Project management in the Wikipedia community

Hang Ung
Hewlett-Packard Laboratories, Palo Alto, CA
and Centre de Recherche en Gestion,
Ecole Polytechnique, Palaiseau, France
hang.ung@hp.com

Jean-Michel Dalle
Université Pierre et Marie Curie, Paris, France
jean-michel.dalle@upmc.fr

ABSTRACT

A feature of online communities and notably Wikipedia is the increasing use of managerial techniques to coordinate the efforts of volunteers. In this short paper, we explore the influence of the organization of Wikipedia in so-called projects. We examine the project-based coordination activity and find bursts of activity, which appear to be related to individual leadership. Using time series, we show that coordination activity is positively correlated with contributions on articles. Finally, we bring evidence that this positive correlation is relying on two types of coordination: group coordination, with project leadership and articles editors strongly coinciding, and directed coordination, with differentiated online roles.

Keywords

Wikipedia, online communities, project-based organizations, leadership.

Categories and Subject Descriptors

K.4.3 [**Organizational Impacts**]: Computer-supported collaborative work

1. INTRODUCTION

Projects are essential to coordination within companies [3]. Company-wide, financial resources, office spaces, workers and managers are often allocated to specific projects, whose outcomes or deliverables are well-defined. At the project level, the manager defines and assigns tasks to his team members, leveraging his leadership to achieve the project's objectives. Thus, projects are organizational subentities within large corporations and this design is thought to provide increased accountability, management performance, and (perhaps consequently) productivity [6], in particular for knowledge-intensive firms [7]. While projects may often involve several functional entities (marketing, engineering, etc.) and sometime span multiple business units, they essentially rely on authority, conveyed by leadership and hierarchy.

In an online community there is, by contrast, no formal hierarchy or at least not one comparable to those found in companies, which are built atop employer-employee contractual relationships [1]. Notwithstanding the absence of such contracts, project-like forms of organization do exist in online peer production systems. For instance, the Apache community, initially focused on a single piece of software, the Apache HTTP server, now develops over 70 projects, some being completely independent, others being interdependent modules of a larger software solution. Yet, all these projects share resources: tools (*e.g.*, code repository, email lists), norms and perhaps most importantly, developers' time.

Another striking example is provided by the online encyclopædia Wikipedia where, among other coordination pages, "WikiProjects" (here simply called *projects*) have become important [5]. Wikipedia defines a project as:

> A collection of pages devoted to the management of a specific topic or family of topics within Wikipedia; and, simultaneously, a group of editors who use those pages to collaborate on encyclopedic work.[1]

In both cases, the project structure is primarily conceived as a tool supporting group self-management and is designed to help group members coordinate their own work at the project level. Such coordination activity typically consists of stating the project's scope and objectives, assigning task priorities and communicating between group members. Kittur et al. point out that the group influences members' behaviors, for instance having them perform certain tasks they would not otherwise be inclined to do [5].

In this context, and following up on the recent literature emphasizing on organizational aspects of online communities [2, 4], it seems that studying project-based organization in online communities could provide a better understanding of how peer production systems successfully achieve the rather complex coordination of numerous volunteers, which in other production systems would rely on either markets or hierarchies [1].

In this paper, we investigate the project-based coordination activity within Wikipedia and its relation to individual and collective production behaviors. More specifically, after presenting our data processing and sample, we characterize the bursty nature of coordination activity. Using time series, we then assess the relation between coordination and

