeholder interaction networks	For each requirement artifact repository, a directed and weighted affiliation network is created. Stakeholders are represented as nodes and are connected by edges if they have interacted on a common requirements artifact, e.g., commented on the same issue or mail thread. To reflect investment and influence, edges are weighted based on the size of each stakeholder's participation
S6	Create influence profiles	To characterize stakeholders' influence on the RE process in the OSS ecosystem, a set of network centrality measures are calculated based on the interaction networks, and used to create an overall influence score. Together, they form an influence profile for each stakeholder. The centrality measures include: – Out-degree centrality – Betweenness centrality – Closeness centrality – Eigenvector centrality
S7	Influence analysis of stakeholder interaction networks	Based on influence profiles, stakeholders are ranked on overall influence score, and cross-compared on the centrality measures. Stakeholders of special interest are investigated further in regard to their relationships. With qualitative analysis of stakeholders' agenda alignment with the focal firm's, stakeholder mapping can be used with the influence/alignment matrix. The analysis should be directed by the purpose defined in **S1**

- a certain component or set of features of the OSS?
- a certain individual or set of stakeholders?
- a certain time period or set of releases?

Mine requirements artifact repositories (S3): In the third step, the goal is *to identify and mine the repositories that are mainly used by the OSS ecosystem*. Examples include issue trackers, mailing lists, IRC logs, source code repositories, and code reviews [6, 7]. When these are identified, the repositories should be mined to collect the necessary data. This can either be done either manually or with the help of existing[3] or custom-made tools.

Classify individuals per their affiliation (S4): In the fourth step, the *individuals that are involved in OSS*

[3] See e.g., https://metricsgrimoire.github.io/.

ecosystem need to be classified in regard to their affiliation. This is a necessary step as firm-affiliated individuals may be assumed to represent the agenda of their sponsor or employer [38, 39]. However, not all individuals involved in an OSS ecosystem have to be affiliated and may rather represent their own personal agenda. These affiliations can be identified and triangulated by qualitative and quantitative means, e.g., through involvement and discussions, and by analyzing meta-data from the requirements artifact repositories and cross-checking against other information sources (e.g., social media and electronic archives) [30, 40, 41].

If no affiliation can be found or exists, the individuals can either be considered as individual stakeholders or as an aggregated group. *For example*, say, John, Mark, Lucy, Kate, and Mary are involved in the Apache Hadoop OSS ecosystem as developers. John and Kate work for a firm called Hortonworks and therefore have a common agenda. They are therefore aggregated and viewed as one stakeholder represented by the firm Hortonworks. Mark, Lucy, and Mary are all independent with the difference that Lucy is a relatively active user in the ecosystem, while Mark and Mary are more involved on a hobby basis. Lucy could, therefore, be seen as an independent stakeholder, while Mark and Mary could be aggregated to one group of hobbyists and be considered as one stakeholder. This type of classification and separation is rather subjective and needs to be done on a case-by-case basis for each ecosystem.

Create stakeholder interaction networks (S5): In the following step, *an interaction network for each requirements artifact repository needs to be created* in order to visualize the interactions between stakeholders. To create these networks, the interactions between the stakeholders to the requirement artifacts within a requirements artifact repository must be identified. As an example, consider a number of individuals (stakeholders) that discuss the need for as well as potential implementations of a new feature in an OSS project. The feature request is represented by an issue (requirements artifact) on the OSS ecosystem's issue tracker (requirements artifact repository). The discussions (interactions) between the individuals concerning the feature's evolution and refinement are recorded and persisted in the issue. This continuous discussion may be referred to as an "event" in social network theory [22]. The individuals partaking in the discussions may be referred to as "participants" of the same event [22].

These events and their participants can furthermore be represented by networks of actors. Two actors within a network are connected by an edge if they have participated in the same event (as a network may include several events). If a network was created based on the previous example, all individuals who partook in the discussion of the issue would be represented by an actor in a network with an edge connecting each one of them. If there was a related discussion of the feature on the OSS ecosystem's mailing list, a similar network may be created based on the concerned mail thread. The two networks could then be analyzed in conjunction to get a more complete overview of the stakeholders to the requirement and their interactions (cf. requirement-central networks [27]).

In a similar fashion, sets of requirements may be analyzed by aggregating requirements artifacts in a repository to a network. Returning to the example, a network could be created that included all of the issues in the issue tracker that are related to a certain release, created in a certain time span, or belonging to the same sub-module. A corresponding network could be created based on the mailing list given that the same conditions apply. By creating corresponding networks of all the relevant requirements artifact repositories, the analyst may get a complete overview of what stakeholders that are involved and how they interact.

It should be noted that one stakeholder's participation in the event (e.g., RE-related discussions of an issue) may be of a relatively different size than the other stakeholders'. A stakeholder with a higher degree of participation may be considered to have a larger investment and interest in the event. These differences in the investment of time and resources need to be considered in order to give a fair view of a stakeholder's stake in a requirement. The relative size of the investment also helps to give a fairer dataset when doing an influence analysis of the interaction networks. As suggested by Orucevic-Alagic et al. [25], weights can be calculated to describe the relative size of the participation to an event.

Following Orucevic-Alagic et al. [25], for a set of stakeholders $V = \{v_1, v_2, \dots, v_k\}$ and a set of requirements artifacts (events) $U = \{u_1, u_2, \dots, u_m\}$, we define a weight W of an edge between one stakeholder v_i and all other stakeholders that collaborate on an artifact u_t as:

$$W(v_i, u_t) = \frac{X(v_i, u_t)}{\sum_{c=1}^{k} X(v_c, u_t)}$$

where $X(v_i, u_t)$ denotes the number of times a stakeholder v_i has participated in the collaboration on the requirements artifact u_t.

