Developing a Cloud Computing Platform for Big Data: The OpenStack Nova case

Jose Teixeira
Information Systems Science Unit
University of Turku
Finland
jose.teixeira@utu.fi

Abstract—New developments in virtualization and cloud computing infrastructure keep leveraging the value of Big Data for organizations. In this research, we explore how competing firms collaborate in the development of the OpenStack Nova open-source cloud computing infrastructure. We employ a mix-methods approach that bridges an ethnographic approach with computerized Social Network Analysis (SNA) that data-mines the OpenStack project source-code and its repositories.

We take a longitudinal analysis approach that allows us to observe how key events in the competitive cloud computing industry have affected the development of the OpenStack project over time. Our findings capture how companies such as HP, Rackspace, IBM, Mirantis, Vmware, Citrix and many others simultaneously collaborate and compete in a high-networked open-source software project over time. After integrating our findings with the current body of theoretical knowledge in R&D Management, Software Engineering and Open-Source Software, we argue that the established coopetition management theories provide powerful lenses for understanding the competitive and collaborative issues that are simultaneously present and interconnected in the OpenStack open-source project.

Keywords—Big Data; Cloud Computing; Open-Source; OpenStack; Open-Coopetition;

I. INTRODUCTION

In an era of recognizable software-crisis [1] the move of firms towards geographically-distributed, and often off-shored, software development teams is being challenged by collaboration issues. On this matter, the open-source phenomenon may shed some light, as successful cases on distributed collaboration in the open-source community have been repeatedly reported [2], [3], [4]. While practitioners face difficulties with globally distributed software development teams there remains a lack of research in academia addressing the collaboration dynamics of large-scale and complex distributed software projects [5], [6]. In this paper, we attempt to explore this gap by exploring the collaboration networks in the development of a complex cloud computing infrastructure for Big Data, the OpenStack Nova open-source project.

II. THE OPENSTACK PROJECT

OpenStack is an open-source software cloud computing platform that is primarily deployed as an Infrastructure as a Service (IaaS) solution. It started as a joint project with Rackspace, an established IT web hosting company, and NASA; a well-known USA governmental agency responsible for the civilian space program, aeronautics and aerospace research. Today more than 200 firms joint-develop OpenStack while contributing to different open-source projects governed by the OpenStack Foundation. Both hardware and software developers affiliated with companies such as AT&T, AMD, Canonical, Cisco, Dell, EMC, Ericsson, HP, IBM, Intel, NEC, NASA and many others work together with independent non-affiliated developers in a scenario of pooled R&D in an open-source fashion.

The OpenStack Nova project, our unit of analysis, is a cloud computing fabric controller, the main part of an IaaS system. It is the biggest and most core project governed by the OpenStack foundation. The project originally started within the NASA Ames Research Laboratory, but it further evolved to an inter-firm and high-networked open-source project developed by dozens of firms and thousands of developers. In the very competitive and expanding industry of cloud computing products and services, the OpenStack Nova project brought together volunteers and firm-sponsored software developers who collaborate over the Internet in an open and transparent manner while giving up many of the traditional intellectual property rights.

III. METHOD

We engaged in a multidisciplinary mix-methods approach employing both qualitative methods and Social Network Analysis. Our network analysis emphasis on the visualization of collaborative activities as in prior studies in Biomedicine [7], Innovation Studies [8], Information Systems [9] and Software Engineering [10].

We made use of ethnographic material to inform a Social Network Analysis (i.e. the collected ethnographic data guided the sampling and design of the more positivist network analysis) and vice-versa (i.e. the network analysis raised new questions that required an qualitative ethnographic approach).

Within Table I, we capture a set of multidisciplinary seminal works that guided our research design.
Table I: Multidisciplinary approach

<table>
<thead>
<tr>
<th>Employed approach</th>
<th>Discipline(s)</th>
<th>Seminal works</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netnography</td>
<td>Marketing</td>
<td>[11] [12]</td>
</tr>
<tr>
<td>Mining of software repositories</td>
<td>Software-Engineering</td>
<td>[13] [14]</td>
</tr>
<tr>
<td>Network analysis of digital trace data</td>
<td>Information-Systems</td>
<td>[15] [16]</td>
</tr>
<tr>
<td>Network analysis with emphasis on the visualization of collaborative activities</td>
<td>Biomedicine, Bibliometrics, Innovation-Studies</td>
<td>[7] [8] [16]</td>
</tr>
<tr>
<td>Network analysis of massive networked data, Use of clustering and sub-community detection algorithms.</td>
<td>Physics, Mathematics, Computer-Science, Anthropology, Bioinformatics</td>
<td>[17] [18] [19] [20]</td>
</tr>
</tbody>
</table>

IV. RESULTS

By mining the OpenStack Nova source-code change-log with Social Network Analysis, we extracted information about software developers (network nodes) and their collaborative behaviors (network edges). Taking a longitudinal approach we observed how the collaborative network evolved over time. As in [7] [8] [16], we emphasized on the visualization of collaborative activities we constructed visualization that capture the evolution of the OpenStack collaborative network release after release. The Figure 1 aggregates visualizations with degree centrality (i.e. with more networked developers at the center) that allow us to depict the evolution of code-collaboration in the OpenStack Nova project.

With very dense networks, the direct interpretation of the visualizations was difficult. Therefore we performed an automated sub-community detection using the state-of-art simmelian backbones extraction method [21]. We opted to use data from the last OpenStack releases (Grizzly, Havana and Icehouse) due to higher project maturity and a steady diminution of group cohesion (i.e. tendency for sub-grouping) as plotted in Figure 2. The emergent sub-communities in Figure 3 surprising reveal a low degree homophily in code-collaboration.

V. CONTRIBUTION

Our social network visualizations capture how features of collaboration (among software developers affiliated with different companies) evolved over time, release after release, event after event. Our findings reinforce prior research addressing how rival firms simultaneously collaborate and compete in the open-source arena [22], [23], [9].

A set of our retrieved network visualizations are available, under public domain, at our project website on the Internet http://www.jteixeira.eu/OpenStackSNA.

REFERENCES

Figure 2: Evolution of group-size and group-cohesion over time

Figure 3: Sub-community detection using Simmelian backbones extraction


