Semantic Analysis of Formal Policies and Community Practices in OSS

Keywords: Governance, OSS, Online Communities, Natural Language Processing, Semantic Embeddings

Extended Abstract

Open Source Software (OSS) is a multi-billion dollar informal industry supporting major contemporary tech enterprises, academia, and scientific R&D. Started in the 1960s as a countermovement to proprietary software [6], it aims for efficient development through scrutiny and quality contributions pooled across communities of users and mostly volunteering programmers. Individuals or groups launch OSS projects out of specific requirements or personal visions and eventually draw other interested volunteers to support further development for general use. To facilitate cooperation and coordination, these communities have often instituted their own self-governance by means of rules and practical norms defining the roles, rights, and responsibilities of those involved. The emergence of OSS and their sustenance without extrinsic financial motivations or conventional firm control has been of increasing interest in collective action scholarship [1, 4].

In the past few decades, several nonprofits have stood up to mentor OSS communities and support them through infrastructure, IP protection, legal aid, and outreach. To streamline management and operations across all the diverse projects they mentor, these organizations institute their own rules and policies. Meanwhile, community-level organizing and self-governance manifest through routines or repeated patterns of actions borne out of established rules as well as strategies learned from experience [3]. We are interested in how these mentor organizations structure their policies around the various internal practices communities adapt to coordinate and realize their development goals. Our work pursues extensive conversation-level semantic analysis to discover and contrast normative routines against practices mandated by the Apache Software Foundation Incubator (ASFI) program, across 214 projects it has overseen to date.

Mailing lists are a critical lifeline and comprehensive log of all activity in OSS communities. Software development teams perform multiple time and context-specific activities that unfold part by part over the course of email dialogue. We thus base our study around performance programs [5] or scripted subroutines (Table 2). Firstly, semantic role labeling (SRL) is used to identify all situated activities in email and ASF policy texts through individual verbs, the corresponding actors, objects, and other constituents indicating the time of the act, manner, direction, goal, purpose, cause, etc. These extracted expressions of subroutines from both policies and emails were next passed through pre-trained encoders to obtain semantic embeddings, followed by fine-tuned density-based clustering. Such aggregation allows us to retain the most relevant, normative behavior. These resulting dense clusters were further assigned topics through a class-based adaptation of TF-IDF, to identify those containing both formal policies and all other related activity around offices, processes, and resources under the ASF governance. Finally, the routines within each governance cluster are directly compared with the corresponding policy rules by means of pairwise semantic scoring. We identify the highest pairwise similarity score (with the closest policy match) for each routine as a measure of its extent of alignment with the governing institution.

Topic-wise distribution of routines across governance themes showed a marked departure between ASF's own attention in terms of policy focus and the actual prevalence of the themes among community norms (Figure 1). While OSS teams are primarily technical communities of practice, ASF policies placed a much higher emphasis on administration, over development processes and standard operating procedures.

We further performed a multivariate regression analysis of the similarity scores of all the extracted routines, with respect to their particular topic, concurrent community network measures, and technical activity, while controlling for the project the routine belongs to. As expected, positive effects from the more regulated managerial topics indicate behavior around administrative functions as being more closely compliant with policy (Table 1). Meanwhile, the generally negative effects from topics concerning core technical tools and software artifacts are a likely outcome of continuous discovery and evolution of routines to accommodate functionalities along new technology and user needs. Overall, topics together accounted for 63.5% of the explained variance and individually exhibited strong, significant effects ($p < 0.001$) against similarity scores, confirming considerable variation in project behavior even along policied themes. While certain individual projects exhibited noticeable variance in behavior, characteristics like network attributes or technical activity did not exert major influence over the extent of a project's divergence or compliance with the policy.