Continuing from Orucevic-Alagic et al. [25], this means that the weight of the edge $W(v_i, v_j)$ for all requirements artifacts that two stakeholders v_i and v_j have collaborated on together equals:

$$W(v_i, v_j) = \sum_{t=1}^{m} W(v_i, v_j, u_t)$$

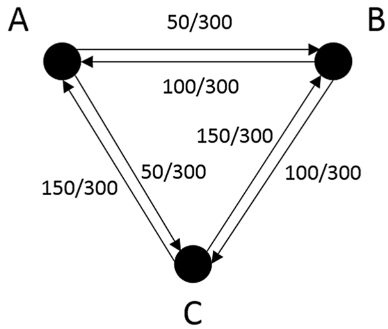


Fig. 2 Example of network with three stakeholders v_A, v_B and v_C, and connecting weighted edges. Adopted from [30]

As an example, when creating an interaction network based on an issue tracker, each issue represents a requirements artifact and number of posted comments may represent the size of participation (X) of a stakeholder. Given that three stakeholders v_A, v_B and v_C comment on the issue, they are all considered as actors in a network with edges connecting them. The weights would, therefore, consider the relative number of comments of each stakeholder as the size of their participation. Say v_A commented 1, v_B commented 2, and v_C commented 3 times. This results in the edge weights:

- $W(v_A, v_B)\&W(v_A, v_C) = 1/5$
- $W(v_B, v_A)\&W(v_B, v_C) = 2/5$
- $W(v_C, v_A)\&W(v_C, v_B) = 3/5$

If two stakeholder participated in an equal number of times, the size of each participation can be made further fine-grained. In another example, when considering an interaction network based on patches submitted to a software code repository, the size of a stakeholder's participation (X) can be quantified with the number of changed lines of code (LOC) of its patches. A simplified example is shown in Fig. 2 where three stakeholders v_A, v_B, and v_C each created various number of patches that were contributed to a certain issue. v_A's patches contain 50 LOC in total. v_B's patches contain 100 LOC in total, while v_C's patches contain 150 LOC in total. Aggregated, 300 LOC were contributed to the issue. Resulting in the following edge weights:

- $W(v_A, v_B)\&W(v_A, v_C) = 50/300$
- $W(v_B, v_A)\&W(v_B, v_C) = 100/300$
- $W(v_C, v_A)\&W(v_C, v_B) = 150/300$

By constructing this kind of networks (i.e., weighted and directed affiliation networks [22, 23]), stakeholders' interaction in an OSS ecosystem's RE process may be visualized on different abstraction levels across the different requirements artifact repositories identified in **S3**.

Create influence profiles (S6): In a network, a stakeholder is more prominent if it has a central position with edges that make it extra visible and important to others [36]. In social networks, centrality measures are commonly used to analyze an actor's position and prominence relative to others [22]. Faust [23] breaks down the notion of centrality into how an actor is central given that they are active in the network, can communicate with others in the network efficiently, are able to mediate and control flows of information between others in the network, and have relationships with others that are central. These four aspects, respectively, relate to the centrality measures of out-degree, betweenness, closeness, and eigenvector centrality. SIA uses these measures as the foundation for analyzing the influence of stakeholders.

These four centrality measures can be adapted in different ways to provide further facets of influence in regard to the interaction networks. As the interaction networks are described in **S5**, the edges that connect two stakeholders have weights attached to them. These weights allow the measures to take account of the relative size of each stakeholder's participation of the requirements artifacts on which the network is based on. For example, out-degree centrality (see Table 2) refers to the sum of weights attached to out-going edges from the focal stakeholder and its adjacent stakeholders [42]. This gives an overall number in regards to the size of the focal stakeholder's participation in the set of requirements artifacts covered by the network. A high out-degree centrality may indicate that the focal stakeholder has a high influence on its adjacent neighbors and is good at communicating its views relative others in the network [25]. However, this way of measuring out-degree centrality does not provide information about the total number of connections of a stakeholder, which may better show the number of collaborations and opportunities to spread one's opinions [43]. Hence, we recommend that the proposed centrality measures are used both in the case where the edges have the relative weights attached to them, and in the case where they are considered either present or not [44].

In Table 2, we describe the foundation for these measures and how they may be interpreted in terms of a stakeholder's influence in the RE process of an OSS ecosystem.

As described by Faust [23], centrality may be broken down into multiple aspects. Centrality measures, in turn, use different definitions and sets of criteria in regard to what classifies an actor's position as central. Hence, one measure can present a different social structure than another and different measures provide different perspectives on who are the most active [25]. In smaller and simpler network structures such measures may co-vary, while in larger and more complex networks, they may characterize actors very differently [45].

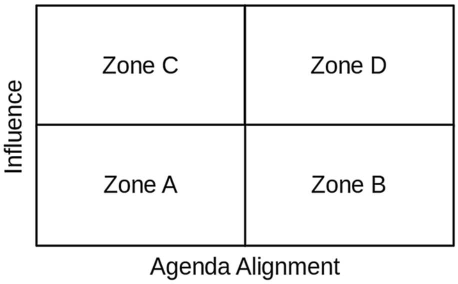


Fig. 3 Influence/agenda alignment matrix to be used for stakeholder mapping. Adapted from earlier work [31]

Therefore, measures presented in Table 2 could be seen as complementary to each other and may be used together to give each stakeholder (v_i) an *influence profile* (IP_{v_i}), a 4-tuple consisting of each centrality measure (i.e., out-degree centrality (ODC_{v_i}), betweenness centrality (BC_{v_i}), closeness centrality (CC_{v_i}), eigenvector centrality (EC_{v_i})).

$$\mathrm{IP}_{v_i} = \left(\mathrm{ODC}_{v_i}, \mathrm{BC}_{v_i}, \mathrm{CC}_{v_i}, \mathrm{EC}_{v_i}\right)$$

Such a profile can then be used when analyzing a stakeholder's interaction network in step **S7**. For example, a stakeholder in a certain interaction network may have

- a high ODC indicating a high activity with many collaborations,
- a low BC indicating that the stakeholder does not have a broker's position, but
- a high CC indicating that the stakeholder can more easily reach out with its communication, and
- a high EC indicating that the stakeholder knows other influential stakeholders.

When comparing stakeholders and their influence profiles, it would be convenient to define, for each stakeholder v_i, an aggregated *influence score* IS_{v_i}. Such a score could be used to divide stakeholders in to two groups, those with a high and low level of influence (see upper and lower zones in Fig. 3). One way to do this aggregation is to simply add the normalized weights of each element in the profile, resulting in a ratio-scale number between 0 and 1, as given by the formula below, and then group stakeholders based on a threshold, e.g., less than or equal to 0.5 denotes low influence:

$$\mathrm{IS}_{v_i} = \frac{1}{4}\left(\frac{\mathrm{ODC}_{v_i}}{\mathrm{ODC}_{\max}} + \frac{\mathrm{BC}_{v_i}}{\mathrm{BC}_{\max}} + \frac{\mathrm{CC}_{v_i}}{\mathrm{CC}_{\max}} + \frac{\mathrm{EC}_{v_i}}{\mathrm{EC}_{\max}}\right)$$

There are other ways of aggregating the different measures, using, e.g., ordinal-scale ranks, a vector space distance metric (e.g., cosine similarity), a normalized exponential function (softmax), or applying some kind of weighting scheme to reflect, e.g., that centrality is considered more interesting. Another option is to qualitatively compare the IS_{v_i} 4-tuple of measures in combination with some visualization technique, such as spider diagrams or similar. Future work should investigate which aggregation method that would best help to partition the stakeholders into high- and low-level category.

