How Social Q&A Sites are Changing Knowledge Sharing in Open Source Software Communities

Bogdan Vasilescu[1,2], **Alexander Serebrenik**[1], **Premkumar Devanbu**[2], **Vladimir Filkov**[2]

[1]Eindhoven University of Technology, The Netherlands, {b.n.vasilescu, a.serebrenik}@tue.nl

[2]University of California, Davis, USA, {devanbu, filkov}@cs.ucdavis.edu

ABSTRACT

Historically, mailing lists have been the preferred means for coordinating development and user support activities. With the emergence and popularity growth of social Q&A sites such as the StackExchange network (*e.g.*, StackOverflow), this is beginning to change. Such sites offer different socio-technical incentives to their participants than mailing lists do, *e.g.*, rich web environments to store and manage content collaboratively, or a place to showcase their knowledge and expertise more vividly to peers or potential recruiters. A key difference between StackExchange and mailing lists is gamification, *i.e.*, StackExchange participants compete to obtain reputation points and badges. In this paper, we use a case study of R (a widely-used tool for data analysis) to investigate how mailing list participation has evolved since the launch of StackExchange. Our main contribution is the assembly of a joint data set from the two sources, in which participants in both the `r-help` mailing list and StackExchange are identifiable. This permits their activities to be linked across the two resources and also over time. With this data set we found that user support activities show a strong shift away from `r-help`. In particular, mailing list experts are migrating to StackExchange, where their behaviour is different. First, participants active both on `r-help` and on StackExchange are more active than those who focus exclusively on only one of the two. Second, they provide faster answers on StackExchange than on `r-help`, suggesting they are motivated by the *gamified* environment. To our knowledge, our study is the first to directly chart the changes in behaviour of specific contributors as they migrate into gamified environments, and has important implications for knowledge management in software engineering.

Author Keywords

Crowdsourced knowledge; social Q&A; mailing lists; open source; gamification.

ACM Classification Keywords

H.5.3. [Information Interfaces and Presentation (e.g. HCI)]: Computer-supported cooperative work

CSCW'14, February 15–19, 2014, Baltimore, Maryland, USA.

ACM 978-1-4503-2540-0/14/02...$15.00.
http://dx.doi.org/10.1145/2531602.2531659

INTRODUCTION

Historically, mailing lists have been the preferred medium for coordinating development and user support activities [16, 31,32]. In particular, mailing lists have been viewed as the *de facto* communication medium between *knowledge seekers* (*e.g.*, users of the software asking for support) and *knowledge providers* (*e.g.*, other users, more knowledgeable about the topic, or the developers themselves) in models of knowledge sharing in open source [32]. The two categories of knowledge actors have been reported to co-exist in a symbiotic relationship, wherein "the community learns from its participants, and each individual learns from the community" [32]. However, their motivations for participation may differ. For instance, knowledge seekers may directly benefit from having their problems solved, while knowledge providers may be motivated intrinsically (*e.g.*, by altruism), or by learning about the problems experienced by other users [20, 32].

Recent years have witnessed the emergence and growing popularity of software-development-related social media sites, such as GitHub[1] (coding), Jira[2] (issue tracking), or the StackExchange network (question and answer websites, *e.g.*, StackOverflow for "professional and enthusiast programmers,"[3] or CrossValidated for "statisticians, data analysts, data miners and data visualization experts"[4]). Such sites are rapidly changing the ways in which developers collaborate, learn, and communicate among themselves and with their users [4, 8, 9, 30, 34]. Moreover, they are offering different socio-technical incentives to their participants, *e.g.*, rich Web 2.0 platforms to store and manage content collaboratively, or a place to showcase their knowledge and expertise more vividly to peers and potential recruiters [8]. In addition, StackExchange sites employ *gamification* [11] to engage users more: questions and answers are voted upon by the community; the number of votes is reflected in the poster's *reputation* and *badges*; exceeding various reputation thresholds grants access to additional features (*e.g.*, moderation rights on topics and posts); reputation and badges can also be seen as a measure of one's expertise by potential recruiters [8], and are known to motivate users to contribute more [1, 2, 10, 42]. Activity on StackExchange sites can also elevate one to celebrity status within the developer community (see, *e.g.*, the discussion around Jon Skeet[5], the most prolific contributor to StackOverflow).

[1]`https://github.com`
[2]`http://www.atlassian.com/software/jira`
[3]`http://stackoverflow.com`
[4]`http://stats.stackexchange.com/`
[5]`http://meta.stackoverflow.com/q/9134`

Naturally, the richer user interfaces, wider audiences, or different incentives and motivations for participation inherent in social Q&A sites are challenging the supremacy of mailing lists as the *de facto* communication medium between knowledge seekers and knowledge providers. For example, StackOverflow is known to provide good technical solutions [25] and to provide them fast [23]. At the same time, mailing list participants are signalling the need for more modern support,[6] and are even promoting a transition to StackExchange.[7] Our goal here is to study in detail the effects of such a transition on contributors and their work. Are mailing list participants transitioning to StackExchange? If so, do they behave differently on StackExchange than on the mailing list?

To study the phenomena associated with such a transition, we need a *longitudinal data set* combining mailing list and StackExchange activity, wherein participants overlap. In this paper we create such a data set for R [28], a popular data analysis software, by integrating mailing list activity and StackExchange activity (the latter under the `[r]` tag). Using this data set, we find that:

- activity on `r-help` (the main user support mailing list for R) has been consistently decreasing since around 2010 (*i.e.*, participants are asking fewer and fewer questions), while at the same time the number of R-related questions asked on the two main StackExchange sites for R (CrossValidated and StackOverflow) has been accelerating.
- participants in the two communities overlap, but different categories of `r-help` contributors are "attracted" differentially by StackExchange. For example, the proportion of mailing list users active on StackExchange is much higher among R developers than non-developers.
- the levels of activity for `r-help` participants who are also active on StackExchange differ relative to those who restrict themselves to either the mailing list or to StackExchange: those participating in *both* `r-help` and StackExchange are more active.
- knowledge providers active in both communities answer questions significantly faster on StackExchange than on `r-help`, and their total output increases after the transition to StackExchange.

Apart from uncovering interesting phenomena, these findings reveal that knowledge management in open source is changing, and that the different socio-technical incentives offered by Q&A sites such as StackOverflow enhance participation in these communities, facilitate user contributions and foster productivity. Therefore, our findings could inspire start-up open source or commercial projects looking to establish user support platforms, or stakeholders interested in knowledge management in general.

The rest of the paper is organised as follows. We first focus on the background underlying knowledge-sharing in open source and present our research questions. Then, we describe the data set and data gathering process, followed by our methods, results, and concluding sections.

[6] `http://goo.gl/0i4j1n`, `http://goo.gl/6h8bMa`
[7] `http://goo.gl/aeoBJA` `http://goo.gl/OtzZT5`

Background

The importance of user support for the adoption, growth, and success of open source projects is well-recognized [3,20,35]. Traditionally, user support was organised through mailing lists, forums, user groups, etc. However, as new venues of information and tools for information access emerge (*e.g.*, weblogs, wikis, social Q&A sites), people's online information seeking behaviour is also evolving [12,29]. When it comes to user support activities around open source software, though, what is common among both traditional and new venues of information is that this "mundane but necessary task" [20] is typically carried out by unpaid volunteers. Often, the developers themselves take part in these activities, but thriving open source projects also succeed in enlisting some of their users to offer assistance to peers. But what motivates knowledge providers to answer other people's questions? Together with online information seeking behaviours, these motives are also evolving.

In traditional information-sharing venues, *knowledge-providers* participate for reasons related to, albeit different, than those of *developers* contributing to open source. Developers contribute for reasons such as: a direct need for the software; enjoyment of the work itself; or the enhanced reputation arising from high-quality contributions [20]. Knowledge-providers, on the other hand, are reportedly motivated by: learning about problems other users are experiencing; an enhanced feeling of being part of a community; personal benefits of learning through teaching; an enhanced likelihood of receiving help in the future; or a sense of obligation from having received help from others in the past [20, 24].

As social Q&A sites, *e.g.*, the StackExchange network, are becoming more popular, instances of what the economic literature calls "signalling incentives" start to better explain what motivates knowledge providers to help others [21], in addition to the previous reasons. Career concerns (*e.g.*, online activities in social Q&A sites are more visible to employers and recruiters, who may use them to find qualified people [8]), or a desire for peer recognition (*e.g.*, reputation building) are among the listed motives [20,21,24]. However, such signalling incentives may not be equally applicable to all knowledge providers. For example, we can expect that more knowledgeable providers would draw greater benefit from signalling, and thus signal more. Therefore, to obtain a more fine-grained understanding of the transition from traditional information venues to social Q&A, it is important to distinguish between different groups of participants (*e.g.*, developers, with likely more knowledge about the software, versus non-developers).