[1] `http://en.wikipedia.org/wiki/WikiProject`

WikiSym '10, July 7-9, 2010, Gdańsk, Poland

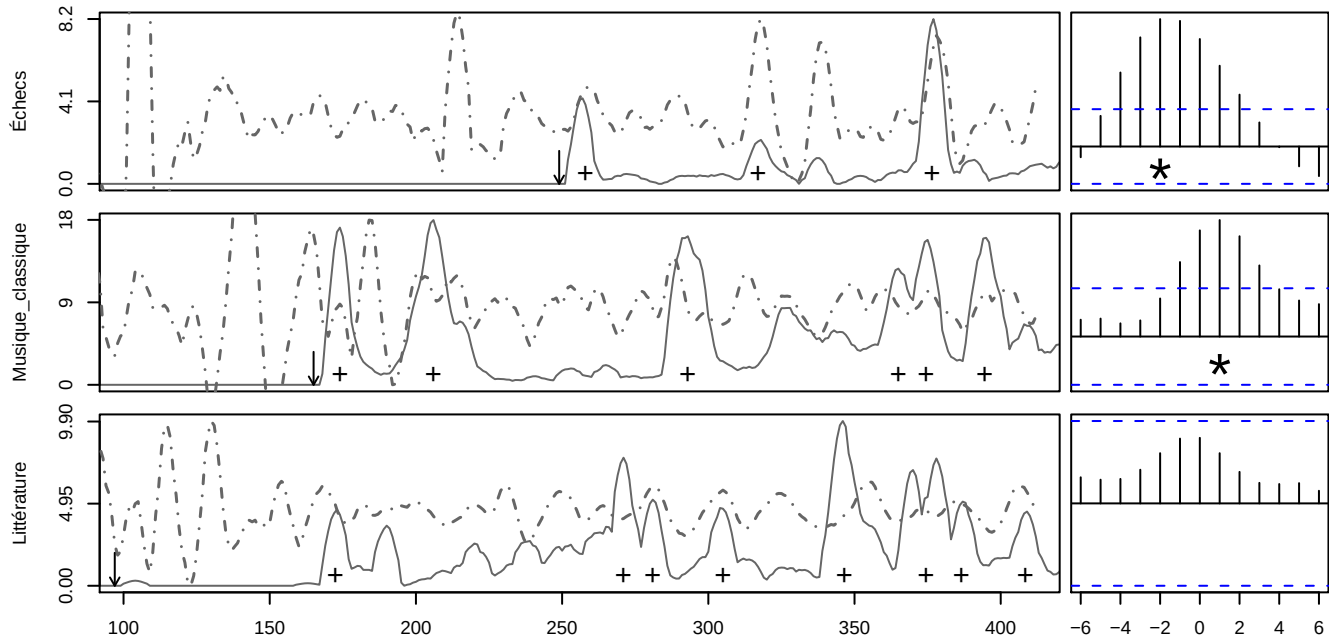

Figure 1: The 3 projects "chess", "classical music" and "literature". X-axis shows time or lag in weeks. Each row shows (left) a project's coordination activity (solid line, in number of contributions per week) where project start is marked by ↓ and peaks by +; production activity (dashed line, arbitrary units), and (right) the cross-correlogram between the two activities where the maximum correlation value is marked by ⋆ when it is above the confidence limit (dashed line).

production activities. We conclude by discussing briefly the results.

2. DATA PROCESSING AND SAMPLE

We retrieved the archive of the French Wikipedia as of March 2009[2], consisting of 3 million pages and 37 million versions, or *revisions* of these pages. The namespace `Projet`[3] contained 18835 pages (or *project-pages*) which are pages used only to manage or coordinate work. For instance, both project-pages entitled `Sport/Participants` and `Sport/Articles_récents` are part of the same `Sport` project, and as their name indicate, the first lists users participating in the project, while the second shows recently created articles belonging to the project. Project-pages were thus aggregated by their title to obtain distinct projects.

This resulted in a list of 833 projects, of which 189 redirect either to other projects or to portals, and 644 can be considered as actual projects. Then, by parsing templates in the talk pages of all articles, we reconstructed for each project a list of *articles* marked by users as belonging to the project. Note that the projects cover a significant set of Wikipedia articles, and in particular the more active ones. Indeed, 28% of all articles belong at least to one project, and these 28% account for 72% of all edits made on articles.