In addition to comparing the stakeholders' influence profiles and overall influence scores within a specific stakeholder interaction network, it is equally important to compare between the networks. For example, if the analysis includes multiple requirements artifact repositories (e.g., issue trackers and mailing lists) or covers multiple releases, these could be cross-compared. A stakeholder may have a high overall influence score in one requirements artifact repository, and less in another. Further, the influence and interactions may shift with time why temporal analysis may give important insights. Also, it may be that one repository is more important than another (e.g., issue tracker over mailing list), as a result, the former should be given more attention in a cross-comparative analysis of a stakeholder.

Influence analysis of stakeholder interaction networks (S7): In the influence analysis, the interaction networks and influence profiles from **S5** and **S6** are used to address the purpose defined in **S1**. First, stakeholders are ranked on their overall influence score to get an overview of the stakeholder population. Stakeholders of interest, e.g., a top list of those most influential, can then be cross-compared based on the centrality measures from their influence profiles, and analyzed in detail, e.g., in regard to their relationships. Table 2 provides descriptions of how the centrality measures may be interpreted in terms of a stakeholder's influence in the RE process of an OSS ecosystem.

As a support in the analysis, and to help address the purpose as defined in **S1**, stakeholder mapping can be applied with the use of an influence/agenda alignment matrix (see Fig. 3). The matrix, based on earlier work [31–33], is adapted to consider the power and politics [14] that play a central part in the RE process of OSS ecosystems [1, 34]. The Y-axis represents the level of influence and the X-axis how well their agenda in the OSS ecosystem aligns with that of the focal firm. Both dimensions range from low to high. The four quadrants Zone A–D in the figure are explained subsequently.

The level of influence of a stakeholder is based on the influence score from **S6**. The threshold for when a stakeholder's influence score ranks as high is set by the analyst in relation to the total number of stakeholders in the network. Agenda alignment, which is the second dimension, is determined by qualitatively investigating the previously identified stakeholders' engagement in the OSS ecosystem, e.g., by reviewing comments made by the stakeholder in the set of issues which the analysis considers (as defined in

Table 2 Network measures described from a general perspective as well as and how they can be interpreted from a RE perspective in regard to stakeholder influence

Measure	Description	Stakeholder Influence Interpretation
Out-degree centrality	Refers to how well connected the focal actor is and considers the out-going edges toward its adjacent actors, where the focal actor is the transmitter (source) for the edges. With weights considered, this measure refers to the sum of weights attached to the outgoing edges of the focal actor [42]. With binary edges considered, this measure refers to the number of outgoing edges between the focal actor and its adjacent actors [44]	Out-degree centrality is generally considered as a measure of activity that can identify "where the action is" and highlight the most visible actor in the network [22]. With weights considered, a high out-degree centrality is an indication of influence on adjacent stakeholders as the focal stakeholder has participated in a large part in the requirement artifacts which they have interacted with [25]. This participation can be viewed as the focal stakeholder's opinions in the RE process of the OSS ecosystem. In both cases of weighted and binary edges, a higher out-degree may also indicate a higher number of options or opportunities for qualitative contacts, i.e., to know the key stakeholders to influence and create traction with on a certain issue. For binary edges specifically, it may further indicate a high level of activity through a number of collaborations, but also to which the focal stakeholder has expressed its opinions
Betweenness centrality	Refers to the extent to which the focal actor lies on the shortest path between pairs of other actors. With weighted edges considered, it refers to the shortest path with the lowest sum of weights [46, 47]. With binary edges considered, it refers to the shortest path in regard to the least number of edges [44]	Betweenness centrality is a measure of control and coordination as it highlights actors who sit on the shortest, and sometimes only, communication paths or resource flow between many others [22]. Hence, stakeholders with a high betweenness centrality may control and coordinate the information flow about requirements, and interactions between other stakeholders. The focal stakeholder could be characterized as having a central position in the ecosystem, e.g., in regard to project management and governance. Others may be dependent on the focal stakeholders to relay the information and to set up connections. Further, the centrality also indicates the ability to act as an intermediary that can influence the content of the information, and whom it reaches and when, to better serve personal priorities. When a stakeholder is the only one, or one of very few, linking two or more parts of a network, they are commonly referred to as brokers as their possibility to influence is very high [24, 28]
Closeness centrality	Refers to the inverse of the sum of the shortest paths from the focal actor to all others in the network. With weighted edges considered, it refers to the shortest path with the lowest sum of weights [46, 47]. With binary edges considered, it refers to the shortest path in regard to the least number of edges [44]. This measure only considers those actors that are connected to the same network as the focal actor [24]. For disconnected actors, the measure in undefined as the distance is infinite	Closeness centrality is a measure of efficiency in contacting others and spreading, but also receiving, information in the network and hence an actors' ability to influence others [24]. Hence, a high closeness centrality indicates that a stakeholder is efficient in spreading and receiving information about a requirement to and from the rest of the network of stakeholders. This efficiency allows the focal stakeholder to more easily communicate its agenda on the requirement and interact with others, e.g., in negotiations and lobbying. The focal stakeholder could, therefore, be characterized as being close to other stakeholders and more independent. This further minimizes the risk of intermediaries influencing the information about the requirement in an unfavorable manner [15]
Eigenvector centrality	Refers to how connected an actor is, similar to out-degree centrality, but considers how well connected the adjacent actors are [48]. The focal actor receives a score based on a sum of its adjacent actors' scores [24]	Eigenvector centrality is a measure of activity and visibility as out-degree centrality, but adds information to whom these attributes connect to. A high value indicates that the actor has important friends who in turn are visible and active [24]. Hence, a high eigenvector centrality indicates that a stakeholder knows and collaborates with other stakeholders who are important and have key positions in the OSS ecosystem [23]. The focal stakeholder is in a position to have a potentially high impact on the RE process in the ecosystem by being able to communicate its agenda to, and influence key actors in the social network [23]

S1 and **S2**). The investigation should seek to answer if the stakeholder and the focal firm want the same thing, and to what extent.

The classification puts a stakeholder into one of four quadrants (A–D) of Fig. 3, each indicating a different relationship and possible engagement that the focal firm should establish and maintain with the stakeholder. Stakeholders with a high level of influence and high level of agenda alignment (Zone D) may pose as (potential) partners, both in regard to general collaboration and RE-related activities and negotiations. Stakeholders with a high level of influence and low level of agenda alignment (Zone C) may pose as the key opponents and may require active engagement in negotiations in the RE process of the OSS ecosystem. Stakeholders with a low level of influence (Zone B and A) may not pose as having high importance, but may still require monitoring as they can move their position with time. Those in Zone B may pose as future collaboration opportunities, while those in Zone A as potential threats.

