coDNA: Visualizing Peer Production Processes

Ofer Arazy
University of Alberta
Alberta, Canada
University of Haifa, Israel
ofer.arazy@ualberta.ca

Henry Brausen
University of Alberta
Alberta, Canada
hbrausen@ualberta.ca

David Turner
University of Alberta
Alberta, Canada
dwt@ualberta.ca

Adam Balila
University of Haifa,
Haifa, Israel
adambalila@hotmail.com

Eleni Stroulia
University of Alberta,
Alberta, Canada
stroulia@ualberta.ca

Joel Lanir
University of Haifa,
Haifa, Israel
ylanir@is.haifa.ac.il

Abstract

Our demo for CSCW2015 is an information visualization tool designed to illustrate the temporal evolution of the peer production process. We combine comprehensive data extraction methods (automated, manual, machine learning) with user-friendly visualization techniques. Our visualization tool – *coDNA* – supports researchers in the development of grounded theory of peer production and allows practitioners to monitor production processes within their online community.

Introduction

Large scale collaborative efforts such as Wikipedia or open source software development projects represent a community-based model for the production of knowledge-based goods. In these peer -production projects, contributions of knowledge are made by volunteers, who self-organize to manage the production process. Such sociotechnical systems are complex assemblages of: (a) human agents and their social organizations; (b) cultural values and organizational norms; and (c) technological artifacts, all of which interact with each other and evolve over time [1]. Recent years have seen a surge in the scale and variety of peer-production projects: from the sharing of volunteers' computer power to produce a distributed super-computer (e.g. SETI@Home[1]) to the community-based design of vehicles (Local Motors[2]).

The overarching objective of this project is to support the investigation of online production communities and enable the generation of grounded theory of peer production. In particular, we focus on the temporal dynamics of the process by which small product elements are contributed by volunteers and then integrated into a unified knowledge-based product and seek to delineate the sequential patterns of collaborative production.

The sheer scale and complexity of peer-production systems present a serious barrier to manual methods for identifying relevant patterns of behavior, thus calling for an automatic method for analyzing knowledge production processes. The availability of temporal data harvested from logs of IT systems supporting peer-production could be employed to track the interactions in socio-technical systems and capture the sequential contributions to a common artifact. Just as the Human Genome Project maps the sequences of genes in the human DNA, we will chart the 'DNA' sequences of computer-mediated collaboration. We, thus, refer to our project as *Collaboration DNA*, or simply *coDNA*.

Visualization makes relevant processes visible that would otherwise be difficult to interpret. This is of particular importance for scholars investigating online production communities and for the administrators and owners of these communities. The software tool we will demo at CSCW2015 was developed to facilitate the exploration and analysis of 'collaboration DNA' sequences. The tool visualizes data collected from a series of peer production process. Our aim is to employ the insights gained through the visualization to the development of a grounded theory [2] of peer-production.

The coDNA visualization tool

The complexity and multi-dimensionality of peer production projects present a major challenge for visualization tools. To date, most of the visualizations have tended to focus on a single dimension presented a static view, capturing metrics such as total number of activities per contributor or the structure of relationships between contributors. We wish to go beyond current practices and develop a visualization tool that would capture the *temporal sequences of activities* in the evolution of peer production across *multiple dimensions*. In what follows, we present the *coDNA* tool (http://codna.org).

Several design principles guided the development of *coDNA*, as described below:

- **Temporal dynamics**. Given our focus on the temporal evolution of peer production, it is essential that we capture the timing of events and that the visualization emphasizes temporal aspects (for example moving the view forward and backward and zooming in/out in time).
- **Nested organizations**. We view online production communities as nested organizations, where each organization operates multiple projects. For example, Wikipedia is an organization that operates multiple projects – each project being the collaborative