Of the 644 projects, 166 were discarded because we could not identify articles belonging to them (articles were not marked or marked with non-standard templates), and an additional 168 were excluded from our sample because of their very limited project activity (less than 200 revisions made to the project-pages). In summary, our sample consists of 310 projects, each of them primarily characterized by:

- A set of constituting project-pages. Their number varies greatly across projects, smaller projects having only a single project-page, and larger ones over 50.
- A set of articles belonging to it, ranging from a few to over a thousand.

3. RESULTS

3.1 Bursty coordination and leadership

Let us first consider a single project with its set of constituting project-pages and its set of articles. Project-based coordination activity occurs when a user modifies a project-page: for example, an item is added to the list of articles to be improved, priorities are updated, etc. Thus, the editing activity occuring on project-pages is a simple proxy for *(project) coordination activity*.

By counting the number of edits on the project-pages per week, we thus obtain a time series reflecting the weekly dynamics of coordination activity in the project. Bots' edits were excluded from this count and all time series were filtered using a moving average with a window of 7 weeks. Figure 1 shows 3 projects with their coordination activity. As one can observe, coordination activity undergoes significant variations in time, with a few pronounced peaks which correspond to "bursts" of coordination activity in the project.

[2]`http://download.wikimedia.org/frwiki/`
[3]Other versions of Wikipedia, like the English Wikipedia, use the `Wikipedia` namespace for projects instead.

Interestingly, these bursts occur not only in the early life of the projects – at the "kick-off" – but also later on.

Thus, to characterize more precisely the bursts, we define them as time intervals during which coordination activity is above a threshold set to twice the average coordination activity (see Fig. 1). We find that 98% of the sample projects exhibit at least one burst in their coordination activity, with 66% showing 2 or more. On average, coordination activity within the bursts represents 69% of a project's total coordination activity.

Then, for each burst i of project p, we identified the most active user u with respect to coordination activity and calculated the share $\alpha_{i,p}$ of the activity of the burst that originate from u. Hence, bursts of activity from a single user are characterized by $\alpha_{i,p} \approx 1$ as opposed to lower values which denote a more collegial activity. We compare $\alpha_{i,p}$ to the share $\bar{\alpha_p}$ of the most active user in the entire project lifetime and find that for 87% of the projects, the average value of $\alpha_{i,p}$ is greater than $\bar{\alpha_p}$. This means that bursts in coordination activity most of the time reflect an initiative from a single user or else that a single user "takes the lead" the project during a limited period of time. Interestingly, we also find that 68% of successive peaks have different "leaders".

3.2 Correlation between coordination and "production"

A relatively straightforward but non-trivial hypothesis is then that the managerial type of activity occuring in project-pages (coordination activity) should be reflected in the articles's editing activity (production activity). Conversely, rejection of this hypothesis would mean than the project structure imposed upon article pages would for instance be devoted to longer term planning rather than to shorter term coordination, or else that both activities would coexist without being really related, which would be the case if editors of articles would not or only weakly take management activity seriously.

To explore this hypothesized correlation between coordination activity and production activity, we constructed for each project a time series measuring the production activity, by counting the number of (non-bot) edits made to articles belonging to the project. Because we are only interested in project-specific variations, this count was normalized by the total Wikipedia production activity. The time series was first filtered with a 7-week moving average window (as for coordination activity) and then low-frequency variations were filtered out using again a moving average, but with a larger window of 21 weeks ("high-pass"). Therefore, the production activity signal shown on Fig. 1 contains essentially variations of the production activity at the time scale of a few weeks, which is similar to the time scale of variations of coordination activity.

To test our hypothesis, we calculate the cross-correlogram of coordination activity and production activity[9], allowing lag times in weeks in the $[-6, +6]$ range (see Fig. 1). For 74% of the projects, we find at least one lag for which there is a positive correlation significant at the 5% level, suggesting that a large part of the variations observed in the production activity on a project is indeed correlated to coordination activity occuring at the project level.