If competitors are identified among those with high influence, this may signal that they have a high interest in the ecosystem and scope of the investigation. If they are found in Zone D, there might be an opportunity for co-opetition. In either case, whether they have aligning agendas or not, consideration should still be taken to the differential value of what is contributed and how resources are invested. By studying stakeholders in Zone C and D, a focal firm can potentially strengthen its own influence by learning from these stakeholders, in how they invest their resources and with whom they collaborate. This may lead to further collaboration and other potential partners, and how interest may overlap between multiple stakeholders.

4 Case study of Apache Hadoop OSS ecosystem

In this section, we describe a first evaluation of SIA in our design methodology. We demonstrate the applicability and utility of SIA in a case study [49] on the Apache Hadoop OSS ecosystem. The case study takes the perspective of a (fictive) focal firm that provides scalable and secure infrastructure on which Hadoop can be deployed for customers. This is a new product offering, and the focal firm is now interested in becoming active in the Apache Hadoop OSS ecosystem. As they are new to the ecosystem, they want to do an initial stakeholder analysis to see if there are any potential partners to collaborate with, and potentially learn from (**S1**). First, they want to get a general overview of the stakeholder population to see who is present and how the ecosystem functions in terms of the power structure and collaboration patterns. Second, they will look for potential partners among those most influential and investigate how they work, and what interests they have in the ecosystem.

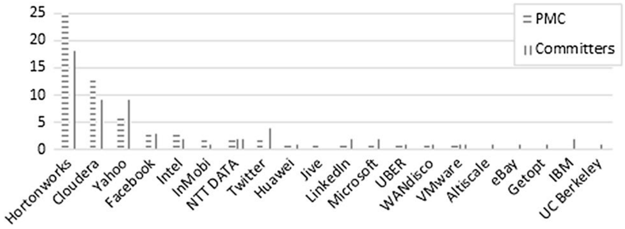


Fig. 4 Number of committers and members in the Apache Hadoop PMC aggregated per firm

The Apache Hadoop project[4] is a widely adopted OSS framework for distribution and process parallelization of large data, originating from *Yahoo* in 2006. The framework consists of four modules: Hadoop Common Modules, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce.

The Apache Hadoop project is part of the Apache Software Foundation which is an umbrella organization for a large number of OSS projects and their ecosystems. A common trait for these projects is the use of meritocracy in terms of culture and governance.[5] This is reflected in the governance structure among the Apache projects, as in Apache Hadoop which is governed by a Program Management Committee (PMC) that consists of representatives from the Apache Software Foundation and of elected members from the project's ecosystem. Further, the PMC members are also classified as committers, i.e., they have been granted write access to the project. A member may be elected as a new committer by the existing ones. Being elected as a committer does, however, not imply membership of the PMC. To become a committer or member of the PMC, an individual need to show merit, e.g., by contributing and actively participating in the development of the project. Hence, power may be earned by showing a long-term commitment and having the competence needed (i.e., meritocracy). In Fig. 4, the distribution of members of the committers and the PMC are presented based on affiliation per firm.

4.1 Overview of stakeholder interaction and influence

To get a recent view on who the most influential stakeholders are, the scope of the analysis is limited to requirements included from release 2.2.0 (15/Oct/13) to 2.7.1 (06/Jul/15) (**S2**). To get a view on both social and technical interaction,

[4] http://hadoop.apache.org/.

[5] https://www.apache.org/foundation/how-it-works.html#meritocracy.

Table 3 Characteristics of comments and patch networks

	Comments network	Patch network
Stakeholders	122	86
Collaborations	1096	260
Per stakeholder	9	3

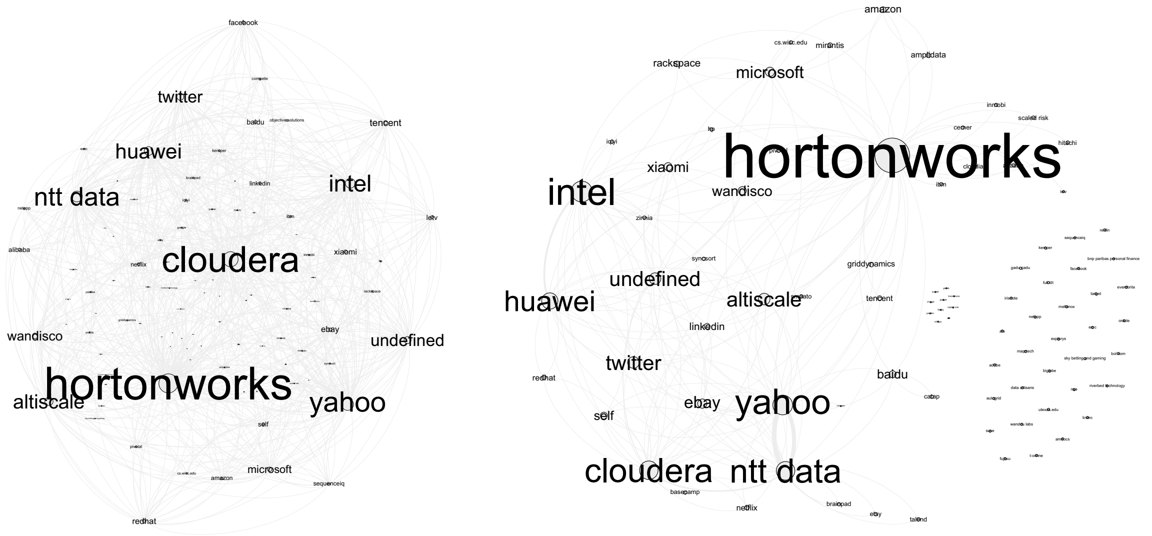
the issue tracker is analyzed in regard to requirements artifact repositories (**S3**). The issues contain both comments (the social dimension) and patches (technical). The patches are committed by authorized users, once they have been approved. To identify the organizational affiliation of individuals that have interacted via the requirements (**S4**), an analysis is done of e-mail sub-domains, complemented with cross-checking against other information sources (e.g., social media and electronic archives) [30, 40, 41]. For a subset of individuals, an organizational affiliation could not be determined. These individuals were aggregated into two separate groups, either independent (if this could be determined) or unidentified.

Creation of Stakeholder Interaction Networks (S5): Based on the scope specified in **S2**, and the repository identified in **S3**, two interaction networks are generated: a *comments network* to include stakeholders who commented on common issues, and a *patch network* to include the stakeholders who contributed patches to the same issues (**S5**). The patch network was presented in earlier work [30], and a similar data collection and cleaning approach were used in order to create the comments network, as is also proposed in SIA (see Sect. 3). The comments network shows activity and collaboration of a stakeholder in regard to the social interaction and discussion that revolves around a certain issue, and the patch network shows same characteristics for a stakeholder in regard to suggesting technical implementations.

