Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem

Asher Trockman,† Shurui Zhou,‡ Christian Kästner,‡ Bogdan Vasilescu‡
†University of Evansville, USA  ‡Carnegie Mellon University, USA

ABSTRACT
In fast-paced, reuse-heavy, and distributed software development, the transparency provided by social coding platforms like GitHub is essential to decision making. Developers infer the quality of projects using visible cues, known as signals, collected from personal profile and repository pages. We report on a large-scale, mixed-methods empirical study of npm packages that explores the emerging phenomenon of repository badges, with which maintainers signal underlying qualities about their projects to contributors and users. We investigate which qualities maintainers intend to signal and how well badges correlate with those qualities. After surveying developers, mining 294,941 repositories, and applying statistical modeling and time-series analyses, we find that non-trivial badges, which display the build status, test coverage, and up-to-dateness of dependencies, are mostly reliable signals, correlating with more tests, better pull requests, and fresher dependencies. Displaying such badges correlates with best practices, but the effects do not always persist. In short, signals mostly reliable.

1 INTRODUCTION
Contemporary software development is characterized by increased reuse and speed. Open-source software forges like GitHub host millions of repositories of libraries and tools, which developers reuse liberally [25], creating complex, often fragile networks of interdependencies [11]. This has earned GitHub a reputation as a one-stop shop for software development [39] and as an influence of practices in both open-source and industry [33]. The DevOps culture [31, 46] also contributes to this acceleration, with its emphasis on automation and rapid deployment. As a result, developers are expected to make more decisions at higher speed, e.g., finding which libraries to depend on and which projects to contribute to.

A key enabler of this decision making process is the transparency provided by social coding platforms like GitHub [20, 21]. The development history of open-source GitHub projects is archived and publicly accessible in a standardized format, and user pages display aggregate information about one’s contributions and social standing in the community (e.g., through stars and watchers). This transparency can enhance collaboration and coordination [21]. Using visible cues—known in the literature as signals—collected from personal profile and repository pages, developers can better manage their projects and dependencies, communicate more efficiently, become informed about action items requiring their attention, learn, socialize, and form impressions about each other’s coding ability, personal characteristics, and interpersonal skills [21, 38, 40, 57].

However, open-source ecosystems are also competitive. In order to survive and thrive, projects must successfully attract and retain contributors, and fend off competitors [16, 36, 42, 45]. In a social coding environment, the visible signals enabled by transparency can, therefore, be seen as a survival mechanism, with high profile signalers benefiting the most. For example, more popular and famous projects attract more contributors [62], coding “rock stars” collect thousands of followers [20], and visible traces of developer actions and interactions are used in recruitment and hiring [13, 37].

Here we focus on repository badges, images such as build status, code coverage information may act as an incentive for contributors to improve the project’s test suite. In summary, badges are a potentially impactful feature in transparent, social coding environments. However, the value and effects of badges are not well understood.

In this paper, we explore two main research questions regarding badges. First, we explore the phenomenon quantitatively and qualitatively, and ask (RQ1): What are the most common badges and what does displaying them intend to signal? Second, we analyze whether badges indeed signal what developers expect, and ask (RQ2): To what degree do badges correlate with qualities that developers expect? To this end, we perform a large-scale mixed-methods empirical study of the badges in npm, a large and vibrant open-source ecosystem for JavaScript with documented interdependency-related coordination challenges [11], wherein many badges originated. We observe the frequency and historical adoption of badges among 294,941 npm packages, we survey maintainers and contributors about their intentions and perceptions, and we build regression models to test hypotheses regarding developer perceptions (collected when exploring RQ1), such as, “coverage badges signal the importance of tests and therefore attract more pull requests with tests.”
Our investigation reveals that badges are popular in npm, adopted in 46% of packages. The most frequent show the build status or version of the latest release, but dependency managers, code coverage, and download counts are also common. Maintainers intend to signal various qualities with badges, and, indeed, we found among others that build-status and coverage badges correlate with larger test suites and encourage external contributors to include more tests, popularity badges correlate with future gains in downloads, and the introduction of dependency-manager badges correlates with a lasting improvement in dependency freshness. Correlations are particularly strong for assessment signals, i.e., badges that test an underlying quality rather than just stating intentions. Our results provide guidance for package maintainers to make more informed decisions about badge adoptions, being more deliberate about what they intend to signal and how that signal is supported. For users and contributors, our results indicate which badges provide reliable signals.

In summary, we contribute (1) a survey among npm developers, (2) a large scale analysis of 294,941 npm packages and their history, and (3) an in-depth analysis (using multiple regression models and time-series regression discontinuity designs) of 8 hypotheses regarding the effect of badges on various qualities, showing many badges are indeed reasonably reliable signals. Furthermore, (4) we frame our discussion in the context of signaling theory and confirm that badges based on assessment signals are more reliable.

2 THEORETICAL FRAMEWORK

We argue that badge are intended to signal an underlying quality about the project to potential users and contributors. In addition, certain badges (e.g., code coverage) may encourage certain kinds of practices, or attempts to improve visible scores. Therefore, we frame our study in the context of signaling theory and gamification.

Signaling theory. Signaling theory is widely applied to selection scenarios in a range of disciplines, from economics [51] to biology [63]. In these scenarios, the signaler benefits from actions taken by the receiver (e.g., being selected over some alternative), which would not have occurred in the absence of the signal. Signals are observable pieces of information that indicate a hidden quality of the signaler. Receiver cost to interpret the signal tends to outweigh signal accuracy, with less reliable but more easily obtainable signals being preferred by receivers over more reliable signals that are costlier to observe or assess [28]. The classical example in economics is job market candidates signaling their quality through education: holding a degree from a prestigious institution is easily observable and communicates to potential employers the candidate’s ability, which is otherwise less readily observable [51]. Selection scenarios occur routinely in open-source, e.g., choosing which libraries to depend on [11], repositories to watch [49], developers to follow [10, 35] or hire [13, 37], and projects to contribute to [9, 14].