3.3 Group coordination vs. directed coordination

A two-fold hypothesis can then be formulated to further assess the nature of this relationship. The first is to view a project as a group coordination tool, with group members using project-pages as a place to coordinate *their own* work. Both coordination and production activities would essentially originate from a single group of users and are hence correlated in time. Alternatively, there could exist a pool of Wikipedia users who are not part of such a core group but are more generally contributing. Think typically of users specialized in certain tasks (translating or correcting articles, adding pictures, etc.), or of less frequent and more "peripheral" contributors. The behavior of such contributors could still be affected by coordination activities occuring on project-pages and their attention thus directed towards the more active ones. These two (non-exclusive) coordination mechanisms could obviously find easy analogues in the management of projects in companies.

To get more insight on this issue, we investigate to what extent users that are responsible for the coordination activity on a given project (*project leaders*) are also users committed to articles belonging to the project (*focused users*), and ask whether this is linked to the presence of the correlation previously highlighted. Thus, for each project p we perform the following analysis:

- We define as a leader a user who contributed either more than 5% of the coordination activity (*i.e.*, edits on project-pages), or more than a hundred times.
- For any given week w, we say that a user u is focused if his contributions on articles belonging to the project represent more than half of his contributions in the same week, in which case we denote $f_{u,w} = 1$.
- We aggregate the weekly focus to obtain a (per-user) project focus: $F_u = \sum_w f_{u,w}$.
- Let n be the number of leaders and ϕ the set of n users with the highest F_u values. If the leaders are also the most focused users, then they should belong to ϕ. To measure how much this is the case, we define the ratio $\rho = (\sum_{u \text{ leader}} F_u)/(\sum_{u \in \phi} F_u)$.

Note that we purposedly make use of the weekly focus $f_{u,w}$ and the project focus F_u, instead of simply measuring each user's share of all contributions made to the project's articles, for the latter is actually less representative of a commitment specific to the project. Indeed, very active users like admins will quite often be the largest contributors to any project's articles, even though their behavior is mainly unrelated to the project, for instance performing large numbers of maintenance edits across the whole encyclopædia.

Figure 2 shows the distribution of ρ. There is clearly a vast majority of projects for which ρ is significantly smaller than 1, meaning that most project leaders do not count amongst the most focused users. Typically, for only 3% of the projects is $\rho > 0.8$ while for 37%, $\rho < 0.2$. This result highlights an interesting heterogeneity amongst the sample projects with respect to the involvement of project leaders in production tasks, which might indeed be coherent with what we have called group vs. directed coordination.

Since only 74% of all projects exhibit a correlation between coordination and production activities, it is then natural to ask whether this correlation could be a characteristic

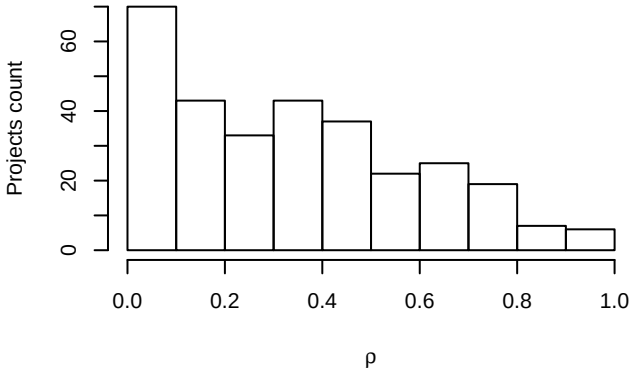

Figure 2: Distribution of ρ among projects. ρ quantifies how much project leaders are also the most focused contributors to articles of the project.