In each of the two networks, a stakeholder is represented by a node, and the collaborations between them are represented by the edges connecting the nodes. The comments network consists of 122 stakeholders, compared to 86 stakeholders in the patch network (see Table 3). In both cases, this includes two groups of developers classified as independent or as unidentified. The comments network has a higher degree of collaboration with an average of 9 collaborations per stakeholder, compared to the patch network, which has an average of 3 collaborations per stakeholder. Both networks are visualized on a high level in Fig. 5a, b. Labels are of firms and of relative size to their weighted out-degree, a reason for which only those with the highest values may be readable.

Creation of influence profiles. (S6): To measure the influence of, and collaboration among, the stakeholders (**S6**), two SNA measures were leveraged: weighted out-degree and betweenness centrality. Other centrality measures presented in Table 2 were excluded due to space considerations in this paper. In Fig. 6, the two measures are presented in two separate diagrams. The diagrams contrast the respective measures for the comments and patch networks in regard to the 15 top stakeholders (considering the overall influence score).

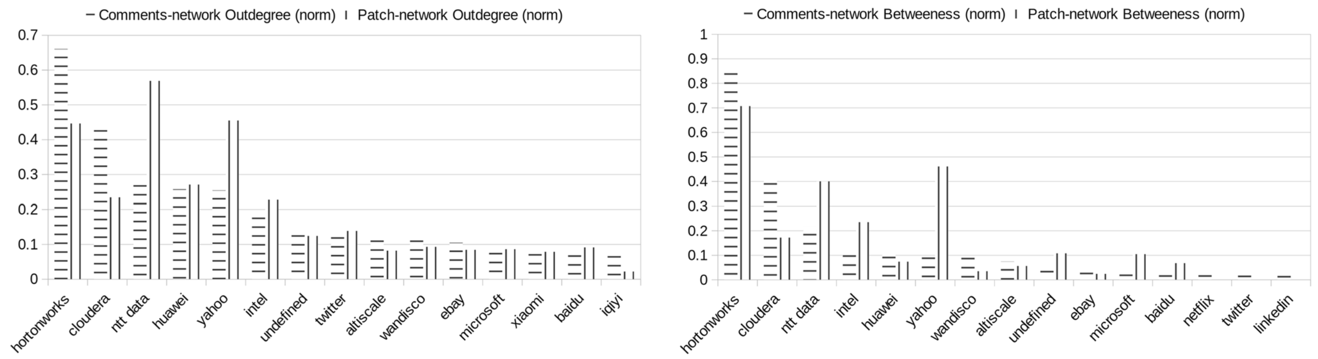
Influence Analysis of the Stakeholder Interaction Networks (S7): As presented in Table 2, the measures measure different aspects of influence and collaboration among



(a) Comments-network.

(b) Patch-network.

Fig. 5 Visualization of the **a** comments and **b** patch networks. Labels are of firms and of relative size of their weighted out-degree to other firms in each network



(a) Out-degree centrality.

(b) Betweeness centrality.

Fig. 6 Visualizations of normalized centrality measures for the 15 top influential firms across the comments and patch networks. Each diagram is sorted in a descending order based on respective centrality measure from the comments network

Table 4 Top ten stakeholders based on influence score based on comment network, considering only the out-degree and betweenness centrality

Stakeholder	Out-degree (norm)	Betweenness (norm)	Influence score
Hortonworks	0.66	0.86	0.76
Cloudera	0.44	0.4	0.42
Ntt data	0.28	0.22	0.25
Huawei	0.26	0.10	0.18
Yahoo	0.25	0.10	0.18
Intel	0.19	0.11	0.15
Undefined	0.11	0.09	0.10
Twitter	0.14	0.06	0.10
Altiscale	0.11	0.07	0.09
Wandisco	0.13	0.02	0.8

the stakeholders. Below, the two measures are compared in regard to the two networks and their stakeholders.

*Out-degree centrality:*Figure 6a illustrates the normalized out-degree centrality which may be considered as rather equal for most stakeholders with the exception of those most influential: NTT Data, Yahoo, Hortonworks, and Cloudera. Both NTT Data and Yahoo have a notably higher influence in regards to technical implementation suggestions, while Hortonworks and Cloudera have a higher influence and activity through social interaction and discussion. Considering the distribution of stakeholders from the different user categories, a heavier representation of product vendors (Hortonworks, Cloudera, and Huawei) can be seen in the top five, in regard to both the comments and patch networks.

Betweenness centrality: In Fig. 6b, it can be seen that the normalized betweenness centrality varies notably between the comments and patch networks for the top stakeholders. Hortonworks has the highest betweenness centrality in regard to both the technical and social aspects and compared to Cloudera and Yahoo, it has double the betweenness centrality in the comments and patch networks, respectively. Contrasting Cloudera and Yahoo, a clear difference in focus and importance is shown. Cloudera values technical implementation suggestions over social interaction and discussion, while Yahoo focuses on social interaction and discussions.

Cross-comparison of centrality measures: To simplify the cross-comparison, the influence score is used to get an overview of the top 10 most influential stakeholders considering the two centrality measures (see Table 4). Comparing the two centrality measures for these ten, both similarities and differences may be found. Although it has higher activity in the comments network, Hortonworks has high influence in regard to both technical and social interaction, if both centrality measures are taken into account. This indicates that Hortonworks has a high impact in regard to what is implemented and how. This firm can be classified as well connected both directly and indirectly and has a good position to act as an authority in regard to information spread and coordination. NTT Data and Yahoo both clearly have a higher degree of activity and influence in the patch network. As with Hortonworks, they also have a similar distribution among both of the two measures. This may indicate that they have a high impact in regard to what is implemented and how, but focus their resources on contributing technical implementation suggestions and solutions. As with Hortonworks, they can be classified as well connected both directly and indirectly, and have a good position to act as an authority in regard to information spread and coordination. Regarding the out-degree centrality, a group of stakeholders forms just below the top.

Considering the influence/agenda alignment matrix (see Fig. 3), these stakeholders could be considered as key stakeholders and qualify for either Zone C or D. They could pose either as potential partners or threats depending on how their agenda aligns. Also, depending on if they are competitors or not, consideration should also be taken when constructing contribution strategies [13]. The focal firm should, therefore,

Table 5 Binary out-degree of Wandisco for Apache Hadoop's four modules. Values aggregated for releases R2.2-2.7 per network type. Relative ranking within parenthesis

	Common	HDFS	YARN	MapReduce
Comments	11 (11/64)	20 (5/48)	4 (32/59)	1 (33/39)
Patches	0	5 (7/24)	0	0

Table 6 Weighted out-degree of Wandisco for Apache Hadoop's four modules. Values aggregated for releases R2.2-2.7 per network type. Relative ranking within parenthesis

	Common	HDFS	YARN	MapReduce
Comments	1.87 (12/64)	2.82 (6/48)	0.73 (19/59)	0.24 (26/39)
Patches	0	2.97 (7/24)	0	0

monitor and form an understanding of how these stakeholders' agendas align with their own.