The phenomenon underlying signaling theory is information asymmetry [52], which occurs between those having all the information and those who could potentially make better decisions if they had all the information. To reduce information asymmetry, actors rely on observable signals (e.g., information in CVs). Information asymmetry also occurs in the open-source selection scenarios above. Even if activity traces are typically publicly accessible, not all information is equally accessible, with some requiring specialized mining, e.g., of git histories and issue trackers. For example, to avoid outdated, possibly vulnerable dependencies, developers may adopt a dependency manager, e.g., Gemnasium, David [11, 41], to receive notifications when a dependency is updated; however, to potential users of the package, this practice is very difficult to recognize unless it is made obvious. A badge reporting the result of the same analysis, e.g., David’s dependencies up to date, indicates publicly that the tool is not only enabled, but also used regularly. That is, we argue that repository badges are signals: by making certain information about the project’s code base or practices transparent, badges contribute to reducing information asymmetry between maintainers (insiders) and users and contributors (outsiders).

Research has confirmed that in transparent, social coding environments, observable signals in online profiles are used as indicators of expertise and commitment [20, 38, 48, 55, 56], e.g., Stack Overflow reputation score, GitHub followers and longest activity streak. We expect that repository badges may have a similar effect.

Assessment signals. The literature distinguishes between conventional signals and assessment signals [24]. Both are used, but the former are not inherently reliable; the quality they indicate is established by convention and the signal is typically cheap to produce, therefore easy to fake. The latter are considered more reliable, because “the quality they signal is ‘wasted’ in the production of the signal, and the signal tends to be more expensive to produce for an individual with less of the quality” [48]. Therefore, for a signal to be effective, it must be both: (i) observable, i.e., readily observable by outsiders (otherwise it might go unnoticed), and (ii) costly to produce, such that only higher quality signalers can absorb the cost to produce it [17]. There is a great diversity of badges which, despite being equally observable on READMEs, vary widely in production cost. Some, e.g., David’s dependencies up to date, indicate relatively deep technical qualities that are achieved with specific, arguably costly, practices. Others indicate technical or non-technical but relatively shallow qualities that are easy to look up elsewhere, e.g., license BSD, npm v3.1.0, and a Star, and others still are mere statements of intentions without any automated validation and thus without any associated costs, e.g., code style standard, PFS, and welcome.

Our goal is to evaluate the reliability of badges as signals, by assessing signal fit [17], i.e., the extent to which signals correspond to the desirable unobservable quality of the signaler, described by the strength of the statistical association between public information—the signal—and private information—the unobservable quality.

Gamification. Using game-design elements in non-gaming contexts [23] is mainstream in “social programming” [6, 54] environments, e.g., on Q&A sites like Stack Overflow. These gamification elements are known to motivate existing users [3, 15, 27] as well as attract new users, at the detriment of other platforms [53, 58].

Despite being voluntarily displayed on project READMEs by maintainers, and not “earned” based on performance and automatically displayed by the platform, we argue that repository badges are also gamification mechanisms. Since badges typically represent best practices and tools (e.g., continuous integration, code coverage, dependency freshness), there is little incentive for maintainers to display badges indicative of “bad” practices (e.g., build failing, coverage 1%). Therefore, we expect that the mere presence of GitHub badges correlates with best practices (i.e., they are in a sense “earned”). Furthermore, as we will discuss, badges such as coverage 94% may
We study the adoption and effects of badges in the npm ecosystem. Many studies have looked into specific practices in software ecosystems, including communication [8, 29, 32, 50], change planning [11, 12, 22, 44], dependency updates [5, 19, 30, 34, 41], static analysis [7, 64], testing and continuous integration [31, 60, 65], and many others. Badges and their underlying tools are also ecosystem-level practices. In contrast to prior work, however, we specifically take a broader view on badges as signals beyond individual tools and practices.

3 RQ1: BADGES ON npm

We study the adoption and effects of badges in the npm ecosystem. npm is a package manager and corresponding repository launched in 2010, currently hosting package releases for 500,000+ distinct packages. It was originally designed for Node.js developers, but is used more broadly by many web developers today. As in other package managers and repositories, e.g., Maven and RubyGems, an npm package bundles files (typically JavaScript) with a README and metadata (package.json), which includes the unique release version and dependencies to other packages. A client installs and updates dependencies from the central npm repository. The npm community is generally considered innovation friendly, frequently adopts external packages, values making it easy to contribute and publish packages, and shows a healthy competition between multiple equivalent packages to solve any single problem [1, 11, 22, 61].

We chose npm because: (1) it provides API access to all package releases and metadata, including download counts, (2) most npm packages point to a GitHub repository, (3) the npm registry and GitHub both prominently show the package’s README file, providing a common place where badges are displayed, and (4) the npm community is innovation friendly and broadly experiments with and adopts developer services [11, 41], including cloud-based continuous integration, dependency managers, and badges.

3.1 Research Methods

To explore which badges are common and what they intend to signal (RQ1), we conducted a survey and mined repositories at scale.

Survey design. To gauge perceptions and anticipated effects we designed two online surveys targeting npm package maintainers and corresponding GitHub contributors; the former focused on what maintainers intend to signal about their packages by displaying badges and what effects, if any, they expect badges would have on their users and contributors, while the latter focused on what inferences contributors make about a package given its badges (for the specific questions see appendix). Both groups were asked to name specific badges when answering. We used mostly open-ended questions with free-text answers, and we piloted the survey first.