Another dimension of social Q&A participation incentives arises from *gamification*, *i.e.*, using elements of game design in non-game contexts [10, 11]. Gamification has been widely used in online platforms in recent years, where it was shown to motivate users to contribute more [1,2,10,42], and there are many psychological theories that can explain why [40]. The blueprint[8] that all current implementations of gamification (including all StackExchange sites) follow as-

[8] `http://codingconduct.cc/Meaningful-Play`

sumes stimulating users to perform a desirable activity by awarding them points. Specifically, users (i) are rewarded with points to encourage the desired behaviour (and may be subtracted points to sanction undesired behaviour); (ii) are awarded badges after collecting sufficiently many points or when performing certain activities; and (iii) have their progress tracked and their achievements displayed publicly in a leaderboard, to create competition between them. On StackExchange sites, the goal of "the game" is to have participants teaching and learning from each other[9] by asking relevant questions and providing helpful, well-documented and clear answers. Users receive reputation points when fellow users vote up their questions and answers. In turn, reputation points or directly generating *good* content translate to badges (*e.g.*, "Famous Question", if a question received 10,000 views, or "Legendary", if the user earned 200 reputation points daily 150 times). Top performing users in terms of their reputation count (either at week, month, quarter, year, or overall level) are displayed in public leaderboards. Exceeding various reputation thresholds grants users additional privileges on StackExchange sites, including at the higher end, the privilege to help moderate the site.

Therefore, coming back to our study of mailing lists, it seems that the different incentives, inherent to social Q&A sites, have the potential to "disrupt" mailing list activity and catalyze a transition to social Q&A, for knowledge seekers and knowledge providers alike. Seekers may turn to StackExchange expecting faster answers and a wider expert audience. Providers may transition to it to satisfy a rising demand for information, or pursuing signaling incentives and recognition. The latter is, for instance, supported by "alpha-male" behaviours self-reported by Apache help forum knowledge providers interviewed by Lakhani and von Hippel [20]: those wanting to be known as "the" expert in a particular aspect of Apache "would strive to answer all questions associated with their area" and seek to "drive out all other information providers from [that] chosen field of expertise". In the presence of gamification, such as on StackExchange sites, such behaviours may be even further accentuated. For example, StackOverflow is said to suffer from the *fastest gun in the West problem*:[10] to maximise their chances of collecting up-votes from their peers, participants would race to answer questions as quickly as possible, rather than as correctly or as exhaustively as possible. The said problem is based on the anecdotal belief that given two answers of comparable quality, it is the earliest that would typically receive the most votes, or that one might refrain from answering altogether if someone else already offered a similar solution.

Research Questions

In this paper, our goal is to analyze a mature and vibrant community where knowledge-transfer is transitioning from an older, mailing list modality to a new, social Q&A modality, and understand some of the effects of that transition. We chose the R software community.

[9] http://goo.gl/4kPzBP

[10] http://meta.stackoverflow.com/q/9731/182512

Research Question 1: Can we find evidence in the R community for a decreasing popularity of the mailing list and an increasing popularity of StackExchange?

RQ1 Discussion. R [28], a popular data analysis software, makes for an interesting case study since there have been initiatives by members of its community to promote a transition to StackExchange.[11] However, to quantitatively assess this phenomenon, we first need to construct a historical data set combining mailing list and StackExchange activity for R, , and identify participant overlap (if any). Given such a data set, we can then evaluate activity in the two communities, and look for transition phenomena. To the best of our knowledge, we are the only ones creating and analysing such overlapping data sets between StackExchange and open source communities [38].

In addition to the potentially different motivations driving participants, other factors may influence these phenomena. The R mailing lists have been around since 1997, while StackOverflow, the first of the StackExchange sites, was only launched in 2008. The relative maturity of the former compared to the latter, as well as the direct contact with R developers, may mean that mailing lists are seen as well-established "educational institutions" by R users, *i.e.*, the default go-to place for requesting R support. Moreover, not all mailing list discussions are equally as well suited for StackExchange, hence not all are at risk of being subsumed. For example, StackExchange discourages questions that are too broad or primarily opinion-based.[12] Such questions may therefore only be suitable for the mailing lists. In other words, despite addressing some mailing lists limitations (*e.g.*, pertaining to user interface), StackExchange may not be a complete substitute.

On the other hand, R users asking questions on StackOverflow may be given answers referring to earlier posts from `r-help`.[13] This suggests that mailing lists may not be as well indexed by search engines (hence solutions posted there may not appear as high up in the results list), or that the mindset of users (asking directly on a StackExchange site) is altogether changing in disfavour of the mailing lists. In addition to the factors above, the two communities also show signs of symbiosis. For example, StackOverflow discussions are being followed by R developers and bug reporters, who use the feedback received from this community to stir up development discussions on `r-devel`,[14] the other main mailing list for R, dedicated to developers. The varying and synergistic activities of developers and knowledge-providers on mailing lists and Q&A sites suggest that some developers and knowledge-providers in R are splitting their time between the two, while others, perhaps, are not. Thus, R, a project with a knowledge-exchange in transition, includes some participants who stay with the old, some who are in transition, and some who go entirely with the new. It would

[11] http://goo.gl/aeoBJA, http://goo.gl/OtzZT5

[12] http://meta.stackoverflow.com/a/10582

[13] http://stackoverflow.com/a/1996404
http://stackoverflow.com/a/4947528

[14] http://goo.gl/6UAUgr
http://stackoverflow.com/a/1321491

be important to understand if this transition is in fact taking place, and to what degree. Community processes for answering user questions on-line are a vital component of the industry, and it is important to understand and promote this important process.

> **Research Question 2:** How do the contributors active both on the mailing list and on StackExchange differ from those focused on a single community (either of the two) in terms of their activity levels?

RQ2 Discussion. The different socio-technical incentives inherent in social Q&A (*e.g.*, related to user interface, potentially wider audiences, and gamification) may result in StackExchange becoming more attractive, *i.e.*, taking over (part of) the mailing list activity, and engaging (part of) the mailing list community. Previous studies (*e.g.*, [8]) argue that one's activity in social media can be used as a signal of qualifications and, similarly, one's reputation in social media as a signal of references from peers. Are there mailing list participants active also on StackExchange sites? Who are they and how do they differ (in terms of their activity levels) from other mailing list or StackExchange participants, who choose to focus solely on the mailing list or solely on StackExchange? Are developers more likely to be active both on `r-help` and on StackExchange (*i.e.*, to "signal" more) than non-developers?

> **Research Question 3:** For the contributors active both on the mailing list and on StackExchange, can we find differences in their behaviour in one community versus the other?

RQ3 Discussion. If StackExchange sites are attracting mailing list participants (*e.g.*, knowledge seekers looking for faster answers or a wider expert audience, or knowledge providers looking to satisfy a demand for knowledge, "signal" their "fitness" as experts, or engage in the game for reputation and badges), we should observe a decreasing trend in the mailing list activity concomitant with an increasing trend in StackExchange activity for those participants (especially knowledge providers) active on both. Moreover, gamification, one of the key differences between `r-help` and StackExchange sites, might influence `r-help` participants active on StackExchange to adopt a different behaviour on StackExchange [1, 2, 10, 42]. For example, in order to "survive" in the game for reputation and badges, they might have to provide faster answers than they do on the mailing list. Can we observe such a trend? Do participants active both on `r-help` and on StackExchange behave differently when on the mailing list than when on StackExchange sites? If answers are indeed quicker, and the speed is somewhat attributable to "game-playing", this might confirm that gamification is a useful adjunct in the design of Q&A fora.

RELATED WORK

Our work touches upon different fields of research. First, mail archives have been studied since as early as 2006 (see recent review by Squire [33]), *e.g.*, to understand knowledge sharing [32] and how users of the software get help [31]. Bettenburg *et al.* [5] identified challenges that arise when using off-the-shelf techniques for processing mailing list data. Singh *et al.* [31] and Guzzi *et al.* [16] qualitatively studied the types of discussions that occur in mailing lists, the types of questions asked and the types of responses that are given. For example, Singh *et al.* [31] reported that about 90% of the questions being asked in the online support forums of a number of open source projects are related to problem solving and information seeking, while the remaining can be classified as social discussions or feature requests.