4.2 Investigating collaborations and agenda of a potential partner

From the previous analysis, the focal firm could identify WANdisco as a stakeholder with a similar business model and a potential partner in terms of collaboration and similar interests (Zone D in Fig. 3). The goal in this second step is to do a more thorough analysis focusing on WANdisco's collaborations and high-level agenda (**S7**).

Looking at WANdisco's influence profile, their overall influence score gives them an ordinal rank of 10 (see Table 4) when analyzing the comments network. They have an equal level of social and technical activity, on similar levels as Twitter, Altiscale, eBay, and Microsoft (see Fig. 6a). They have a relatively high level of control and coordination considering the betweenness centrality. All things considered, they have a relatively high influence and interest in Apache Hadoop, but much lower than the key-stone players, Hortonworks, Cloudera, NTT Data, Huawei, Yahoo, and Intel.

WANdisco entered the Apache Hadoop ecosystem in 2012 by acquiring AltoStar. Their product is a platform that allows for distribution of data over multiple Apache Hadoop clusters, similar to that of the focal firm. WANdisco has 14 active developers in the investigated set of releases in regard to comments and patch contributions. One developer is also a member of the PMC and Committers group.

To learn more about WANdisco's interests in Apache Hadoop and its collaborators, the focal firm investigates if WANdisco has shown a special focus in regard to any of the four modules of Apache Hadoop: Common, HDFS, YARN, and MapReduce (**S2**). The analysis is still focused on requirements included in releases R2.2-R2.7. Regarding **S3–4**, they are identical to the previous example. When creating the interaction networks (**S5**), one patch network and one comment network are created for each of the modules.

Table 7 Top collaborators with Wandisco in the comments network of the HDFS module

Stakeholder	Number of comments	Total number of comments	Weight
Hortonworks	227	1109	0.20
Cloudera	98	663	0.15
Intel	91	679	0.13
Pivotal	42	79	0.53
Yahoo	34	313	0.11

When creating the influence profiles (**S6**), the analysis is limited to examining the out-degree to get a view of their activity and comprehension of their relative influence in regard to the modules. Values for binary and weighted out-degree are presented in Tables 5 and 6, respectively. The former specifically indicates the number of other stakeholders that WANdisco has interacted with, and the latter a better relative measure of their influence. As can be noticed for values regarding the patch network, it can be concluded that WANdisco has a specific interest in the HDFS module. The out-degree values for the comments network further confirm a specific interest in the HDFS module with a relative ranking of 5 and 6, respectively, out of 48. Some interest can also be observed for the Common module.

Regarding collaboration, the analysis is limited to the HDFS module as this is where their main interest of WANdisco lies. In regard to the patch network, there are only five collaborators, as indicated by the binary out-degree in Table 5. These consist of Hortonworks, Huawei, Intel, Yahoo, and Intel. In regard to comments network, WANdisco had interacted with 20 other stakeholders. Out of these, Hortonworks, Cloudera, Intel, Pivotal, and Yahoo were the top five in regard to the number of comments made by WANdisco on common issues, see Table 7. The table further presents the weight of the outgoing edge from WANdisco to each respective stakeholder. In the example of Pivotal, this may be interpreted as WANdisco having made 53 percent of the total number of comments on issues where both WANdisco and Pivotal have collaborated on.

An outcome of this analysis is that WANdisco holds their main interest and invest their resources in the HDFS component, both from a technical and social perspective. Considering the influence/agenda alignment matrix in **S7** (see Fig. 3), a qualitative investigation needs to be performed, e.g., of their comments and code commits, in order to determine e.g., what features they value or prioritize. Such an investigation will help to determine the agenda alignment further and if WANdisco belongs to Zone C or D, i.e., if they make up a potential opponent or partner. Based on their active collaborations, Pivotal should be investigated further in terms of their interest and activity.

5 Discussion

Below we first discuss different alternatives to characterizing a stakeholder's influence, followed by a discussion regarding limitations and threats to validity in our demonstration of SIA's utility in the analysis of the Apache Hadoop ecosystem stakeholder influence.

5.1 Alternatives to characterizing a stakeholder's influence

The three questions stipulated by Frooman [14] highlight the strategic importance of stakeholder identification and analysis: Firms need to identify and characterize present stakeholders in terms of their influence, identify their agendas primarily in terms of alignment with one's own, and how they are planning to achieve it. The latter of the three is important as it informs of a firm's possible strategies for contribution and interaction. SIA helps to address these questions by characterizing the stakeholders' collaboration and influence within the OSS ecosystem. The quantitative outcome which is generated is best complemented with qualitative insights which may be gained through observing or even taking part in the communication of the ecosystem.

In the stakeholder mapping process, which is part of the influence analysis of SIA (**S7**), both the quantitative and qualitative aspects are needed. In the proposed influence/agenda alignment matrix, the influence profiles and influence scores may be used to determine the level of influence. As mentioned in Sect. 3 and the description of (**S6**), the proposed influence score is one approach to get a simplified overview of which stakeholders are the most influential. However, as the different centrality measures provide different aspects, these measures should still be investigated qualitatively to get a more fair view over how influential the stakeholders can be considered to be relative others.

On a general level, one can also look at which stakeholders hold a seat in the different committees of the ecosystem governance structure. However, these do not have to align as a firm can influence both by having representatives in and outside leadership positions, of which the latter is the more common [50]. This phenomenon can be noted when comparing firms with members on the PMC and Committers group in the Apache Hadoop OSS ecosystem (see Fig. 4) with firms that have the overall highest influence score (based on out-degree and betweenness centrality, see Table 4). NTT Data, with a relatively high activity both in regard to the patch and comments networks (see Fig. 5), have a very limited number of places on both the PMC and Committers group. Furthermore, when changing the scope of the analysis (e.g., a certain set of releases or a component—see Sect. 4.2), the governance structure may give an even less representative overview as different stakeholders have different interests, why the network approach proposed by SIA may prove more valuable.

Another approach to measuring influence with other than centrality measures would be to use pure count-based measures of a developer's activity, e.g., number of comments and code commits. As highlighted by Joblin et al. [51], these, however, give a simplified view of a developers position and do not consider the inter-developer relationship. By considering the latter, an analysis can investigate, for example, how active the developers are, how efficiently they can communicate with others, how they are able to mediate and control the flow of information between others, and have relationships "with others that are themselves" central [23]. As further shown [51], network-based measures are equally good, and in certain cases better than count-based measures at describing how influential a developer is. Certain count-based aspects are however included in SIA as it does recommend the use of binary edges as a complement to the weighted edges. As is mentioned in S6 (see Sect. 3), a high out-degree centrality based on weighted edges may indicate that the focal stakeholder has a high influence on its adjacent neighbors and is good at communicating its views relative others in the network [25]. However, this way of measuring out-degree centrality does not provide information about the total number of connections of a stakeholder, which may better show the number of collaborations and opportunities to spread one's opinions [43].