We extracted contact information and commit counts per person from the git logs of the 10,000 most popular npm packages by downloads, and classified developers as maintainers (≥ 33 % of project commits) or contributors (< 10 %). We resolved multiple aliases using standard heuristics about common first name/last name/email formats [59]. We then randomly sampled two mutually exclusive sets of contributors and maintainers, 300 each, and sent personalized invitation emails (580 succeeded, 294 to maintainers and 286 to contributors). We received 32 maintainer and 57 contributor responses, for a total response rate of 15.3 %. Our respondents have a median 5 years of experience with open source and many surveyed maintainers have contributed to dozens of packages. We analyzed the textual responses using standard open-coding techniques.

Repository mining. We collected a multidimensional longitudinal data set of 294,941 npm packages, as follows. We started mining all npm packages on July 11, 2017 (512,834), then kept only 346,369 that: (1) had metadata on downloads, releases, dependencies, and maintainers, and (2) linked to a GitHub repository. In 11,316 cases when multiple packages linked to the same GitHub repository,1 we kept only the most downloaded one, which further reduced the size of our sample to 322,734. Next, we attempted to clone all GitHub repositories locally, and succeeded for 294,941 packages; the others’ repositories were either private or missing.

To identify badges and their evolution, we used the git history of each repository’s README file.2 This was iterative: We began by matching the markdown expression typically used to insert an SVG badge, i.e., [[Badgen Name](Image URL)](Service URL), but discovered packages with badges added as plain HTML, markdown reference links, and PNG. Consequently, we converted the markdown HTML and matched img tags. To reduce false positives (not all images are badges), we curated a list of services frequently associated with badges, e.g., Travis and Coveralls, and devised regular expressions for classification. We then split all images into other-badge or other-images and iteratively refined the classification rules to define badge classes for specific services, such as Travis. To support the iterative process, we generated web pages showing all found badges per class to inspect them manually for accuracy. As needed, we refined our classification until we reached stability (very few false positives in each class) and until the other-badge category comprised only obscure badges and non-badge images. Overall, we identified 88 kinds of badges (examples in Table 1). By analyzing READMEs longitudinally (following only first parents in the commit history so as not to detect temporary discrepancies between the mainline and other branches), we could identify the dates when badges were introduced or removed. Note that we analyzed badges on GitHub, not npm, GitHub provides a finer temporal granularity of commits, whereas npm typically shows the same README files but only updates them with each new release.

Threats to validity. As typical for a survey, our sample may suffer from selection bias. One should be careful when generalizing the results beyond the studied corpus of npm packages with corresponding GitHub accounts. The npm community has certain characteristics and results may differ in other communities; e.g., Python and Java developers may adopt innovations at a different pace and may use different kinds of tools; not all package repositories show READMEs with badges as prominently as npm does. Results may also differ outside of an open-source context, e.g., when using badges to advertise practices among corporate teams.

1In most cases, code is developed together in one repository, but deployed as multiple packages to reduce user download size when only parts of the project are needed.

2We considered all GitHub-supported filename extensions for markdown files (see github.com/github/markup), most commonly README.md.
We distinguish badges with different signaling types [24] depending on whether they show results of deeper analyses (assessment signal), summarize readily available information, or merely state unvalidated information (both conventional signals).

Most importantly, the survey provides insights into what maintainers intend to signal and what consequences they expect, as well as insights into how contributors and users might interpret badges:

**Quality assurance.** All surveyed maintainers and most contributors (89%) mentioned specific badges related to quality assurance. Most maintainers (84%) stated explicitly that they intend to signal code and development quality—from “having any tests and running them regularly,” to signaling good quality assurance practices more generally, including striving for high coverage, standardizing code layout, and using static analysis tools. Respondents often intended to signal quality broadly, e.g., stating that their badges show that their code was “built with love” or “well written” by an “experienced developer” who pays “attention to quality.”

\[H_1.\] The adoption of quality-assurance badges correlates with other indicators of code quality (metric: test suite size).

**Dependency management.** Dependency-management badges (mentioned in 26% of all responses) indicate whether direct and indirect dependencies refer to outdated versions or even versions with known vulnerabilities. Respondents indicate that good dependency management practices (signaled with these badges) reduce the chance of “conflicting versions of nested dependencies” and indicate attention to updates and security patches.

\[H_2.\] The adoption of dependency-management badges correlates with increased user confidence and attractiveness (metric: downloads).

**Information and navigation links.** Though mentioned by many (56%), usually as convenient shortcuts, we expect that link-related badges do not provide much signaling impact beyond other statements or links in the package description. Respondents suggested the following potential effects: Showing the latest npm version of a package as a badge might encourage more users to more quickly update to the latest version. Badges linking to npm, Heroku, CDNJS, or other external sites make it convenient for users to download and experiment with a package and may thus attract more users. Badges indicating high code quality (including style conventions) might encourage more users to attempt to read the code. Badges indicating explicitly that the package is open for contributions may lower the bar for new contributors. Badges indicating licenses can make it easier for users to make adopting decisions. Nonetheless, following signaling theory, we expect at most marginal effects.

\[H_3.\] The adoption of a link-related badge does not correlate with either popularity or code quality.

**Popularity.** Popularity badges were mentioned in 25% of responses. Five maintainers explicitly mentioned that they intend to signal package popularity, to “instill confidence in new users,” and several contributors indicated that popularity is an important signal when deciding between similar packages, because the “wisdom of
the crowd” has deemed a package safe to use or of high quality. In addition, popularity is seen as a sign for likely sustainability. Especially the npm-downloads badge was mentioned frequently as an important addition to GrtHtus signals like stars and number of contributors. While downloads can be looked up on npm, a badge was often mentioned as a more convenient and direct means.

**H₆.** The adoption of popularity-related badges in popular packages correlates with more future downloads (metric: monthly downloads).

**Support.** Support badges were mentioned infrequently (9%), but some interpret them to signal “dedication to offering support.”