Second, StackExchange (in particular StackOverflow) is receiving increasing attention from researchers, as witnessed by the growing number of papers using StackExchange data published each year,[15] as well as the yearly mining challenge of the International Working Conference on Mining Software Repositories (MSR) choosing StackOverflow as its topic in 2013. For example, Anderson *et al.* [1] investigated the dynamics of the StackOverflow community and found, *e.g.*, significant assortativity in the reputations of co-answerers, or relationships between reputation and answer speed. Vasilescu *et al.* [36, 37] studied the representativeness and activity of genders on StackOverflow compared to traditional mailing lists, and found StackOverflow to be a relatively "unhealthy" community, in which women disengage sooner although their activity levels are comparable to men's. Also related to the current topic is our previous study [38] of the interplay between StackOverflow Q&A activities and the development process, reflected by code changes committed to GitHub by developers also active on StackOverflow. There we found that the more prolific StackOverflow experts (*i.e.*, those providing the most answers) are also very active GitHub committers, and that in general participating in StackOverflow catalyses committing to GitHub.

Third, numerous studies focused on the motivations of participants in knowledge-sharing (online) communities. For example, Lin [22] analysed the effects of extrinsic and intrinsic factors in explaining the knowledge sharing intentions of a number of employees from fifty large organisations in Taiwan. Hendriks [17] proposed a theoretical model linking the variables and motivators involved with sharing knowledge using ICT. Sowe *et al.* [32] discussed the altruistic sharing of knowledge between participants in the Debian mailing lists. Deterding *et al.* [10] presented a number of views on how gamification impacts participation of users in online communities. Zhuolun *et al.* [42] studied the value of using badges in StackOverflow, and found that this reward system helps cultivate users' loyalty to the community.

Finally, R has been the object of two recent studies [13, 41]. German *et al.* [13] found differences in the growth rate as well as the way in which active contributors are attracted between user-contributed packages and core packages. Voulgaropoulou *et al.* [41] analysed the quality of 508 R packages and found, *e.g.*, that changes in social attributes such as the number of developers do not influence the code quality.

Source	Period	#messages #threads	#participants #unique	Multiple aliases
r-help	4/1997–	344,854	33,338	≃15%
	3/2013	97,125	28,096	
StackOverflow [r]	9/2008–	67,248	10,534	≃2%
	3/2013	24,957	10,284	
CrossValidated [r]	7/2010–	7,351	2,312	≃1%
	3/2013	3,208	2,285	

Table 1. Basic statistics for the two data sources.

METHODOLOGY

As case for our case study we selected R [28], a popular data analysis software, for several reasons: First, R is a typical example of an open-source software ecosystem, comprising a (relatively) closed *core of developers providing the basic functionality* and coordinating new releases, various developers *contributing patches and bug fixes*, numerous developers *contributing packages* (plugins) that extend the functionality beyond that provided with each release, and a *plethora of users*. Other examples of such ecosystems include the Eclipse IDE and its third-party plugins, or the Python/Ruby/LaTeX programming languages and their various contributed packages/gems. Second, R has been evolving for almost 20 years, and its entire history of mailing list communication is archived and publicly available. Third, R has been the recent subject of an extensive study of its evolution [13]. Fourth, R promises to provide broader relevance outside the software developers' community, as many users and contributors are data analysts from different domains such as economics or biology, with none or limited software engineering experience.

Data extraction. We integrated data from two different sources: the community behind the main R user support mailing list (r-help), and the StackExchange R subcommunities behind StackOverflow and CrossValidated, the StackExchange websites containing the most R-related ([r]-tagged) questions and answers.

r-help is the principal mailing list for user support in R. It hosts discussions about problems and solutions using R, as well as announcements about the availability of new functionality and documentation. Moreover, it mirrors announcements from r-packages on new or enhanced contributed packages, or major developments from r-announce. The other major mailing list for R is r-devel, targeting developers, testers and bug reporters, with topics that are considered "too technical" for r-help's audience. The r-help archives can be downloaded in standard mbox format.[16] We wrote Python scripts to automatically download, extract, and parse the archives, which date as far back as April 1997.

For each r-help participant we recorded their role: (i) *core developer*, *i.e.*, the 22 developers with "write access to the R source" since its inception;[17] (ii) *peripheral developer*, *i.e.*, the 43 developers who "contributed by donating code, bug fixes and documentation";[17] (iii) *package author/maintainer*, *i.e.*, the 2617 developers maintaining or having authored packages on CRAN, the largest R package repository;[18] (iv) *user*, *i.e.*, those not fitting in any of the previous categories.

StackExchange is a network of Q&A sites started in 2008, now comprising more than 100 sites on different topics, such as English language, video gaming, photography, or parenting. All have similar look and feel, and function by the same principles: participants can ask and answer questions; questions are organised by tags (*e.g.*, [r] for R-related questions) and voted upon by members of the community, with votes translating to reputation points and badges; questions and answers can be edited and improved by other members with sufficiently high reputation; each participant has a dedicated profile page, combining the representation of oneself (self-determined, *e.g.*, name, location, or personal website) with activity-related information (automatically provided, *e.g.*, a list of the answers provided, the total reputation count, or a list of badges); users can subscribe to tags and receive email notifications when new questions are being asked (less practical for high-volume tags such as [c#], [java], or even [r], the latter receiving around 500 questions per week at the time of writing), or simply browse the website.

StackExchange releases quarterly data dumps in XML format from all the websites under its umbrella. We explored the one dated March 2013,[19] using Python scripts to parse the XML archives. We restricted our analysis to StackOverflow (a generic programming Q&A site, and the first and largest StackExchange member) and CrossValidated (dedicated to statistics). This left out a number of other StackExchange websites where questions tagged [r] occasionally (*i.e.*, very infrequently) pop up, such as TeX or GIS, as well as all the activity on StackOverflow and CrossValidated which did not occur within the [r] tag. The oldest StackOverflow questions tagged [r] are dated September 2008. CrossValidated hosts [r]-tagged questions dating as far back as July 2010. The [r] activity on both websites mainly targets R users as opposed to R developers (hence the comparison to r-help is sound), although R developers may also follow it, as discussed in the introduction.

The organization of StackOverflow and CrossValidated in terms of questions, answers, tags, badges and reputation points is the same, with the only difference being the target audience, programmers in the former vs. data analysts in the latter. Therefore, we expect that differences between StackOverflow and CrossValidated should not affect data collection, analysis and interpretation. Table 1 lists some basic statistics about the data sources.

Identity merging. One of the biggest challenges when mining software repositories is *identity merging* [6, 14, 19, 39]. Both within a single repository (*e.g.*, mailing lists), as well as across repositories (*e.g.*, mailing lists and StackExchange), the same person may use different aliases, *i.e.*, different (*name, email*) tuples. For instance, *John Smith* may go by (*John Smith, johnsmith@gmail.com*), (*John, john@smith.com*), etc. The extent of the problem is unpredictable, *e.g.*, one of the Gnome developers reportedly used 168 different aliases in the source code repository [39].

[15] http://meta.stackoverflow.com/q/134495

[16] http://www.r-project.org/mail.html

[17] http://www.r-project.org/contributors.html

[18] Comprehensive R Archive Network http://goo.gl/hHJROZ

[19] http://www.clearbits.net/torrents/2121-mar-2013

A solution is to merge identities, and existing approaches are very diverse. For example, Bird *et al.* [6] try to match full names or email addresses shared by different aliases, and use heuristics to "guess" email prefixes based on combinations of name parts (*e.g.*, *jsmith* is likely to belong to *John Smith*). Kouters *et al.* [19] use Latent Semantic Analysis (LSA), a popular information retrieval technique, and report better results in presence of very noisy data. However, all existing approaches are known to produce false positives and false negatives [14]. We followed different approaches for `r-help` and for StackExchange.

StackExchange. The email addresses of the StackExchange participants are not publicly available for privacy reasons, but their MD5 hashed versions are offered instead. Since it is highly unlikely for two different email addresses to share an MD5 hash, we performed identity merging on the StackExchange data if two MD5 hashes coincided. This process resulted in a reduction in the number of participants of approximately 2% on StackOverflow and 1% on CrossValidated.

We decided to limit the identity merging *solely* to the case above due to the following reasons. First of all, on StackExchange as opposed to the mailing lists users have *accounts* and can log in using a password. This suggests that the incidence of multiple aliases (*i.e.*, accounts) by the same person should be much lower than on the mailing lists, where one is more likely to send an email, *e.g.*, from the account which is most at hand at any given time. While it is common for people to own multiple email accounts (*e.g.*, employer-related, open-source-related, private, etc.), we believe it to be less common for the same person to own multiple accounts on the same StackExchange site (at least because one cannot integrate the activity and reputation points earned using each). Moreover, since the email addresses are not publicly available, identity merging would rely solely on names, increasing the likelihood of false positives.