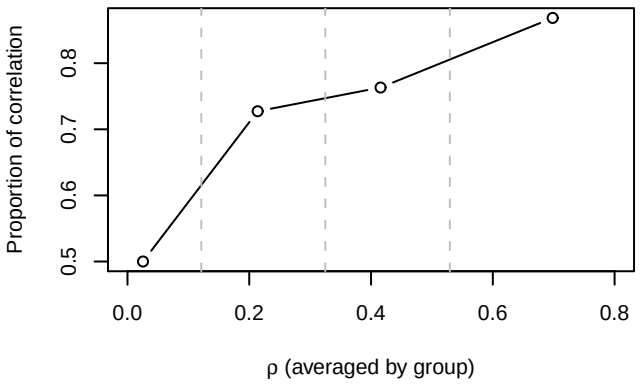

Figure 3: Proportion of projects that exhibit a significant positive correlation between coordination and production activities, as a function of ρ. Vertical lines shows the quartiles used to group projects.

of projects with a higher involvement of leaders in production, *i.e.* with a higher ρ. To explore this hypothesis, we subdivided our sample into four equally-sized groups according to the value of ρ. Then, for each group, we simply calculate the share of projects that exhibit a significant positive correlation.

As shown in Fig. 3, projects with higher ρ are more likely to show a significant positive correlation. In other words, when project leaders and its core group of focused contributors coincide, a positive correlation between coordination and production is more likely to occur: 87% of projects in the upper quartile (with respect to ρ) exhibit this positive correlation. Clearly, this is consistent with the former view of projects as group coordination processes.

However, 50% projects in the lowest quartile exhibit a similar positive correlation, suggesting the existence also of a more "directed" form of coordination, in which leaders and contributors do not coincide, but where the managerial and coordination activity of leaders in project-pages influences the activity of contributors to articles. To further verify this point, we recalculated production activity without the contributions of leaders and computed cross-correlograms in a similar way as described above. We still find a significant correlation for 58% of the projects, supporting the existence of a more directed type coordination, perhaps involving contributors whose focus would shift from a project to another or whose involvement in a given project could be increased under the leadership of other users.

4. CONCLUSION

We showed that in Wikipedia, project-based coordination exhibits a bursty pattern of activity, possibly as a result of a form of leadership. We demonstrated that for most projects, the coordination activity (on project-pages) is positively correlated with the production activity (on articles), supporting the general hypothesis of a role of projects in coordinating individual users' contributions to the encyclopædia. Finally, we found that two types of coordination are likely to coexist: group coordination where the project leaders coincide with the users who are the most focused on the projects' articles, and directed coordination, with more distinct roles.

Together, our results emphasize the heterogeneity of projects and of users' behaviors. On the one hand, a project can be a group of users who know each other and work solely on the articles in the scope of their project. These "island" projects would strongly involve social identification [5]. On the other hand, a project might also function by eliciting the attention of the community as a whole and attracting temporarily the efforts of less topic-focused users. Thus, this distinction also points to the heteregeneity of users with respect to their level of focus, which could be a important characteristic of their behavior [8].

5. REFERENCES

[1] Y. Benkler. Coase's penguin, or, linux and the nature of the firm. *Yale Law Journal*, 112(3):367–445, 2002.

[2] M. den Besten, J.-M. Dalle, and F. Galia. The allocation of collaborative efforts in open-source software. *Information Economics and Policy*, 20(4):316–322, Dec. 2008.

[3] P. O. Gaddis. The project manager. *Harvard Business Review*, 37(3):89–97, 1959.

[4] A. Kittur, B. Lee, and R. E. Kraut. Coordination in collective intelligence: the role of team structure and task interdependence. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1495–1504, Boston, MA, USA, 2009. ACM.

[5] A. Kittur, B. Pendleton, and R. E. Kraut. Herding the cats: the influence of groups in coordinating peer production. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, pages 1–9, Orlando, Florida, 2009. ACM.

[6] C. J. Middleton. How to set up a project organization. *Harvard Business Review*, 45(2):73–82, 1967.

[7] W. H. Starbuck. Learning by Knowledge-Intensive firms. *Journal of Management Studies*, 29(6):713–740, 1992.

[8] H. Ung and J.-M. Dalle. Characterizing online communities with their "signals" (accepted). In *European Academy of Management*, Rome, Italy, 2010.

[9] W. W. Wei. *Time series analysis: univariate and multivariate methods*. Addison-Wesley, 2006.