**H₇.** The adoption of a support-related badges correlates with more responsive support (metric: issue closing time).

Too many badges. An interesting facet to explore that came up a few times in the survey is that packages with too many badges can be perceived as cluttered or “trying too hard” and may be taken less seriously: “People tend to overwhelm visitors with too many (useless) badges, thus creating a contra effect and loosing the initial purpose of having useful information.” Hence our final hypothesis: H₈. The number of badges correlates non-linearly with popularity.

### 3.3 Badge Popularity and Adoption

Of our 294,941 npm packages, 46% have at least one badge. Adoption statistics per badge (Table 1 for the most popular) reveal that only few badges are broadly adopted. Build status and version badges are by far the most common, followed by dependency managers, test coverage, and download statistics. A longitudinal analysis of badge adoption (Fig. 1) shows that quality-assurance badges (TRAVIS CI primarily) were adopted early and quickly, but seemed to have reached saturation (roughly every third new package adds a quality-assurance badge). Other kinds of badges have been adopted later and at lower rates; most seem to have reached saturation as well.

Badges tend to be adopted in groups and are not frequently changed afterward. Of 136,865 packages with badges, 66% adopted multiple kinds, of which 82% did so within 24 hours. Combinations of badges follow their overall popularity and typically involve quality-assurance, information, and dependency-manager badges. While there are often multiple badge-related commits when badges are first adopted (including temporary removal), only 13% of packages changed any badge more than 15 days after adopting their first. Permanent badge removal is also rare (11% of packages).

### 4 RQ₂: EFFECTS OF BADGES

After providing an overview of badge adoption in practice and collecting hypotheses about what their effects might be, we can now test these hypotheses. In particular, we want to test to what degree the presence of badges correlates with certain expected qualities of the package, which we operationalize with measures such as downloads or rate of external contributions with tests.

#### 4.1 Data and Methods

**Data analysis.** To evaluate the badges’ signaling reliability we proceed in three complementary steps per hypothesis.

**Step 1: Correlation.** We look for correlations between presence of badges and differences in the quality they are signaling. This analysis takes the outsider’s perspective of somebody looking at a repository now, and explores whether badges are reliable signals for certain qualities, independent of causal relationships, confounds, or historic trends. In line with our hypotheses, we typically analyze categories of badges together, as badges within a category can be expected to represent similar signals. We use the non-parametric MW test to compare distributions and report Cliff’s delta.

**Step 2: Additional information.** Badges may correlate with various qualities, but still be redundant or weaker predictors of those qualities, compared to other signals. Here we explore whether badges add information to explain the qualities beyond readily-available signals, e.g., stars and issues shown on GrtHtus, downloads and dependent packages shown on npm. To assess the information gain with badges, we model the variability in the underlying quality using hierarchical linear regression. Specifically, we compare the fit of a base model, which includes only readily available signals and control variables, and a full model, which adds badge predictors. The difference, i.e., the added explanatory power attributable to badges, suggests their association with deeper-level qualities.

We use different types of linear models depending on the response variable (details with each result below), but always follow the same procedure for model fit and diagnostics. First, we conservatively remove outliers for predictors with exponential distributions, i.e., those values exceeding $k(1 + 2/n)\text{median}(x) + \theta$ \cite{43}, where $\theta$ is the exponential parameter \cite{47}, and $k$ is computed such that not more than 1% of values are labeled as outliers; among these are high leverage points that disproportionately affect regression slopes, affecting robustness. Second, we diagnose the models, checking for multicollinearity (variance inflation factor, or VIF \cite{2} below 3, except between the interaction terms and their comprising factors, which is expected), and checking if modeling assumptions hold (no significant deviation from a normal distribution in QQ-plots, randomly distributed residuals across the range). Finally, we consider model coefficients important if they are statistically significant at 0.05 level, and we estimate their effect sizes from ANOVA analyses.

**Step 3: Longitudinal analysis.** Previous steps look at differences between packages that have badges now. A longitudinal analysis could reveal whether packages with badges evolve differently than without, and whether introducing a first badge (the intervention) has an observable effect on the package’s quality as the package evolves. Here we use a powerful time series analysis method—time series regression discontinuity design (RDD) \cite{18}—to evaluate longitudinal effects of displaying the first badge. With RDD, we estimate the magnitude of a function’s discontinuity between its values at points just before and just after an intervention. RDD is based on the assumption that in the absence of an effect, the function’s trend after the intervention
would be continuous in the same way as prior to the intervention. We consider the earliest display of a badge as the intervention, and compare data about the signaled underlying qualities in 18 monthly windows, 9 months on each side, centered around the adoption month. Aligning the history on the intervention date, we can compare 9-month trends before/after an intervention across many packages, looking for sudden jumps at the intervention and long-term differences in trends. We use multiple regression to estimate the trend in the response before the badge adoption (variable time in our models, e.g., Table 2), and the changes in level (variable intervention) and trend (variable time_after_intervention) after the badge adoption, cf. [65]. By controlling for confounds in the multiple regression (including presence of other badge classes), we evaluate whether the change could be attributed to other factors than the intervention.

To account for projects that adopt multiple badges (§3.3), we align on the first badge adoption but add controls for the adoption of other badges if they occur within 15 days of the first badge adoption. As usual, the controls allow us to isolate the effects of individual badges. The intuition is that, given the resolution of our analysis, interventions within 15 days can be considered as simultaneous. Later adoptions are not considered (existing only in 17,217 packages), but also do not systematically bias any specific month.

**Repository mining and operationalization.** Using the same dataset (294,941 packages; §3.1), we collected, besides badge adoption, data from three sources, both for a current snapshot for Steps 1 and 2 (July 2017) and longitudinally for Step 3 (monthly, January 2010—July 2017): (1) package metadata on npm using the npm API (e.g., downloads, releases, dependencies), (2) GitHub project data using the GitHub API and GHTorrent [26] (e.g., contributors, issues, pull requests), and (3) the package’s git repository, cloned locally (e.g., code size, badges, tests). Specifically, we collected the following data. Representing readily-available signals of an npm package, we collected download statistics (npm), the number of stars and issues (GitHub), commit counts (git), and the size of the README file in bytes (git). As further controls, we collected the package’s age (npm), the size of the code base in bytes (git), the number of dependencies per package (npm), and the number of dependents, i.e., other packages depending on the package (npm).