Mailing list. For `r-help` we performed a fixed-point computation: for each email address, we (i) collected all other email addresses having the same prefix (*e.g.*, for *john.smith@gmail.com* we collected *john.smith@hotmail.com*), as long as the prefix did not consist of a single word (*i.e.*, contained either a dot or an underscore character); and (ii) collected all the different names associated with it (*e.g.*, *John Smith* and *John* if both used the email address *john.smith@gmail.com*). Then, for each of these names we collected the different email addresses with which they were associated (*e.g.*, *John Smith* might be associated with both *john.smith@gmail.com* and *johnny@gmail.com*), and repeated the process until a fixed-point was reached. Checks for a minimal length of the email prefixes and names were used to limit the number of false positives. Applying this technique resulted in a reduction in the number of participants of approximately 15%. Among the users with the most aliases were, *e.g.*, an active participant with ten different emails under the same name, or one of the peripheral R developers with two email addresses and nine different variations of his name.

Intersecting the two data sets. A prerequisite for studying migration phenomena from the mailing lists to StackExchange was computing the overlap between the two communities. First, we made use of the fact that email addresses are available for the mailing list participants, as opposed to only MD5 hashes for the StackExchange users. As a result, we merged a mailing list user with one on StackExchange if the MD5 hash computed for any of the former's email addresses (there might have been multiple after identity merging on the mailing list) was identical to the MD5 email hash of the latter. Then, to expand the merges beyond only matching email addresses, we followed a fixed-point approach similar to the one described above, using names and email address prefixes (*e.g.*, *John Smith* from StackExchange with *John Smith* from `r-help`, despite the latter not having an email address that matches the former's MD5 hash). As a side effect, at this step two StackExchange users could have been merged to the same `r-help` individual, hence also among themselves (in addition to the merges using MD5 hashes discussed in the previous paragraph), if their MD5 hashes matched email addresses known to belong to this `r-help` individual.

Using this approach we found that approximately 15% (or 3,894) of the `r-help` unique participants (*i.e.*, after identity merging) have StackOverflow accounts, and 5% (or 1,159) have CrossValidated accounts.[20] These numbers are influenced by the difference in age between the two communities, since `r-help` started in 1997, while StackOverflow only in 2008 and CrossValidated in 2010: part of the overall mailing list population is not active anymore, as observed also for other open source projects [7]. The size of the overlap increases (*e.g.*, to slightly over 20% for StackOverflow) when considering only the `r-help` participants active starting with September 2008. Variations may also occur between different user groups (*e.g.*, knowledge seekers and knowledge providers). For example, the activity of open source contributors typically follows very skewed distributions [39], *i.e.*, most have very few contributions. Hence, it seems more likely that active knowledge providers would participate in StackExchange than one-time knowledge seekers.

Comparing apples and oranges. Our goals include understanding how the amount of activity differs between mailing lists and StackExchange, and whether mailing list participants are migrating to StackExchange. However, *activity* is organised in different ways in the two communities. On StackExchange users ask questions and typically receive several answers from other members of the community. Answerers "compete" with each other at offering good answers, as reflected by the up votes received from other members of the community. If a question was ill-formulated, *e.g.*, as signalled by other users through comments, the original poster has the option of editing and improving it any number of times. Similarly, if the answers were incomplete, the answerers or any other members of the community have the option of updating them. Questions may remain unanswered. Acknowledgements for answers that solve the original problem can be made by the original poster by *accepting* one answer using a dedicated button, and/or *posting comments* to the others (we do not consider comments). The standard structure

[20] Having a StackExchange account does not guarantee having engaged in Q&A there, hence a smaller fraction will have actually been *active* on StackExchange.

of a Q&A post is therefore *one question followed by zero or more answers, typically offered by different people*.

The equivalent communication structure on mailing lists is represented by *discussion threads*. The first new message with a particular topic (subject) is the one starting the thread (*i.e.*, the question). Similarly to StackExchange, it can remain unanswered, or it can receive responses identifiable as messages sent In-Reply-To the original message, or to other messages within the same discussion thread. However, as opposed to StackExchange, the original poster and the answerers can all appear multiple times within a thread.

Therefore, to avoid comparing apples to oranges when comparing activity on the two communication media, we (i) grouped related email messages into discussion threads, operation known as *threading*; and (ii) discounted multiple answers by the same person, and discounted the original poster as answerer within the same thread. Due to numerous caveats, threading is a non-trivial operation, typically taken lightly in the literature. We used a slightly modified version of Zawinski's algorithm[21] (based on Kuchling's Python implementation),[22] to the best of our knowledge the leading publicly-available threading algorithm. The number of distinct threads obtained for `r-help` is displayed in Table 1. For the StackExchange websites, the number of threads is the same as the number of questions, by definition.

User survey. To better understand the context of our research, related to changes in information seeking and information providing behaviours in the R community, we augmented the quantitative study with a user survey,[23] used to triangulate the quantitative findings. We asked the participants for: background information about their occupation, experience with R and involvement in the development of R (*e.g.*, as core developers, package maintainers, users); their information seeking behaviour (*e.g.*, how frequently they need information, where and how they search for it, how satisfactory what they find is, what their preferred information seeking medium is, and whether anything has changed in their seeking behaviour over the years); and their information providing behaviour (how often and by what means they share information with others, and what motivates them to do so). Survey participants were recruited by posting an ad about the questionnaire on various channels, such as `r-help`, the StackOverflow Meta (the site for meta-discussion of the StackExchange family of Q&A websites, where the maintainers of StackExchange also hang out), the StackOverflow R chat room, Google groups, other French, Russian or German fora, our own LinkedIn, Twitter, Google+ or Facebook profiles, or by contacting R developers directly.

RESULTS

Overall knowledge seeking activity. We start by analysing the activity on `r-help` (Figure 1). After an initial increase in the number of threads started (questions asked) each month, we observe drops in recent activity (since 2010). It is interesting to note that the 2010 inflection point is synchronised with initiatives to promote StackExchange among R users,[24] as illustrated by this excerpt from a blog entry:[25] "*Young experts don't want to have to monitor email all day to be part of the discussion. Their answers belong on a website with a normal content management system, with good search functions and user interactions. Go* [to StackExchange and] *sign up.*" On the other hand, `[r]` activity on StackExchange and in particular StackOverflow (Figure 1) grows at an increasing rate. The fraction of `[r]` questions relative to all questions grows linearly[26] (*i.e.*, `[r]` gains popularity).

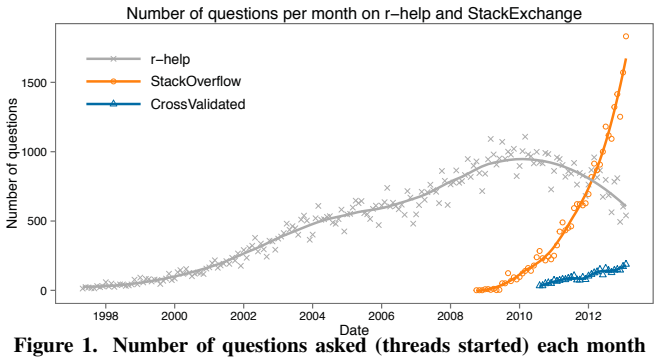

Figure 1. Number of questions asked (threads started) each month on `r-help` and StackExchange (StackOverflow and CrossValidated) in the `[r]`-tag. The trend curves are Loess curves with 0.5 span.

> **RQ1.** Knowledge seeking on `r-help` decreases sharply since around 2010. At the same time, the number of R-related questions asked on StackOverflow and CrossValidated grows at an increasing rate.

Structure of the community. Recall that we have distinguished between different roles within the `r-help` population. Here we wish to understand the knowledge seeking and knowledge providing activities of the different roles. Developers (mostly package authors/maintainers) are responsible for starting only a small fraction of the discussion threads (approximately one tenth in recent years), as depicted in Figure 2, top. Non-developers, the vast majority of question askers, are most responsible for the decreasing activity on `r-help` since 2010. The situation is reversed for thread answerers (Figure 2 bottom). The R core developers and the package authors/maintainers are responsible for most replies to threads started on `r-help`. However, since the decrease in new `r-help` threads started in 2010, the answering activity of developers (package authors/maintainers in particular) has also decreased the most. Both users and developers may be tempted by StackExchange, *e.g.*, one in search of better or faster answers, the other—of recognition.