Furthermore, it may be noted that there are other centrality measures available [22] than those proposed in the CSF. We based our choice of out-degree, betweenness, closeness, and eigenvector centrality measures on the suggestion of Faust [23] as explained in Sect. 3. These are generally adopted in explaining the centrality and importance of an actor when analyzing OSS ecosystems (e.g., [25, 52, 53]).

5.2 Limitations and threats to validity

As a proof of concept demonstration that SIA is functional and practical in stakeholder analysis in a large ecosystem, we described a case study on the Apache Hadoop OSS ecosystem. The ecosystem has a community-managed governance model, meaning that the OSS project is owned and managed by the community [54], and a meritocratic authority structure, meaning that influence is gained by proving merit [2] and by establishing a symbiotic relationship with the ecosystem [11]. Another important characteristic of the chosen ecosystem is the high concentration of firms among its stakeholders, as we are interested in identifying and analyzing stakeholder on the organizational level.

This application of SIA in our case study, however, is not without a number of threats to validity. A threat to the internal validity concerns the way how weights are calculated (see S5, Sect. 3). The consideration taken to the relative size in regard to changed lines of code does account for the net amount (i.e., added and removed lines), but commits containing larger amounts of non-meaningful content may give a non-fair view. Thus, it may be valuable to compare interaction networks and influence profiles based on both weighted and binary edges. Also, comparing networks based on different requirement artifact repositories (as exemplified in Sect. 4) can help to give a more nuanced view.

A related threat is that we consider issues in general as "requirements," which may be further extended in our reasoning of requirements artifacts in general. This is based on the nature of RE in OSS as informal and decentralized [6]. Requirements consist of fragmented representations, such as issues, mail thread discussions, and commits [7]. Further mitigation of this threat could include textual and natural language processing of the content in each of the requirements artifacts. This is a vibrant topic in the research field of mining software repositories. However, we consider this topic as out of the scope of SIA as we focus on the identification and stakeholder analysis process in its form and structure. We do acknowledge the topic as complementary quality aspects that should be further researched and integrated with our proposed process in future research.

A further threat concerns the determination of organizational affiliation of individuals in the OSS ecosystem. We adopted a heuristic approach as suggested by earlier research [40, 41], starting with an analysis of e-mail subdomains and complementing with second- and third-level sources such as social network sites as LinkedIn and Facebook, as well as blogs, community communication (e.g., comment-history, mailing lists, IRC logs), web articles and firm websites. We acknowledge this is a delicate and complex process that is best mitigated by "knowing" the ecosystem and actively interacting with its communication channels. In SIA, we recommend using a mix-method triangulation with both qualitative and quantitative approaches.

The case study we described exemplifies how the different steps of SIA can be applied and the insights that can be gained. We acknowledge that a single case study is not sufficient to prove validity in terms of repeatability and utility, and that this is only a first step in the artifact validation phase of our design science research [20]. The characteristics of the Apache Hadoop OSS ecosystem, i.e., community-managed, meritocratic, and multi-vendor, do however indicate what types of OSS ecosystems might benefit from a stakeholder analysis using SIA. Further investigation of SIA's utility and repeatability is out of the scope of this study and instead left for future work. Future research should consider applying SIA from a focal firm's perspective and study different types of OSS ecosystems with a more nuanced authority structure, e.g., as both autocratic, democratic, and meritocratic coordination processes can act in parallel [55]. This falls naturally in the design science research approach as it is an iterative search process for an artifact that will solve the stated problem [21].

6 Conclusions

This study proposes the stakeholder influence analysis (SIA) method which aims to help firms involved in OSS ecosystems to characterize ecosystem's stakeholders according to their level of influence on the ecosystem's RE process. This is an important attribute due to the collaborative and informal nature of the OSS ecosystem's RE processes, and often meritocratic governance structure. SIA, therefore, allows firms to see in which requirements a stakeholder holds a certain interest, and thereby create an overview of a stakeholder's agenda. This also allows firms to understand how stakeholders invest their resources, and with whom they collaborate according to their agenda. Thus, SIA offers input to how firms involved in OSS ecosystems should construct their contribution strategies and act in the politics and negotiations of the ecosystem's RE process in order to align it with their internal RE process and product planning. It can be concluded that SIA shows potential through the case study on the Apache Hadoop OSS ecosystem, while further work is needed in regard to external validity.

In future work, we therefore aim to refine and validate SIA quantitatively and qualitatively through further design cycles involving additional OSS ecosystems, and from existing focal firms' perspectives. Further, we aim to investigate how the stakeholder analysis processes resulting from SIA may be used as an input to the construction and execution of contribution strategies [13].

Acknowledgements We would like to thank Dr. Alma Orucevic-Alagic and the anonymous reviewers for their valuable feedback and inputs.