We operationalize the qualities in our hypotheses with the following metrics for which we collect both current and historic values:

- **Indicator of code quality (H1):** We measure the (relative) size of the test suite, by identifying files and directories that refer to test code, reusing the detector maintained by the package search service npmjs.io. We measure size in bytes, as it is robust to different test frameworks and file formats.
- **Indicator of contribution quality (H2):** We measure how many pull requests contain test code (GitHub), using the same mechanism to identify test-related files.
- **Indicator of users and popularity (H3, H6, H5, H5):** We collect download counts from npm.
- **Indicator of fresh dependencies (H4):** Inspired by recent work [19], we design a freshness score that performs a similar analysis to dependency managers, based on how many dependencies declared in a package have a newer version that existed on npm at the time (git, npm). For each package, we compute the average freshness of all direct dependencies, i.e., the average distance between the specified version and the most recent release. Assuming that larger version changes require more effort to update, we assign a distance of 1 for each change of a patch version, 5 per minor version change, and 20 per major version change. An up-to-date dependency has distance 0. Details of the metric, including how version ranges are handled, are described in the appendix.
- **Indicator of support (H7):** We collect the average time between when an issue is first posted and when it is closed, for all GitHub issues closed in July 2017, as the package’s average issue latency.

**Threats to validity.** Imperfect measures. Our operationalized measures can only capture some aspect of the underlying quality indicated in the survey. For example, large test suites are an indicator of good testing practices, but neither the only nor the most reliable indicator. However, while individual packages may vary in their practices, the large size of the data sets we study implies a reduction to the mean in terms of individual behavior. Therefore, we expect that by averaging over thousands of packages in our regression models, our (imperfect) measures will meaningfully reflect the intensity and directionality of underlying relationships between badges and project qualities.

To ensure internal validity, in our analysis Steps 2 and 3, we remove outliers and explore different operationalizations, whenever practical, to improve robustness. We further account for many covariates, especially other readily available signals, but we may miss other confounds easily observable by humans but hard to assess automatically, e.g., the quality of the documentation beyond our proxy of README file size. Consequently, one must be careful when generalizing our results beyond the studied measures.

**Badges vs. practices.** Typically, we cannot distinguish effects of practice adoption from effects of badge adoption; hence, our results can only be interpreted as exploring the reliability of the signal that a badge provides. Our analysis also does not consider the specific value shown on the badge (e.g., current coverage). Still, we expect that badges are usually adopted to signal good practices, a badge highlighting that a practice is not followed (e.g., low coverage) might have a negative effect. We control for this indirectly in many models, e.g., by controlling for popularity in our analysis of downloads (§4.3), but a more detailed analysis is outside the scope of this paper.

**Beyond correlations.** None of our three analysis steps can establish a causal relationship between badges and the studied qualities. Still, note how our three steps investigate each hypothesis from complementary perspectives: Step 1 checks for basic correlations, Step 2 explores the role of covariates and whether badges provide additional insights, and Step 3 looks at whether adopting badges leads to observable differences over time. The careful multi-faceted analysis, spearheaded by the sophisticated time-series regression discontinuity design, can indicate that correlations are not spurious and associate with some underlying phenomenon.

### 4.2 Signals of Updated Dependencies (H4, H5)

We explore our hypotheses grouped by response variable and start with a discussion of dependency freshness, as it clearly illustrates our 3-step analysis. We expect that dependency-manager badges correlate with more up-to-date and secure dependencies (H4), operationalized with our freshness metric (see Sec. 4.1), and expect at most a marginal effect from information-related badges (H5).
An Empirical Study of Repository Badges in the npm Ecosystem

### Table 2: Dependency freshness models.

<table>
<thead>
<tr>
<th></th>
<th>Basic Model</th>
<th>Full Model</th>
<th>RDD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>response: freshness = 0</td>
<td>response: freshness = 0</td>
<td>response: log(freshness)</td>
</tr>
<tr>
<td>17.3% deviance explained</td>
<td>17.4% deviance explained</td>
<td>R² = 0.04</td>
<td>R² = 0.35</td>
</tr>
<tr>
<td>Coefs (Err)</td>
<td>LR Chisq</td>
<td>Coefs (Err)</td>
<td>LR Chisq</td>
</tr>
<tr>
<td>(inter.) 3.54 (0.03)**</td>
<td>3.50 (0.03)**</td>
<td>1.45 (0.09)**</td>
<td>1.49 (0.09)**</td>
</tr>
<tr>
<td>Dep = -1.78 (0.01)**</td>
<td>-2077.8***</td>
<td>-1.79 (0.01)**</td>
<td>-3292.8***</td>
</tr>
<tr>
<td>RDep = 0.22 (0.04)**</td>
<td>610.3***</td>
<td>0.21 (0.01)**</td>
<td>560.6***</td>
</tr>
<tr>
<td>Stars = -0.08 (0.00)**</td>
<td>301.4***</td>
<td>-0.09 (0.00)**</td>
<td>311.2***</td>
</tr>
<tr>
<td>Cout = -0.24 (0.01)**</td>
<td>500.3***</td>
<td>-0.25 (0.01)**</td>
<td>548.7***</td>
</tr>
<tr>
<td>lastU = -0.65 (0.03)**</td>
<td>12208.9***</td>
<td>-0.64 (0.01)**</td>
<td>11537.9***</td>
</tr>
<tr>
<td>hasDM = 0.24 (0.03)**</td>
<td>116.1***</td>
<td>0.45 (0.08)**</td>
<td>2.43</td>
</tr>
<tr>
<td>hasInf = 0.11 (0.02)**</td>
<td>48.3***</td>
<td>0.04 (0.05)</td>
<td>0.45</td>
</tr>
<tr>
<td>hasInf; hasDM = 0.09 (0.04)</td>
<td>1.9</td>
<td>-0.32 (0.10)**</td>
<td>hasOther = 0.01 (0.01)</td>
</tr>
<tr>
<td>time = 0.03 (0.00)**</td>
<td>0.93 (0.03)***</td>
<td>1373.22***</td>
<td></td>
</tr>
<tr>
<td>time; after intervention = 0.11 (0.00)**</td>
<td>455.56***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time; after intervention; hasDM = -0.10 (0.01)</td>
<td>230.36***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time; after intervention; hasInf = -0.00 (0.01)</td>
<td>1.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time; after intervention; hasDM; hasInf = 0.03 (0.01)</td>
<td>10.62**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Dep: dependencies; RDep: dependencies; Cout: contributors; lastU: time since last update; hasDM: has dependency-manager badge; hasInf: has information badge; hasOther: adopts additional badges within 15 days.