Activity of contributors engaged in both communities. In this section we focus on those mailing list participants who were also *active* on StackExchange. To control for the difference in age between `r-help` (started in April 1997) and StackExchange (started by StackOverflow in September 2008), we only consider the mailing list participants starting

[21] `http://www.jwz.org/doc/threading.html`
[22] `https://github.com/akuchling/jwzthreading`
[23] `http://goo.gl/mZtz9X`
[24] `http://goo.gl/0i4j1n`, `http://goo.gl/jSZQFZ`
[25] `http://goo.gl/12nJxe`
[26] `http://goo.gl/gWKxbQ`

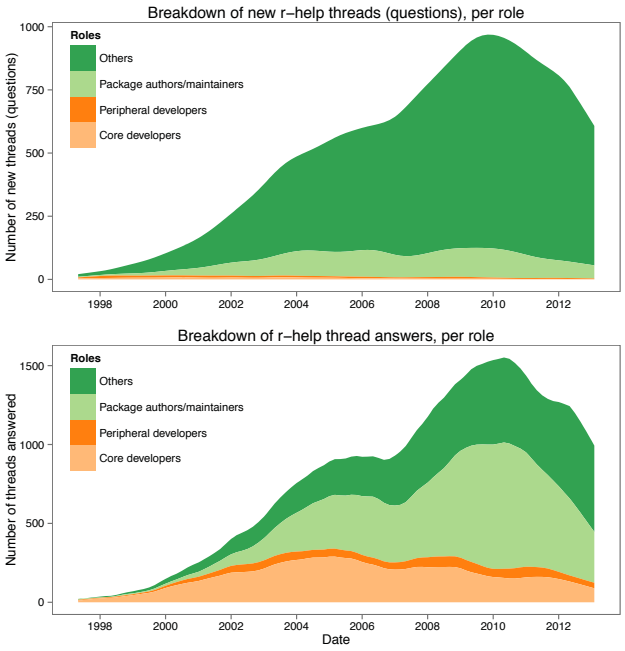

Figure 2. Breakdown of new threads (top) and thread answers (bottom) on `r-help` by initiator role. Only the *trend* component of each time series is displayed, after using a seasonal-trend decomposition procedure based on Loess. The top plot corresponds to Figure 1.

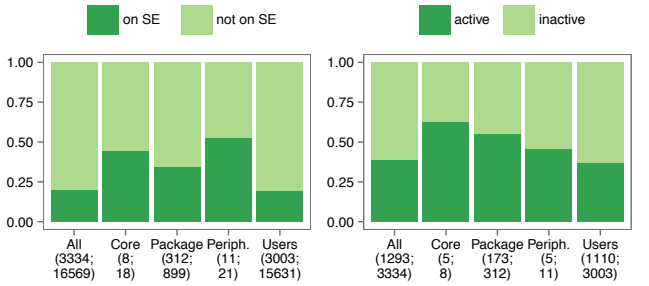

Figure 3. *Left:* Fraction of `r-help` participants (starting with September 2008) with StackExchange accounts, by role. *Right:* Among those with StackExchange accounts, the fraction that asked or answered at least one question. The absolute values are displayed under each label.

Outcome	Devs (D)	Users (U)	χ^2 (p-val)	D-U (CI)	RR (CI)
Have SE accounts	331/938 (35.28%)	3,003/15,631 (19.21%)	141.28 (<2.2E-16)	.16 (.13, .19)	1.83 (1.67, 2.01)
Active on SE	183/331 (55.3%)	1,110/3,003 (36.9%)	41.39 (1.2E-10)	.18 (.12, .24)	1.49 (1.34, 1.66)

Table 2. Results of statistical testing for the Figure 3 cases. D-U (CI): Estimate and 95% confidence intervals for the difference of proportions. RR (CI): Risk ratio by unconditional maximum likelihood estimation and 95% confidence intervals using normal approximation.

with September 2008. While we were able to link approximately 20% (from September 2008, or 15% overall) of the `r-help` participants with StackExchange accounts (either StackOverflow, CrossValidated, or both), we also observed that not all of them have actually engaged in `[r]`-tagged Q&A on the two StackExchange sites. Indeed, only 7.8% (or 1,293 out of 16,569) have asked or answered at least one `[r]`-tagged question on StackOverflow or CrossValidated.

Different population subgroups again exhibit different behaviour (Figure 3): the fraction of participants with StackExchange accounts is much higher among developers (core, peripheral, or package authors/maintainers) than non-developers (left); within those with StackExchange accounts, developers are also more likely to have been active (*i.e.*, to have engaged in Q&A) than non-developers (right). The groups of core and peripheral developers are too small to enable statistically significant comparisons (*e.g.*, we linked only 8 out the 18 core developers active on `r-help` since September 2008 with StackExchange accounts; out of these, only 5 asked or answered at least one question on StackExchange). However, when considering the core, peripheral, and package developers together, the differences in joining and contributing to StackExchange between developers and the users become statistically significant and the effect sizes practically significant (Table 2). For example, developers (183 active out of 331 with StackExchange accounts) have between a 1.34 and 1.66 times higher chance of being active on StackExchange than users (second row).

The more things you do, the more you can do. Actively contributing to both communities requires dividing one's time between them. Do mailing list participants contribute more or less to `r-help` if they are also active on StackExchange? To answer this question, we recorded for each `r-help` participant a flag indicating whether or not he or she was active on StackExchange in the `[r]` tag, the number of threads started and the number of threads answered on the mailing list. To control for differences in age, we again restricted the counts to posts after September 2008. Depending on their `r-help` activity types, three groups of participants emerged (Table 3): (i) those who only started threads; (ii) those who only replied to threads; and (iii) those who both started and replied to threads. For each group we compared the amount of activity (number of threads started for (i) and (iii), and number of threads answered for (ii) and (iii)) between those active and those inactive on StackExchange using the Wilcoxon-Mann-Whitney test (since activity is not normally distributed). In all cases, the results reveal that the mailing list participants who also contributed to StackExchange are more active than those who did not. For example, the median number of threads answered is 4 among those who ask and answer on `r-help` *and* engage on StackExchange, compared to 1 for the others.

Next we compared the StackExchange activity for `r-help` participants also active there to that of the rest of the StackExchange `[r]` community: which were more prolific? Similarly, we distinguished between exclusive askers, exclusive answerers, and users who both asked and answered questions (Table 4). Similar comparisons using Wilcoxon-Mann-Whitney tests show that among the StackExchange `[r]` population, those who also participate in `r-help` are more active. For example, the median number of `[r]` answers for those who ask and answer on StackExchange *and* contribute to the mailing list is 3, as opposed to just 1 for the others.

These results show that the most prominent mailing list participants "signal" the most both on the mailing list and on StackExchange. The group of users who participate in the R mailing lists *and* engage in `[r]` Q&A on StackExchange are the most active contributors relative to the other mail-

Activity on r-help\SE	No account or inactive (N)	WMW comparison (p-val)	Active (A)
Only ask	11,724	Asking: A>N (4.7E-14)	790
Only answer	1,413	Answering: A>N (4.2E-3)	123
Ask, answer	2,139	Asking: A>N (2.83E-8)	380
		Answering: A>N (<2.2E-16)	
Total	15,276		1,293

Table 3. Activity comparisons for three groups of r-help participants also active on StackExchange relative to the other r-help participants (Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction). StackExchange = StackOverflow + CrossValidated.

Activity on SE \r-help	Inactive (N)	WMW comparison (p-val)	Active (A)
Only ask	5,464	Asking: A>N (5.25E-5)	794
Only answer	2,824	Answering: A>N (<2.2E-16)	338
Ask, answer	1,470	Asking: A>N (3.24E-12)	564
		Answering: A>N (<2.2E-16)	
Total	9,758		1,696

Table 4. Activity comparisons for three groups of StackExchange contributors also active on r-help relative to the other StackExchange participants (Wilcoxon-Mann-Whitney tests with Benjamini-Hochberg correction). StackExchange = StackOverflow + CrossValidated.

ing list participants and, similarly, the most active contributors relative to the other StackExchange participants. This result is in line with a recent observation made by Posnett *et al.* [27] in their study of expertise on StackExchange: "*expertise is present from the beginning* [of one's participation]*, and doesn't increase with time spent with the community. [...] In other words, experts join the community as experts, and provide good answers immediately.*" Assuming the number of answers one provides to be a proxy for their expertise, we showed, *e.g.*, that StackExchange experts (*i.e.*, heavy answerers) stem from mailing list experts. Recently we made a similar observation about GitHub developers active on StackOverflow [38]: those who provide the most answers are also those who perform the most commits.