References

1. Munir H, Wnuk K, Runeson P (2016) Open innovation in software engineering: a systematic mapping study. Empir Softw Eng 21(2):684–723
2. Nakakoji K, Yamamoto Y, Nishinaka Y, Kishida K, Ye Y (2002) Evolution patterns of open-source software systems and communities. In: Proceedings of the international workshop on Principles of software evolution, pp 76–85. ACM
3. Jansen S, Brinkkemper S, Finkelstein A (2009) Business network management as a survival strategy: a tale of two software ecosystems. In: Proccedings of the 1st international workshop on software ecosystems, pp 34–48
4. Glinz M, Wieringa RJ (2007) Guest editors' introduction: stakeholders in requirements engineering. IEEE Softw 24(2):18–20
5. Alspaugh T, Scacchi W, et al. (2013) Ongoing software development without classical requirements. In: 21st IEEE international requirements engineering conference, pp 165–174. IEEE
6. Ernst N, Murphy GC (2012) Case studies in just-in-time requirements analysis. In: IEEE second international workshop on empirical requirements engineering, pp 25–32. IEEE
7. Scacchi W (2002) Understanding the requirements for developing open source software systems. In: Software, IEE proceedings, vol 149, pp 24–39. IET
8. German DM (2003) The gnome project: a case study of open source, global software development. Softw Process Improv Pract 8(4):201–215
9. Laurent P, Cleland-Huang J (2009) Lessons learned from open source projects for facilitating online requirements processes. In: Glinz M, Heymans P (eds) Requirements engineering: foundation for software quality. Springer, Berlin, pp 240–255
10. Baars A, Jansen S (2012) A framework for software ecosystem governance. In: Cusumano MA, Iyer B, Venkatraman N (eds) Software business. Springer, Berlin, pp 168–180
11. Dahlander Linus, Magnusson Mats G (2005) Relationships between open source software companies and communities: observations from nordic firms. Res Policy 34(4):481–493
12. Jensen C, Scacchi W (2007) Role migration and advancement processes in ossd projects: a comparative case study. In: 29th international conference on software engineering, 2007, pp 364–374. IEEE
13. Wnuk K, Pfahl D, Callele D, Karlsson E-A (2012) How can open source software development help requirements management gain the potential of open innovation: an exploratory study. In: Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement, pp 271–280. ACM
14. Frooman J (1999) Stakeholder influence strategies. Acad Manag Rev 24(2):191–205
15. Rowley TJ (1997) Moving beyond dyadic ties: a network theory of stakeholder influences. Acad Manag Rev 22(4):887–910
16. Milne A, Maiden N (2012) Power and politics in requirements engineering: embracing the dark side? Requir Eng 17(2):83–98
17. Aurum A, Wohlin C (2003) The fundamental nature of requirements engineering activities as a decision-making process. Inf Softw Technol 45(14):945–954
18. Pacheco C, Garcia I (2012) A systematic literature review of stakeholder identification methods in requirements elicitation. J Syst Softw 85(9):2171–2181
19. Freeman RE (1984) Strategic management: a stakeholder approach. Cambridge University Press, Cambridge
20. Wieringa RJ (2014) Design science methodology for information systems and software engineering. Springer, Berlin
21. Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. MIS Q 28(1):75–105
22. Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge
23. Faust K (1997) Centrality in affiliation networks. Soc Netw 19(2):157–191
24. Newman M (2010) Networks: an introduction. Oxford University Press, Oxford
25. Orucevic-Alagic A, Höst M (2014) Network analysis of a large scale open source project. In: 40th EUROMICRO conference on software engineering and advanced applications, pp 25–29, Verona, Italy, 2014. IEEE
26. Teixeira J, Robles G, González-Barahona JM (2015) Lessons learned from applying social network analysis on an industrial free/libre/open source software ecosystem. J Internet Serv Appl 6(1):1–27
27. Damian D, Marczak S, Kwan I (2007) Collaboration patterns and the impact of distance on awareness in requirements-centred social networks. In: International requirements engineering conference, pp 59–68. IEEE
28. Marczak S, Damian D, Stege U, Schroter A (2008) Information brokers in requirement-dependency social networks. In: International requirements engineering, 2008, pp 53–62. IEEE
29. Bhowmik T, Niu N, Singhania P, Wang W (2015) On the role of structural holes in requirements identification: an exploratory study on open-source software development. ACM Trans Manag Inf Syst 6(3):10:1–10:30
30. Linåker J, Rempel P, Regnell B, Mäder P, (2016) How firms adapt and interact in open source ecosystems: analyzing stakeholder influence and collaboration patterns. In: Daneva M, Pastor O (eds) Requirements engineering: foundation for software quality, REFSQ, (2016) Lecture Notes in Computer Science, vol 9619. Springer, Cham
31. Johnson G, Scholes K, Whittington R (2008) Exploring corporate strategy: text & cases. Pearson Education, London
32. Newcombe Robert (2003) From client to project stakeholders: a stakeholder mapping approach. Constr Manag Econ 21(8):841–848
33. Mendelow A (1991) Stakeholder mapping. In: Proceedings of the 2nd international conference on information systems. Cambridge, MA
34. Munir H, Linåker J, Wnuk K, Runeson P, Regnell Björn (2018) Open innovation using open source tools: a case study at sony mobile. Empir Softw Eng 23(1):186–223
35. Mitchell RK, Agle BR, Wood DJ (1997) Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts. Acad Manag Rev 22(4):853–886
36. Barnett GA (2011) Encyclopedia of social networks. Sage Publications, Thousand Oaks
37. Damian D, Kwan I, Marczak S (2010) Requirements-driven collaboration: leveraging the invisible relationships between requirements and people. In: Mistrík I, Grundy J, Hoek A, Whitehead J (eds) Collaborative software engineering, Springer, Berlin, Heidelberg
38. Henkel J (2008) Champions of revealing-the role of open source developers in commercial firms. Ind Corp Chang 18(3):435–471
39. Dahlander L, Wallin MW (2006) A man on the inside: unlocking communities as complementary assets. Res Policy 35(8):1243–1259
40. Bird C, Nagappan N (2012) Who? Where? What?: examining distributed development in two large open source projects. In: Proceedings of the 9th IEEE working conference on mining software repositories, pp 237–246. IEEE Press
41. Gonzalez-Barahona JM, Izquierdo-Cortazar D, Maffulli S, Robles G (2013) Understanding how companies interact with free software communities. IEEE Softw 30(5):38–45

42. Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci U S A 101(11):3747–3752
43. Opsahl T, Agneessens F, Skvoretz J (2010) Node centrality in weighted networks: generalizing degree and shortest paths. Soc Netw 32(3):245–251
44. Freeman LC (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239
45. Hanneman RA, Riddle M (2005) Introduction to social network methods. University of California Riverside, Riverside
46. Brandes U (2001) A faster algorithm for betweenness centrality*. J Math Sociol 25(2):163–177
47. Newman MEJ (2001) Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Phys Rev E 64(1):016132
48. Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92(5):1170–1182
49. Runeson P, Höst M, Rainer A, Regnell B (2012) Case study research in software engineering—guidelines and examples. Wiley, Hoboken
50. Schaarschmidt M, Walsh G, von Kortzfleisch HFO (2015) How do firms influence open source software communities? A framework and empirical analysis of different governance modes. Inf Organ 25(2):99–114
51. Joblin M, Apel S, Hunsen C, Mauerer W (2017) Classifying developers into core and peripheral: an empirical study on count and network metrics. In: Proceedings of the 39th international conference on software engineering, pp 164–174. IEEE Press
52. Bird C, Gourley A, Devanbu P, Gertz M, Swaminathan A (2006) Mining email social networks. In: Proceedings of the 2006 international workshop on mining software repositories, pp 137–143. ACM
53. Hossain L, Wu A, Chung KKS (2006) Actor centrality correlates to project based coordination. In: Proceedings of the 2006 20th anniversary conference on computer supported cooperative work, pp 363–372. ACM
54. O'Mahony S (2007) The governance of open source initiatives: what does it mean to be community managed? J Manag Gov 11(2):139–150
55. Shaikh M, Henfridsson O (2017) Governing open source software through coordination processes. Inf Organ 27(2):116–135