### Correlation.
Among the analyzed packages that had any dependencies, 37% had all up-to-date dependencies (freshness = 0). Supporting H1 and, surprisingly, contradicting H5, Fig. 2a reveals a small, but statistically significant difference: packages with a dependency-manager badge or an information badge tend to keep their software updated at a generally faster rate than packages without. To further explore these differences, we also find that dependency-manager badges are proportionally adopted by packages with more dependencies.

### Additional information.
To test if the presence of badges correlates with deeper-level freshness indicators beyond other readily available signals, we fit a hurdle regression: a logistic model of the likelihood of freshness = 0 and a linear regression model of packages with outdated dependencies. This hybrid modeling approach is necessary due to the bimodality of the data (Fig. 2a). As described in §4.1, the base model attempts to explain freshness given readily-available signals (stars, dependencies, contributors) and a control for time since package was last updated; the full model additionally models the presence of dependency-manager badges and information badges and their interaction, with controls for other badges adopted within 15 days.

We show the base and full logistic regression model (predicting whether a package has any outdated dependencies) in Table 2. The base model explains 17.3% of the deviance; the full model explains 17.4%. The difference is small but statistically significant (DeLong’s test for correlated ROC curves p < 0.001). The number of dependencies and the time since the last update explain the majority of the deviance, but dependency-manager badges add explanatory power: the odds of having fresh dependencies increase by 27% (exp(0.24) 1.27) for packages with dependency-manager badges (H1). Surprisingly, the effect of information badges is comparable: a 17% increase in odds (H2). For the linear regression (predicting the severity of outdated dependencies for packages with outdated dependencies), we see a similar small but significant difference between base (22.4%) and full models (22.8%), and similar behavior of the badge predictors.

### Longitudinal analysis.
We collect a sample of 3,604 packages that had dependencies and satisfy the RDD requirements (9 months activity before and after the adoption of their first dependency-manager badge), and keep 1,761 that had at least one month with freshness ≠ 0 during the +/- 9 (to avoid data bimodality issues). A trend is already visible from the longitudinal freshness data plotted for those packages in Fig. 3a, but a corresponding RDD model controlling for confounds (column RDD3 in Table 2) confirms that: The adoption of (any) badges correlates to a strong improvement in freshness (see the intervention term in the model), by about a factor 2.5 on average, after which freshness slightly decays again over time (the interpretation derives from the sum of the coefficients for time and time after intervention in the model, cf. RDD [65], which expresses the slope of the post-intervention trend). As hypothesized, the adoption of a dependency-manager badge is associated with a longer-lasting effect on freshness than other badges (see the interaction time after intervention * hasDM in the model); ≈ 80% slower decay). The interaction effect of information badges is negligible.

### Discussion.
Results from all three steps confirm H4 that dependency-manager badges signal practices that lead to fresher dependencies. However, the effect is not exclusive to dependency-manager badges; we speculate that any maintenance task involving README updates with more badges might involve other project cleanup. Still, the effect of dependency-manager badges is both stronger and longer lived, as signaling theory would predict given the assessment-signal nature of dependency-manager badges. The results are similar for a security score that counts known vulnerabilities, similar to the Snyk and nsp services (not shown due to space restrictions).

### 4.3 Signals of Popularity (H2, H5, H6, H8)
We expect that adopting quality-assurance and popularity badges correlates with increases in downloads (H2, H6), and at most a marginal effect from information-related badges (H5). We follow the same three steps, analyzing monthly download counts as response.

### Correlation.
Supporting our hypotheses, comparing downloads for packages with and without badges shown in Figure 2b, we can see that for all categories of badges, those packages with a badge tend to skew toward more downloads than packages without. The differences are generally small, but statistically significant. As hypothesized, the effect for information badges is smaller than for quality assurance and popularity badges.

### Additional information.
We fit nested negative binomial regression models to explain the information added by badges for explaining downloads over readily-available signals and controls. Since already popular packages might benefit from badges in a different way than less popular ones, we also model a dummy variable is-Popular, that indicates whether the package was among the 10% most downloaded packages in the prior month (June 2017). In the model with badges, we then explore interactions with the is-Popular dummy. Both models (see appendix) explain downloads well: The basic model explains 66% of the deviance (with expected behavior of all main predictors), and the full model explains 86%. The difference is statistically significant; all badge categories have significant effects. Modeling interactions reveals that quality-assurance badges

---

1. Note that all packages modeled in the RDD adopted some badge during the alignment month, hence the control hasOther is subsumed by experimental design.
2. exp(0.24) factor decrease in freshness score; note the log-transformed response, hence the exponentiation here.
have a stronger effect in already popular packages: holding other variables constant at their mean values, popular packages with a quality-assurance badge tend to have about 2.2 times more downloads than comparable packages without; the effect is marginal for less popular packages. For popularity badges and information badges, we see smaller effects on popular packages (1.2× more, and 0.78× fewer downloads respectively). Again, the models show that badges explain additional aspects of popularity.