[1] http://setiathome.berkeley.edu/

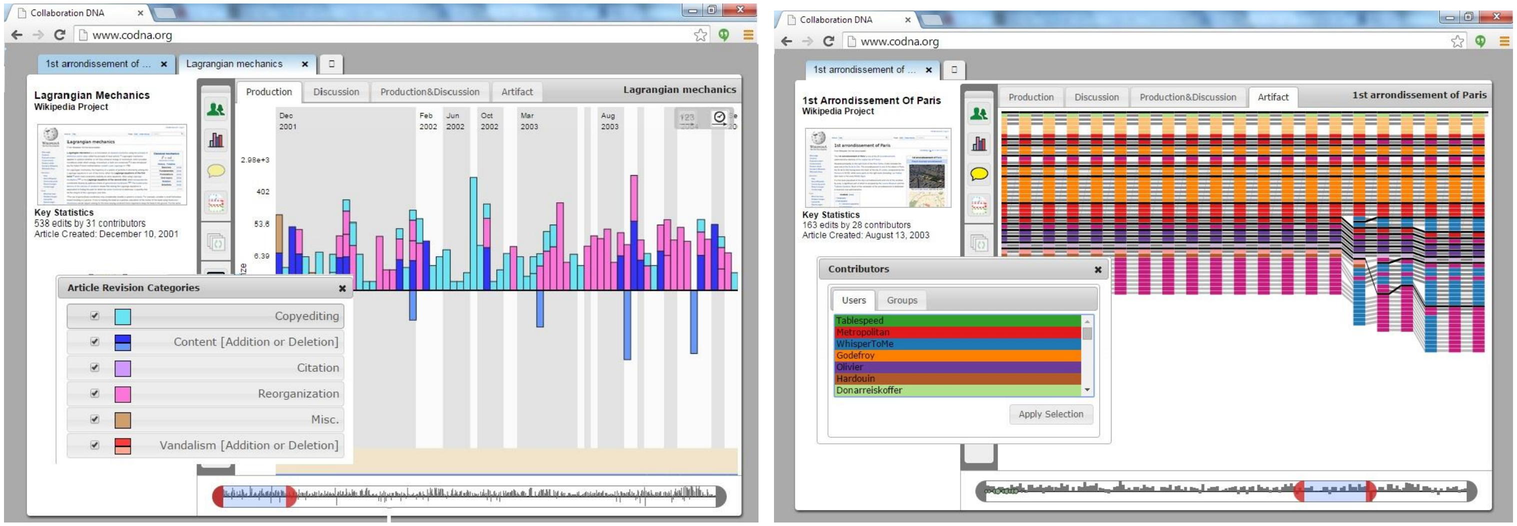
[2] https://localmotors.com/

authoring on an encyclopedic entry. Our tool captures and visualizes these relationships.

- **Project vs. Contributor focus**. When studying collaboration patterns we may focus on a particular project, mapping all participants' activities in this project. Alternatively, we may want to focus on the activities of a single contributor across multiple projects. The visualization tool should allow switching between these two modes.
- **Production & Coordination**. Activities in online production communities include the contribution of product elements (e.g. editing a Wikipedia page, contributing code to open source projects), as well as coordination activities (i.e. discussions around task allocation, negotiating conflicting views). It is important that we capture and represent both dimensions.
- **Activities**. The basic building blocks of 'collaboration DNA' are activities, specifically production and coordination activities. It is important that we capture and represent key attributes of those activities such as: date & time, activity type, scope of activity, and the contributor
- **Contributors and roles**. Contributors play different roles in the organization, often moving between roles. The visualization tool should record and present the organization role of a contributor (at the time when making the making each contribution).
- **Process vs. Product focus**. The outcome of a peer production process in a knowledge-based product, an artefact. We want to be able to capture and represent the architecture of this product, i.e. the ways in which various modules are organized , as well as the relationships between contributions of small elements and the product's architecture.

Data was collected from several peer production organizations, and from multiple projects within each organization. We tried to capture the key attributes specified by the design principles above. Some attributes of the data were harvested from the logs of systems supporting the collaborative production process (e.g. timing of each activity, contributor's ID), while other attributes required additional processing. In some cases we developed automated tools to calculate certain values (e.g. the scope of a Wikipedia 'edit' was measured through the Levenshtein distance and we developed an algorithm to perform this calculation; tracking the relationships between contributions and modules of the product required the development of a tool). In other cases, we relied on manual analysis, for example in determining the type of activities (for both production and coordination activities). Where possible, we used the manually-annotated data set for training a machine learning algorithm, and then automated the task for subsequent data sets.

The tool is a browser based SVG visualization that was built using several programming languages (javascript, HTML and CSS, and for the server side API we used PHP. The specific javascript libraries used are D3, jQuery, Backbonejs and Underscorejs.



Lagrangian Mechanics
Wikipedia Project

Key Statistics
538 edits by 31 contributors
Article Created: December 10, 2001

Article Revision Categories

- ☑ 🟦 Copyediting
- ☑ 🟦 Content [Addition or Deletion]
- ☑ 🟪 Citation
- ☑ 🟪 Reorganization
- ☑ 🟫 Misc.
- ☑ 🟥 Vandalism [Addition or Deletion]

1st Arrondissement Of Paris
Wikipedia Project

Key Statistics
163 edits by 28 contributors
Article Created: August 13, 2003

Contributors

Users | Groups

- Tablespeed
- Metropolitan
- WhisperToMe
- Godefroy
- Olivier
- Hardouin
- Donarreiskoffer

Screenshots of coDNA showing (a) 'Production' view with a filter on the type of activities and (b) 'Artefact' view with a filter on the contributors

Relevance to the CSCW audience

Researchers studying peer production are overwhelmed with the complexities of these projects. While the availability of data harvested from systems' logs makes it easy to track the activities that transpired. Nonetheless, the sheer amounts of the data make it difficult to identify patterns and draw insights. As a result, existing approaches have restricted the analysis to data that could be extracted automatically, tended to focus on a single dimension, and presented a static view capturing the production at a single point in time.

Our approach focuses on the temporal evolution of the peer production process, capturing its multiple dimensions (e.g. production, coordination, artefact). We combine comprehensive data extraction methods (automated, manual, machine learning) with user-friendly visualization techniques. The resulting information visualization – *coDNA* - provides an excellent tool for researchers interested in understanding online collaborative production processes, insight generation, and the development of grounded theory of peer production. Our intention is to make this tool available to CSCW research community, hoping that others would contribute additional data sets, thus allowing the comparison of collaboration patterns across settings. Our vision is to build a community of researchers sharing data, using the tool, and contributing to the enhancement of *coDNA*.

We strongly believe that our visualization system is also valuable to practitioners administering peer production projects. Owners and administrators of such projects often lack tools to monitor processes within the community. Using *coDNA* they could follow the activity profiles of contributors as they evolve, better allocate tasks between people, and open career paths within the community to relevant contributors. *coDNA* could also facilitate the tacking of projects' evolution, allowing early detection when projects go off tracks, and thus enabling project administrators to take preventive measures

<u>References:</u>

1. Geels, F.W. From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research policy*, 33, 6 (2004), 897-920.
2. Glaser, B.G. *Emergence vs forcing: Basics of grounded theory analysis*. Sociology Press, 1992.