> **RQ2.** The more things you do, the more you can do: Contributors to both communities (more likely developers than non-developers) are more active than those who focus on just one.

Behavioural differences. We compared the question answering activity on r-help for two groups of knowledge providers: those with (denoted "r-help and StackExchange") and those without (denoted "only r-help") StackExchange accounts (Figure 4). For each group we plotted the total number of answers given to r-help threads each month (denoted "On r-help"); for the "r-help and StackExchange" group, we also plotted the number of answers given to StackExchange questions each month (denoted "On StackExchange").

We draw the following observations. The inflection point (mid 2010) in the number of answers given to r-help threads by the "r-help and StackExchange" group coincides with the inflection point in general activity trend on r-help, depicted in Figure 1 top. This should not come as a surprise. In the previous sections we have seen that it is those mailing list participants that are also active on StackExchange that are most active, *i.e.*, the activity drivers or trend setters (both relative to the other mailing list participants as well as to the other StackExchange participants). Therefore, it is natural that they also exhibit a decreasing trend in activity since mid 2010. However, it is interesting to observe that the activity trend of the r-help knowledge providers without StackExchange accounts remains relatively constant (or decreases much slower). Corroborated with the increasing trend in answering activity on StackExchange by mailing list participants active on StackExchange, this suggests that R mailing list experts are migrating to StackExchange.

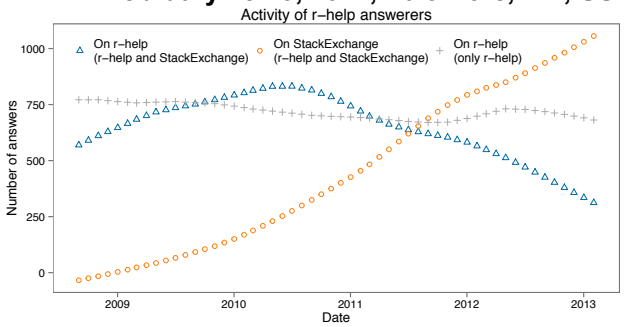

Figure 4. Number of questions answered on r-help (after September 2008) and StackExchange each month: participants exclusive to the mailing list versus those also active on StackExchange. Only the *trend* component of each time series is displayed, after using a seasonal-trend decomposition procedure based on Loess.

However, not all members of the "r-help and StackExchange" group exhibit similar patterns (Figure 5). For example, participant 3513 (topmost subfigure) *is* the norm: singlehandedly responsible for approximately one fifth of the total number of answers, he exhibits a similar decreasing trend in activity as the entire group. In contrast, participant 9440 (second topmost subfigure) has migrated entirely to StackExchange in the second half of 2010. Participant 209 (second bottommost subfigure) has been active on both r-help and StackExchange ever since the beginning, albeit not as intensively on StackExchange; over the years he has become increasingly disinterested in r-help, without becoming more interested in StackExchange. Finally, participant 9859 (bottommost subfigure) is barely active on StackExchage, but is becoming increasingly more active on r-help.

Next we investigated whether there are significant differences between the speed with which r-help participants also active on StackExchange answer questions on r-help versus on StackExchange. If the incentives on StackExchange (*e.g.*, gamification, more attractive user interface) do not influence behaviour, then we should not observe any meaningful differences. If instead users are engaging in the race for reputation and badges, then they might try to answer more questions (we have already seen this in the previous section) as well as answer questions faster on StackExchange than on the mailing list, where they are not "rewarded" for their haste. To test this hypothesis, we compute for each member of the "r-help and StackExchange" group the time intervals between their first answer within a thread and the thread start (for all threads for which they provide answers), on the one hand, and between their first answer to a StackExchange question and the question date (for all questions they answer), on the other hand. Then, us-

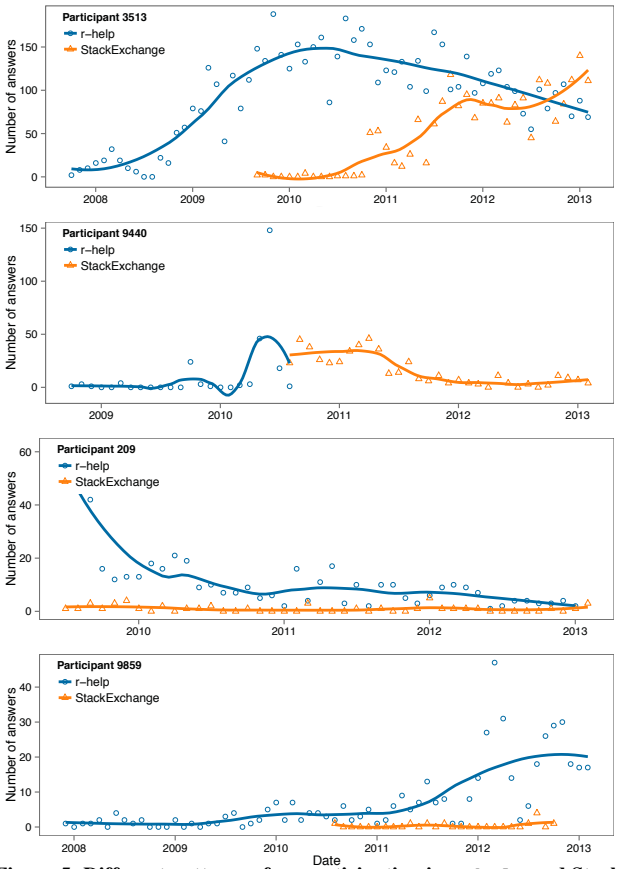

Figure 5. Different patterns of co-participation in `r-help` and StackExchange for knowledge providers (Loess curves with 0.5 span).

ing a Wilcoxon-Mann-Whitney test we compare two large groups of `r-help` and StackExchange time deltas obtained by concatenating the intervals for all "`r-help` and StackExchange" members (Figure 6).

Our results show that the StackExchange answers are significantly faster ($p < 2.2\text{E-}16$), with a median of 47 minutes versus 3 hours on `r-help`. This confirms that `r-help` participants behave differently when on StackExchange, where they are rewarded for their efforts more. To further put these results into context, note that on mailing lists a knowledge provider is *passively* (automatically) provided with opportunities to answer questions (by receiving emails with new questions directly, or in the form of a digest); on StackExchange, although subscribing to tags (to receive notifications of new questions) is possible, knowledge providers will frequently *actively* pursue new questions being asked by browsing the site,[27] and will rush to answer them.

> **RQ3.** Knowledge providers active both on `r-help` and on StackExchange (*i.e.*, the mailing list experts) are migrating to StackExchange, where they answer questions significantly faster than on the mailing list.

Triangulation

In this subsection we compare the quantitative findings with the user survey results. We received 115 responses. One

[27] `http://meta.stackoverflow.com/q/40927/182512`

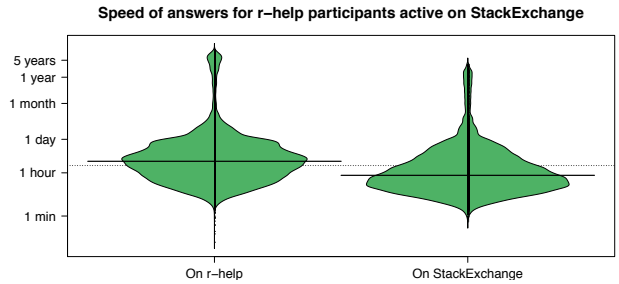

Figure 6. Answer speed for `r-help` users active on StackExchange. Beans: corresponding density shapes. Longer horizontal lines: medians per group. Dashed line: median over all groups.

respondent participated twice, and two respondents did not consent to participate in the survey, leaving 112 valid responses, mostly from academics (32% of the respondents), statisticians (35%), students (14%), and software engineers (6%). Most respondents described themselves as R users (51%), or authors or maintainers of R packages (35%). We also received two responses from the core developers. We stress that the survey was not intended for quantitative analysis; rather, it served to triangulate the earlier findings.