Codified, established rules serve as an invaluable resource for structuring and preserving functional institutions. Yet, designing optimal governance is challenging in fluid and technologically fast-paced organizations. Our results indicate that compliance with governance is strongly associated with the extent of policy regulation. Identification, assessment, and eventual legitimization of effective practices as recorded recommended norms will not only bolster project performance and sustenance but also recognize and encourage continual learning and innovation [2]. While supporting policy analysis, our work also opens up future directions for deducing implicit norms and routines through natural language dialogue, over the course of organizational learning and adaptation. Such compilation of effective processes evolving across communities is expected to act as a knowledge repository for growing projects, as they learn to coordinate and mobilize teams, recruit volunteers, and launch new products.

References

[1] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. 2006.

[2] John Seely Brown and Paul Duguid. "Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation". In: *Organization science* 2.1 (1991), pp. 40–57.

[3] Richard M Cyert, James G March, et al. *A behavioral theory of the firm*. Vol. 2. 4. Englewood Cliffs, NJ, 1963.

[4] Charlotte Hess and Elinor Ostrom. "A Framework for Analyzing the Knowledge Commons: a chapter from Understanding Knowledge as a Commons: from Theory to Practice." In: (2005).

[5] JG March and HA Simon. "Organizations." In: (1958).

[6] Eric Raymond. "The cathedral and the bazaar". In: *Knowledge, Technology & Policy* 12.3 (1999), pp. 23–49.