Separately, we also fit a negative binomial regression to model the effect of the number of badges displayed (and controls). The model (see appendix) suggests a nonlinear relationship in popular packages, with a predicted inflection point at 5 badges, which supports H₅: Packages with many badges tend to have fewer downloads. The effect for less popular packages is negligible.

Longitudinal analysis. We compile a set of 1,762 packages that satisfy the RDD requirements and have monthly download counts after March 2015 (a limitation of the npm API). Specifically, we align on the adoption month of their first badge for categories pertaining to our hypotheses. Of these, 1,414 packages adopted a quality-assurance badge, 892 an information badge, and 366 a popularity badge. Prior to modeling, we inflation-adjusted each package’s monthly download counts to account for npm’s natural growth over time, based on the average download growth of 10,000 randomly-sampled packages with at least 10 dependents each, that existed the entire period.

A visual inspection of trends around the first badge adoption (Fig. 3b) shows potential intervention effects; an RDD model controlling for confounds (see appendix) suggests a small positive trend prior to the intervention, a sizeable positive discontinuity in download counts at badge adoption (33% increase on average), and, surprisingly, a small negative slope after the intervention. That is, badge adoption correlates with a sudden popularity boost, but the acceleration is not sustained over time. The post-intervention decay for quality-assurance badges is slower than average (8.8%)—in other words, the positive intervention effect lasts longer.

Discussion. Together, our three analysis steps paint a mixed picture for our hypotheses: All badges correlate with more downloads in general, and a longitudinal analysis shows also positive intervention effects for quality-assurance and popularity badges (H₂, H₅, H₆), but the pace is not sustained over time. Still, not all packages with badges show similar effects: Those with assessment-signal badges (namely quality-assurance) tend to maintain the popularity boost that correlates with badge adoption longer, as signaling theory would predict. At the same time, too many badges may seem counterproductive (H₆).

4.4 Signals of Test Suite Quality (H₁, H₃)

We expect that the adoption of quality-assurance badges correlates with increases in test-suite quality (H₁), operationalized as the size of the test suite, and again at most a marginal effect from information badges (H₃). We follow our common 3-step analysis.

Correlation. The distribution of test suite sizes across packages is bimodal (Figure 2c), as we cannot detect any test code in 39.3% of the packages. Packages with quality-assurance badges almost always had tests (93.5%). Among the packages with tests, those packages with quality-assurance badges tend to skew toward larger
test suites, with statistically significant differences, supporting our hypothesis. Surprisingly, so do packages with information badges.

**Additional information.** Given the bimodal distribution of test suite sizes, as for dependency freshness (Sec. 4.2), we fit a hurdle regression (see appendix), modeling separately the likelihood of having any tests (logistic regression) and the test-folder size for packages with non-empty test suites (negative binomial regression). In addition to the usual controls, we control for the package size (larger packages are expected to have larger test suites).

The base logistic model is plausible for explaining whether a package has any tests (4.4 % deviance explained). The full model with quality-assurance and information badges fits the data significantly better (22.6 % deviance explained), most being attributed to quality-assurance badges. The full negative binomial regression model shows a small improvement for explaining test-suite size (78.2 % to 78.4 %). In both cases (having any tests and size of test suite), the models show a strong positive effect of quality-assurance badges ($H_1$): on average, the odds of having tests increase by a factor 18 for packages with quality-assurance badges; among packages with tests, those with quality-assurance badges are expected to have 18.3 % larger test suites; other variables held constant. Information badges have a marginal effect on their own but interact with quality-assurance badges, strengthening their effect slightly ($H_3$).

**Longitudinal analysis.** To avoid technical problems with the bimodal distribution, we study only how first badge adoption correlates with the growth of an existing test suite, not whether it coincided with creating a test suite in the first place. To this end, we assemble a longitudinal sample of 2,855 packages that adopted a badge ($H_2$): on average, the models show a small improvement for explaining test-suite size (78.2 % to 78.4 %). In both cases (having any tests and size of test suite), the models show a strong positive effect of quality-assurance badges ($H_1$): on average, the odds of having tests increase by a factor 18 for packages with quality-assurance badges; among packages with tests, those with quality-assurance badges are expected to have 18.3 % larger test suites; other variables held constant. Information badges have a marginal effect on their own but interact with quality-assurance badges, strengthening their effect slightly ($H_3$).

Discussion. As expected by the surveyed maintainers, all three steps indicate that quality-assurance badges are a good signal for a project having some tests, though they are a weaker signal for the size of the test suite ($H_1$). Since the adoption of quality-assurance badges is correlated with an intervention effect on test suite size, we conclude that they may act as gamification mechanisms. Again, we see a marginal change at the intervention for information badges; the former can likely be explained through some overall perception of well-managed projects ($H_3$).

4.6 Signals of Support ($H_3$)

We expect that the adoption of support badges correlates with better support ($H_3$), operationalized as the average time to close issues on GitHub.

Unfortunately, there are relatively few packages that have adopted support badges (see Tab. 1). We limited our analysis to 826 packages with at least 100 issues overall, and having at least one issue created after July 1, 2017 and closed before Aug. 20; 397 of these packages have support badges. In contrast to our hypothesis, we observe that projects with a support badge have, on average, 20 % longer closing times than those without; however, we also observe that projects with support badges receive twice the number of issues on average. When controlling for confounds, including the number of issues (base model 21.2 %, full model 21.3 %), we find a similar effect: Projects with support badges have a 30.2 % higher latency on average. Removing packages that do not fulfill the RDD constraints leaves only 76 packages. Controlling for the same confounds, the RDD model only suggests that issue latency decreases with time, but does not show an effect of support badges.