Our first empirical finding was the evidence of a transition in seeking support from the mailing list to StackExchange: while activity on `r-help` decreases sharply since around 2010, the number of R-related questions asked on StackExchange grows at an increasing rate. In order not to impose our perception of transition on the survey participants, we opted for an open question, and asked whether the participants have experienced changes in their information seeking behaviour over the years. On the one hand, we observed that all survey participants reporting changes related to the mailing lists indicated disengagement, *e.g.*, "*Google is getting better at finding answers related to R so I use it more. I rely less on going directly to mailing lists now*" (user and documentation contributor, statistician/data analyst, 6 years of experience), or "*`r-help` used to be very helpful. But as the number of posts has gone up, I find that reading it is not as useful as it had been*" (package maintainer, academic, 8 years of experience). On the other hand, when the survey participants reported changes related to StackExchange sites, they are much more positive, *e.g.*, "*StackExchange has become more prevalent and more useful in the last two years*" (user, academic, 5 years of experience), or "*I started using RSeek.org, but currently I prefer stackoverflow.com, whose question database is increasing*" (user, statistician/data analyst, 6 years of experience). Still, these observations should be placed in the right context: more than half of the respondents are satisfied with the quality of the answers found in the mailing list archives, and more than a third will occasionally share their knowledge through mailing list discussions.

Our next empirical finding was that developers are more likely to be active on StackExchange than non-developers. To confirm it, we asked the survey participants what motivates them to contribute their knowledge. While numerous reasons such as reciprocity have been named both by the users and by the developers, developers are the only ones that mentioned StackExchange and gamification explicitly: "*In case of StackExchange, the reputation ratings are a nice little*

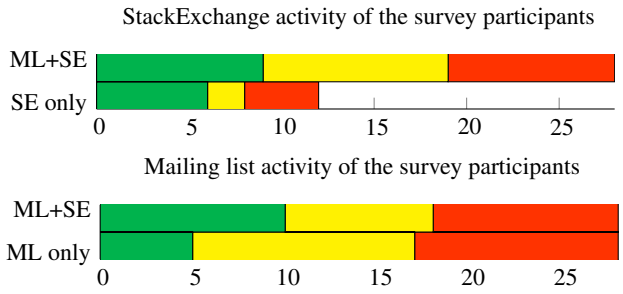

Figure 7. Number of survey participants frequently (green), occasionally (yellow) and rarely (red) sharing knowledge about R on StackExchange (SE) and mailing lists (ML).

incentive," (academic, package maintainer, 6 years of experience), or "[I am motivated by] *peer recognition/gamification within StackOverflow*" (academic, contributes to documentation, submits bugs, 5 years of experience). Moreover, developers put more stress on being motivated by the desire to enhance one's own reputation ("*wanting to be seen as a relative expert in some sub-domain*"; academic, contributes to documentation, submits bugs, 5 years of experience) and on evangelisation ("*I like to promote R because I think it's a great tool to learn about the principles of statistics and programming*"; academic, submits bugs, 4 years of experience). We believe that StackExchange is a better platform for these goals than mailing lists, since StackExchange quantifies one's reputation and makes it visible to everyone (see previous discussion of the gamification blueprint), as opposed to the mailing lists lacking commonly agreed upon and readily accessible representations of reputation. Moreover, StackExchange is a multi-topic community, hence better suited to evangelisation than "preaching to the choir" on a mailing list.

The empirical study also showed that contributors active both on the mailing list and on StackExchange are more active than those focused on only one of the two. To triangulate this finding, we asked the survey participants how often they share information about R by replying to threads on R mailing lists, and by answering questions on StackExchange. Figure 7 shows that survey participants active in both communities are more actively participating in the mailing list discussions than those active only on the mailing lists, *i.e.*, it supports the findings of the empirical study. A similar observation can be made for the StackExchange activity, although disparity between the numbers of survey respondents active only on StackExchange (12) and active in both communities (28) hinders interpretation in this case.

Finally, in the empirical study we observed that knowledge providers active both on `r-help` and on StackExchange (*i.e.*, the mailing list experts) are migrating to StackExchange and are answering questions faster there than on the mailing list. To confirm, we revisited the survey questions about changes in information seeking behaviour and motivations for sharing knowledge. Half of the survey participants reporting changes related to the mailing lists explicitly indicated their transition to StackExchange websites, *e.g.*, due to an easier user interface, a friendlier community, or the sites being better indexed by the search engines. In addition, although not explicitly commenting on the speed with which they provide answers, respondents who contribute their knowledge to StackExchange acknowledged gamification (*i.e.*, reputation building) as important, *e.g.*, "[I am motivated to answer questions on StackExchange because] *it's a game, which also serves a good purpose*" (academic, package maintainer, contributes to documentation, submits bugs, 6 years of experience). Not everyone is attracted by the StackExchange incentives, though. R users might still prefer to offer support via the mailing lists, *e.g.*, "*mailing lists are very nice to read and to reply to, due to their text-only policy*" (academic, package maintainer, 11 years of experience).

IMPLICATIONS

In this section we review the implications of our study for Q&A sites design, knowledge communities and future research.

Implications of our work for *Q&A site designers* are twofold. First of all, we have observed that the movement to social, gamified Q&A is correlated with an increase in the engagement of knowledge providers, and the rapidity of response. This finding suggests that *Q&A site designers* should consider gamification elements to increase engagement of the participants, and, indirectly, popularity of their sites. Second, we have seen that users and developers exhibit different behaviour (*e.g.*, developers are more likely to be active on StackExchange). This means that the Q&A site designers should cater for different groups of community members (*e.g.*, developers and users) having different needs and expectations, by providing different knowledge sharing channels involving different participation and reward mechanisms. This is also why we do not believe that `r-help` will eventually die off, as StackOverflow and `r-help` users have explained: "*If you have a problem and you are* completely stuck*, ask a question on the mailing list*"[28]and "*Although many people there [StackOverflow] gave very detailed answers, I have the feeling that there is much more wisdom on the subject that is still only available in this mailing list*".[29]

For *knowledge communities* our findings suggest that a move to gamified social Q&A can be a beneficial strategy when searching for a better visibility and contributors' activity. We believe that gamification may be of particular interest in closed environments, such as commercial corporations, where knowledge is proprietary, and the set of potentially knowledge providers is closed. Furthermore, the aforementioned distinction between different groups of community members, and different knowledge sharing channels required, suggests that knowledge communities might prefer to combine different knowledge channels.

Finally, our findings provide new insights for *researchers* of collaborative software engineering. While there is a significant body of work on why gamification features should positively influence participation in knowledge communities, facilitate user contributions and foster productivity, our study adds concrete evidence that gamification features or community design affect productivity of the contributors.

[28] `http://stackoverflow.com/a/3382477`
[29] `http://goo.gl/CoJIyp`

THREATS TO VALIDITY

Despite our detailed efforts in experimental design, data gathering and data analysis, we do note several threats to the validity of our methodology and conclusions. The core step in the data gathering, extracting information from the StackExchange data dump and the mail archives, is subject to threats to validity akin to those identified for digital trace data [15, 18]. Following the classification of Howison [18], *system and practice issues* in our work may be related to communication through other mailing lists than `r-help` and other StackExchange sites than StackOverflow and CrossValidated. Indeed, advanced R packages such as `lme4` have separate mailing lists,[30] and experts working with these packages might prefer not to migrate to StackExchange. However, `r-help`, StackOverflow and CrossValidated are the biggest platforms of their kind. We also knowingly omitted other venues where knowledge exchange happens, *e.g.*, Google+ groups or individual blogs. Our results are very strong, and give us confidence that StackExchange is a good representative of the missing Q&A communities.

Moreover, the data extracted is subject to potential *reliability* threats arising from missing messages from mail archives, and questions and answers being deleted from StackExchange. Indeed, as indicated by one of the survey participants, StackExchange sites encourage the participants "*to ask well-formed questions, leading to well-formed answers*", while ill-formed questions are being removed. Question or answer removal might lead to underestimating the activity of the StackExchange contributors, as well as the overlap between StackExchange and `r-help`. Similarly, removal of StackExchange accounts might lead to underestimation of the overlap. Another reliability threat is related to multiple system representations of a single individual, which again might incur an underestimate of the overlap between StackExchange and `r-help`. We have explicitly addressed this threat when discussing identity merging in the methodology section: while no data gathering is perfect, our efforts are particularly detailed, and have been shown elsewhere to work satisfactorily. Finally, reliability might be threatened by noise in the data, such as spam messages stored in the mail archives.

The next group of threats to validity refers to *representation of the data extracted in the model* of one question followed by multiple answers. While the distinction and relation between questions and answers on StackExchange is explicit in the data organisation, the natural counterpart in the mailing lists is the discussion thread (see "Comparing apples and oranges" in the methodology section). Recognition of the discussion thread starter as the asker, and other thread participants as answerers can, however, be threatened by the thread starter posting an announcement rather than a question, as well as by thread participants trying to clarify the intention of the thread starter rather than answering her questions. We have performed an informal evaluation of the discussion threads in `r-help` and observed that the lion's share of threads adhere to the "question and answers" model similar to StackExchange.