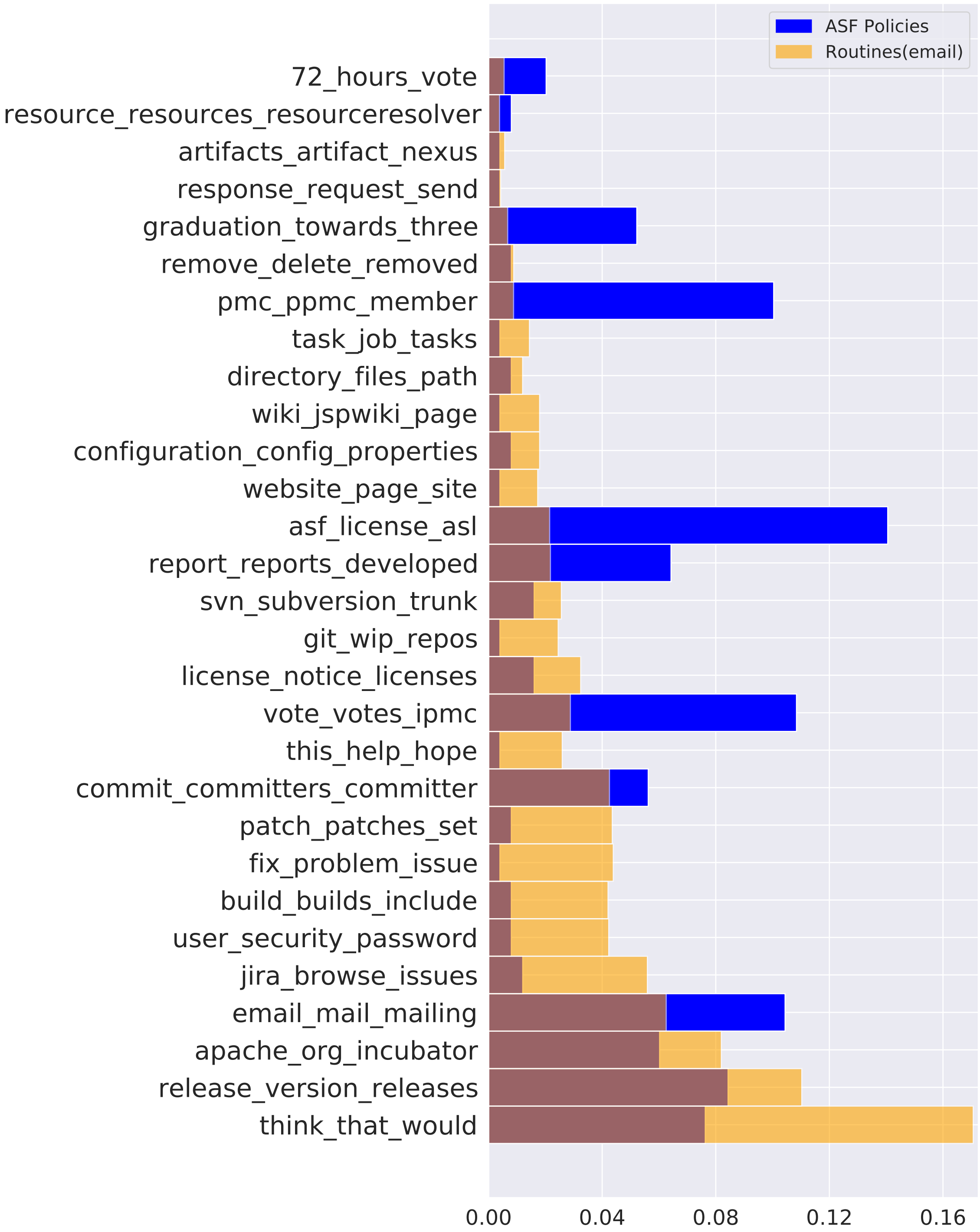

Figure 1: Distribution of ASF rules and recurring email expressions across topics. Heavily regulated themes were found to have significant positive effects and higher resulting similarity between policy and observed routines.

Topic: top 3 words	**Coefficient**
72_hours_vote	0.8654***
pmc_ppmc_member	0.8345***
vote_votes_ipmc	0.7489***
asf_license_asl	0.6745***
commit_committers_committer	0.6640***
release_version_releases	0.5706***
graduation_towards_three	0.5662***
report_reports_developed	0.4651***
apache_org_incubator	0.4614***
svn_subversion_trunk	0.3590***
license_notice_licenses	0.2593***
email_mail_mailing	0.2048***
jira_browse_issues	0.1133***
remove_delete_removed	0.0723***
patch_patches_set	-0.0568***
resource_resources_resourceresolver	-0.2266***
artifacts_artifact_nexus	-0.3268***
configuration_config_properties	-0.4582***
build_builds_include	-0.4772***
wiki_jspwiki_page	-0.5100***
fix_problem_issue	-0.5543***
directory_files_path	-0.5589***
user_security_password	-1.2435***
task_job_tasks	-1.2716***
response_request_send	-1.2952***
git_wip_repos	-1.3603***
website_page_site	-1.3871***
this_help_hope	-2.1057***

Table 1: ASF Governance themes and corresponding effects from regression analysis. *,** and *** indicate significance levels $p < 0.05$, $p < 0.01$ and $p < 0.001$ respectively. Table does not include effects from project controls, STS and technical activity

Original Policy:
’After a vote has finished, the ipmc must send a notice email to the board and then wait for 72 hours before inviting the proposed member’

Semantic Role Parsing:
’ARG0’: [’the ipmc’], ’ARGM-MOD’: [’must’], ’V’: [’send’], ’ARG1’: [’a notice’], ’ARGM-DIR’: [’email’], ’ARG2’: [’to the board’], ’ARGM-TMP’: [’after a vote has finished’]
’ARG1’: [’the ipmc’], ’ARGM-MOD’: [’must’], ’V’: [’wait’], ’ARGM-TMP’: [’after a vote has finished’, ’then’, ’for 72 hours’, ’before inviting the proposed member’]

Performance Programs (After reconstitution):
’After a vote has finished the ipmc must send a notice email to the board’
’After a vote has finished the ipmc must then wait for 72 hours before inviting the proposed member’

Table 2: Policy describing steps in admission to the Incubator Project Management Committee (IPMC), which was parsed into semantic roles and reconstructed into ’performance programs’. These subroutines may individually unfold over emails across the duration of the full process. ARG0 denotes agent, ARG1-ARG5 are patients, ARG-MOD are modals while ARG-TMP and ARG-DIR are the temporal and directional arguments respectively.