Overall, our findings for hypothesis $H_3$ are negative and opposite to our expectations. The results in Steps 1 and 2 are statistically significant, but all steps suffer from small data sets. We have no explanation for this effect beyond conjecturing that external support platforms (GITTER and SLACK) handle support requests in a way that is not captured by our operationalization with GitHub issues.
5 DISCUSSION AND CONCLUSION

We studied repository badges, a new phenomenon in social coding environments like GitHub, with previously unknown effects.

Research questions. We answered two research questions. First, exploring the most common types of badges and their intended signals ($RQ_1$), we found a diversity of badges and signals (Table 1): some are merely static displays of (existing) information, others aggregate information that is otherwise much harder to observe, e.g., reflecting build status, up-to-dateness of dependencies, and test coverage. As predicted by signaling theory ($§2$), our survey revealed that package maintainers displaying badges have clear signaling intentions doing so, and many contributors interpret badges when evaluating packages. Second, exploring the fit of the signals to underlying qualities hypothesized by our survey participants ($RQ_2$), we found that packages with badges tend to skew towards having more of the quality they signal, with stronger effects for the non-trivial quality-assurance and dependency-manager badges. Moreover, the presence of badges consistently adds explanatory power, albeit little, to readily-available signals. Time-series analysis further revealed that the introduction of quality-assurance badges tends to correlate with positive intervention effects: The underlying qualities they signal tend to improve immediately, especially improved dependency freshness and more tests in pull requests.

Gamification effects. Our results also revealed gamification effects ($§2$). Clear examples are the quality assurance badges displaying test coverage percentages ($§4.5$), which we found to correlate with developers increasing the size and, arguably, quality of their test suites. Dependency-management badges ($§4.2$) could also be seen as a gamification mechanism: by making the up-to-dateness of dependencies noticeable, they create an incentive for developers to make the most out of their dependency-management tools by staying up to date. One can imagine other badges with gamification value, e.g., around bug fixing, being used in the future to encourage desirable practices. However, we advocate caution in implementing gamification and acknowledge the risk of creating the wrong incentives, e.g., writing tests only to maximize coverage.

The power of assessment signals. As discussed in $§2$, signaling theory makes an interesting distinction between conventional signals and assessment signals, where the latter require that the signaler actually possesses the signaled quality. Although a few survey participants suggested intended signals for information badges and support badges ($§3.2$), the theory predicts that, without an associated cost or analysis, they are less reliable signals. Our results seem to confirm that quality assurance, dependency manager, and popularity badges (mostly assessment signals) provide more reliable signals than information badges (mostly conventional signals). For example, the effect on freshness is stronger and longer lasting for the assessment signal. Interestingly, information badges often do signal something, even outside their domain of cost (in fact, in data exploration we often also found other small effects of other badges on various qualities). We speculate that badges are often adopted during a general maintenance phase in which also test suites or dependencies are improved, but we expect that only assessment signals correlate with lasting change ($§4.2$).

The results encourage more badges to be designed as assessment signals. For example, several badges that currently only state intentions, such as code style standard or gather join chat, could be redesigned to report analysis result for the underlying quality, such as past conformance to coding standards or responsiveness to support requests. Such badges may encourage stronger conformance and even accrue gamification benefits. Interestingly, Slack offers multiple kinds of badges, including one inviting users to join ($slack$ join) and one that shows the number of currently active and registered users ($slack$ 6/150). In practice though, most package maintainers with a Slack badge adopt the former (conventional signal) rather than the latter (assessment signal). Our results indicate that maintainers, when they have the choice and are serious about signaling their dedication, should adopt the assessment-signal badge.

Badges vs practices. We emphasize that in most cases effects associated with non-trivial badges are inextricably linked to effects associated with the badges’ corresponding tools or practices. For many kinds of badges, the badge is the only (easily) externally observable indicator of the tool’s use (e.g., dependencies up to date, docs). While effects associated with adopting continuous integration [31, 60, 65], static analysis [7, 64], and dependency-management tools [41] on software development practices have been studied by prior work, our approach is unique in that we focus on the signaling dimension added by repository badges to these and other tools, previously overlooked in the software engineering signaling literature (e.g., [20, 38, 48, 55, 56]). Further delineating effects associated with badges from effects associated with the tools themselves goes beyond the scope of this work, but some preliminary experiments with the continuous integration service Travis, which is detectable independently of its badge through a Travis.yml configuration file in the repository, suggest that while the observed effects of adopting the tool are similar to the effects at the badge adoption time, badges seem to have a small amplifying effect. Pending further validation, this is encouraging for researchers, in that badges might be a reliable indicator for longitudinal studies of practices that are hard to detect otherwise, because there are no clear traces of the practice in the repository, beyond the badge.

Implications for practitioners. Our results provide guidance on which qualities are usually signaled with badges and which signals tend to be more reliable. Therefore, package maintainers can make more deliberate choices about badges (e.g., limiting conventional signals), service developers can design badges more carefully (e.g., providing an assessment signal based on some analysis of past conformance), and package users and contributors can decide which badges to use as indicators of underlying practices and as starting points to investigate deeper qualities. Overall: badges, improve reliable.

Acknowledgements. Many thanks to respondents to our survey! Trockman was supported through Carnegie Mellon’s Research Experiences for Undergraduates in Software Engineering. Kästner and Zhou have been supported in part by the NSF (awards 1318808, 1552944, and 1717022) and AFRL and DARPA (FA8750-16-2-0042). Vasilescu has been supported in part by the NSF (award 1717415).

A ADDITIONAL MATERIAL

Plots, models, and exact survey questions are available online at https://github.com/CMUSTRUDEL/npm-badges.
An Empirical Study of Repository Badges in the npm Ecosystem

ICSE ’18, May 27–June 3, 2018, Gothenburg, Sweden

REFERENCES