[30] `http://goo.gl/ofUT3L`

Temporal aggregation threats arise when aggregating events that occur at different points in time (cf. Figures 1 and 4). Indeed, inappropriate choice of the time granularity might have made our work subject to ecological fallacy [26]. However, our choice for month as the basic time unit stems from the way traditional mail archives are presented (per month) as well as the literature.

CONCLUSIONS

Using an *aligned, longitudinal* data set which connects the same people over time in different knowledge-transfer venues, we have been able to examine the phenomena associated with the transition to StackExchange from the mailing lists.

Future research directions would involve designing and setting up follow-up interviews with R developers and users to address the causality of user migration between different mailing lists and StackExchange: which incentives such as more modern user interface, potentially wider audiences, and gamification are perceived as most important for different subgroups within the R community? Understanding causality could result in quantitative predictive models of user behaviour in the presence of various incentives.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their numerous remarks that helped us improve the paper significantly. Vasilescu gratefully acknowledges support from the Dutch Science Foundation (NWO) through the project NWO 600.065.120.10N235. Filkov gratefully acknowledges support from the US Air Force Office of Scientific Research through the project FA955-11-1-0246. Devanbu gratefully acknowledges support from the NSF, grants CISE SHF MEDIUM #0964703 and CISE SHF EAGER #1247088. Part of this research has been carried out during Vasilescu's visit at the University of California, Davis, USA.

REFERENCES

1. Anderson, A., Huttenlocher, D. P., Kleinberg, J. M., and Leskovec, J. Discovering value from community activity on focused question answering sites: a case study of Stack Overflow. In *Proc. KDD*, ACM (2012), 850–858.
2. Antin, J., and Churchill, E. F. Badges in Social Media: A Social Psychological Perspective. In *Proc. CHI*, ACM (2011).
3. Bagozzi, R. P., and Dholakia, U. M. Open source software user communities: A study of participation in Linux user groups. *Management Science 52*, 7 (2006), 1099–1115.
4. Begel, A., Bosch, J., and Storey, M.-A. Social networking meets software development: Perspectives from GitHub, MSDN, Stack Exchange, and TopCoder. *IEEE Software 30*, 1 (2013), 52–66.
5. Bettenburg, N., Shihab, E., and Hassan, A. E. An empirical study on the risks of using off-the-shelf techniques for processing mailing list data. In *Proc. ICSM*, IEEE (2009), 539–542.
6. Bird, C., Gourley, A., Devanbu, P. T., Gertz, M., and Swaminathan, A. Mining email social networks. In *Proc. MSR*, ACM (2006), 137–143.
7. Bird, C., Gourley, A., Devanbu, P. T., Swaminathan, A., and Hsu, G. Open borders? Immigration in open source projects. In *Proc. MSR*, IEEE (2007), 6.
8. Capiluppi, A., Serebrenik, A., and Singer, L. Assessing technical candidates on the social web. *IEEE Software 30*, 1 (2013), 45–51.
9. Dabbish, L. A., Stuart, H. C., Tsay, J., and Herbsleb, J. D. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proc. CSCW*, ACM (2012), 1277–1286.

10. Deterding, S. Gamification: designing for motivation. *Interactions 19*, 4 (2012), 14–17.
11. Deterding, S., Sicart, M., Nacke, L., O'Hara, K., and Dixon, D. Gamification: Using game-design elements in non-gaming contexts. In *Proc. CHI*, ACM (2011), 2425–2428.
12. Gentle, A. *Conversation and Community: The Social Web for Documentation*. XML Press, 2009.
13. German, D. M., Adams, B., and Hassan, A. E. The evolution of the R software ecosystem. In *Proc. CSMR*, IEEE (2013).
14. Goeminne, M., and Mens, T. A comparison of identity merge algorithms for software repositories. *Science of Computer Programming* (2011).
15. Goggins, S., Mascaro, C., and Valetto, G. Group informatics: A methodological approach and ontology for understanding socio-technical groups. *JASIST 64*, 3 (2013), 516–539.
16. Guzzi, A., Bacchelli, A., Lanza, M., Pinzger, M., and van Deursen, A. Communication in open source software development mailing lists. In *Proc. MSR*, IEEE (2013), 277–286.
17. Hendriks, P. Why share knowledge? The influence of ICT on the motivation for knowledge sharing. *Knowledge and Process Management 6*, 2 (1999), 91–100.
18. Howison, J., Crowston, K., and Wiggins, A. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems 12* (2011).
19. Kouters, E., Vasilescu, B., Serebrenik, A., and van den Brand, M. G. J. Who's who in Gnome: Using LSA to merge software repository identities. In *Proc. ICSM*, IEEE (2012), 592–595.
20. Lakhani, K. R., and von Hippel, E. How open source software works: "free" user-to-user assistance. *Research Policy 32*, 6 (2003), 923–943.
21. Lee, S., Moisa, N., and Weiss, M. Open source as a signalling device - an economic analysis. Working Paper Series: Finance and Accounting 102, Department of Finance, Goethe University Frankfurt am Main, 2003.
22. Lin, H.-F. Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. *J. Information Science 33*, 2 (2007), 135–149.
23. Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. Design lessons from the fastest Q&A site in the west. In *Proc. CHI*, ACM (2011), 2857–2866.
24. Oram, A. Why do people write free documentation? Results of a survey. `http://www.onlamp.com/lpt/a/7062`, 2007.
25. Parnin, C., Treude, C., Grammel, L., and Storey, M.-A. Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow. Tech. rep., Georgia Institute of Technology, 2012.
26. Posnett, D., Filkov, V., and Devanbu, P. T. Ecological inference in empirical software engineering. In *ASE* (2011), 362–371.
27. Posnett, D., Warburg, E., Devanbu, P., and Filkov, V. Mining Stack Exchange: Expertise is evident from earliest interactions. In *Proc. ASE SocialInformatics*, IEEE (2012), 199–204.
28. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.
29. Shah, C., Oh, S., and Oh, J. S. Research agenda for social Q&A. *Library & Information Science Research 31*, 4 (2009), 205–209.
30. Singer, L., Filho, F., Cleary, B., Treude, C., Storey, M., and Schneider, K. Mutual assessment in the social programmer ecosystem: An empirical investigation of developer profile aggregators. In *Proc. CSCW*, ACM (2013).
31. Singh, V., Twidale, M. B., and Nichols, D. M. Users of open source software - How do they get help? In *Proc. HICSS*, IEEE (2009), 1–10.
32. Sowe, S. K., Stamelos, I., and Angelis, L. Understanding knowledge sharing activities in free/open source software projects: An empirical study. *JSS 81*, 3 (2008), 431–446.
33. Squire, M. How the floss research community uses email archives. *IJOSSP 4*, 1 (2012), 37–59.
34. Storey, M.-A. D., Treude, C., van Deursen, A., and Cheng, L.-T. The impact of social media on software engineering practices and tools. In *Proc. FoSER*, ACM (2010), 359–364.
35. Swisher, J. Open source user assistance: Ensuring that everybody wins. *Open Source Business Resource* (2010).
36. Vasilescu, B., Capiluppi, A., and Serebrenik, A. Gender, representation and online participation: A quantitative study of StackOverflow. In *Proc. ASE SocialInformatics*, IEEE (2012), 332–338.
37. Vasilescu, B., Capiluppi, A., and Serebrenik, A. Gender, representation and online participation: A quantitative study. *Interacting with Computers* (2013), 1–24.
38. Vasilescu, B., Filkov, V., and Serebrenik, A. StackOverflow and GitHub: Associations between software development and crowdsourced knowledge. In *Proc. ASE SocialCom*, IEEE (2013), 188–195.
39. Vasilescu, B., Serebrenik, A., Goeminne, M., and Mens, T. On the variation and specialisation of workload–A case study of the Gnome ecosystem community. *Empirical Software Engineering* (2013), 1–54.
40. Vassileva, J. Motivating participation in social computing applications: a user modeling perspective. *User Modeling and User-Adapted Interaction 22*, 1-2 (2012), 177–201.
41. Voulgaropoulou, S., Spanos, G., and Angelis, L. Analyzing measurements of the R statistical open source software. In *Proc. SEW*, IEEE (2012), 1–10.
42. Zhuolun, L., Huang, K.-W., and Cavusoglu, H. Can we gamify voluntary contributions to online Q&A communities? Quantifying the impact of badges on user engagement. In *Proc. WISE* (2012).