GitHub and Stack Overflow: Analyzing Developer Interests Across Multiple Social Collaborative Platforms

Roy Ka-Wei Lee(✉) and David Lo

School of Information Systems, Singapore Management University, Singapore, Singapore
{roylee.2013,davidlo}@smu.edu.sg

Abstract. Increasingly, software developers are using a wide array of social collaborative platforms for software development and learning. In this work, we examined the similarities in developer's interests within and across GitHub and Stack Overflow. Our study finds that developers share common interests in GitHub and Stack Overflow; on average, 39% of the GitHub repositories and Stack Overflow questions that a developer had participated fall in the common interests. Also, developers do share similar interests with other developers who co-participated activities in the two platforms. In particular, developers who co-commit and co-pull-request same GitHub repositories and co-answer same Stack Overflow questions, share more common interests compare to other developers who co-participate in other platform activities.

Keywords: Social collaborative platforms · Online communities

1 Introduction

Software developers are increasingly adopting social collaborative platforms for software development and making a reputation for themselves. Two of such widely adopted and studied social-collaborative platforms are *GitHub*[1] and *Stack Overflow*[2]. GitHub is a collaborative software development platform that allows code sharing and version control. Developers can participate in various activities in GitHub, for example, developers may *fork* (i.e., create a copy of) repositories of other developers. Stack Overflow is a community-based website for asking and answering questions relating to programming languages, software engineering, and tools. Although the two platforms are used for different purposes, developers can utilize both platforms for software development. For example, a developer who has interests in Java programming language may fork a Java project in GitHub and answer Java programming questions in Stack Overflow.

We broadly define the interests of a developer as the programming related topic domains of GitHub repositories and Stack Overflow questions that he or

[1] https://github.com/.
[2] http://stackoverflow.com/.

G.L. Ciampaglia et al. (Eds.): SocInfo 2017, Part II, LNCS 10540, pp. 245–256, 2017.
DOI: 10.1007/978-3-319-67256-4_19

she has participated. For instance, when a developer answers questions tagged with *javascript*, *jquery*, and *angularjs*, we deduce that the developer is interested in the three technologies. Similarly, when a developer forked repositories in GitHub which description contains keywords such as *javascript* and *ajax*, we could estimate that the developer is interested in the two technologies.

The learning of developers' interests could provide new insights to how developers utilize the two social collaborative platforms for software development. For example, if developers share similar interests in GitHub and Stack Overflow, the two platforms may be used to complement each other for software development. Conversely, if the developers display differences between their interests in GitHub and Stack Overflow, the two platforms may have been used in a disjoint manner. The social and community-based element in GitHub and Stack Overflow also adds on to the dynamics when studying developer's interests; developers may find themselves sharing similar interests with other developers who also co-participated in a common repository or question. Thus, it would be interesting to investigate the interests of developers within and across the two platforms. In particular, we ask the following research questions: Does an individual developer share similar interests in his GitHub and Stack Overflow accounts? (**RQ1**), and does an individual developer share similar interests with other developers who co-participated activities in GitHub and Stack Overflow? (**RQ2**).

Our research in this paper is thus divided into two main parts: In the first part, we propose similarity scores to measure the developer's interests within and across social collaborative platforms. In the second part, we applied the propose measures on large GitHub and Stack Overflow datasets and conduct an empirical study to answer the two research questions listed earlier.

Contributions. This work improves the state-of-the-art of inter-network studies on multiple social collaborative platforms. Key contributions of this work include: Firstly, to the best of our knowledge, it is the first research attempt to study similarity of developer interests across GitHub and Stack Overflow using large datasets. Second, we proposed several scores to measure the similarity in developer interests within and across social collaborative platforms. The proposed similarity scores are also applied in an empirical study to quantify the similarity in developer's interests within and across Stack Overflow and GitHub.

2 Data Preparation

2.1 Dataset

There are two main datasets used in our study; we retrieve activities from October 2013 to March 2015 of about 2.5 million GitHub users and 1 million Stack Overflow users from open-source database dumps [5][3]. As this study intends to investigate developer interests across GitHub and Stack Overflow, we further identify developers who were using both platforms. For this work, we used

[3] https://archive.org/details/stackexchange.

the dataset provided by Badashian et al. [1], where they utilized GitHub users' email addresses and Stack Overflow users' email MD5 hashes to find the intersection between the two datasets. In total, we identify 92,427 developers, which forms our *base developer* set. Subsequently, we retrieved the platform activities participated by the base developers. In total, we have extracted 416,171 *Fork*, 2,168,871 *Watch*, 846,862 *Commit*, 386,578 *Pull-Request*, 277,346 *Ask*, 766,315 *Answer* and 427,093 *Favorite* activities. Our subsequent analysis will be based on this group of activities participated by the base developers.

2.2 Estimating Developer Interests

We estimate developer interests by observing the group of activities they participated in GitHub and Stack Overflow. To estimate developer interests in Stack Overflow, we use the descriptive tags of the questions that they asked, answered and favorited. For example, consider a question q related to mobile programming for Android smartphones which contain the following set of descriptive tags: {*Java, Android*}. If a developer d asked, answered, or favorited that question, we estimate that his interests include *Java* and *Android*. GitHub does not allow users to tag repositories but it allows users to describe their repositories. These descriptions often contain important keywords that can shed light to developer interests. To estimate developer interests from the repositories that a developer had participated, we first collect all descriptive tags that appear in our Stack Overflow dataset. Subsequently, we perform keyword matching between the collected Stack Overflow tags and a GitHub repository description. We consider the matched keywords as the estimated interests. We choose to use Stack Overflow tags to ensure that developer interests across the two platforms can be mapped to the same vocabulary.

We denote the estimated interests of a developer given a repository r that he or she forked, watched, committed or pull-requested in GitHub as $I(r)$. Similarly, we denote the estimated interests of a developer given a question q that he or she asked, answered, or favorited in Stack Overflow as $I(q)$. Since the estimated interests given a repository or a question is the same for all developers participated in it, we also refer to $I(r)$ and $I(q)$ as the interests in r and q. For simplicity, we also refer to them as r's interests and q's interests respectively. Developer d's overall interest in GitHub and Stack Overflow, denoted by $I^{GH}(d)$ and $I^{SO}(d)$, is the union of his/her interests over all the repositories and questions group of activities that d has participated in.

3 Measuring Developer Interests Similarity

3.1 Developer Interests Similarity Across Platforms

One way to measure the similarity in an individual developer's interests across platforms is to take the intersection of his interests in Stack Overflow ($I^{SO}(d)$) and his interests in GitHub ($I^{SO}(d)$). However, this simple measure considers

all interests to have an equal weight. In reality, a developer may ask much more questions related to a particular interest than other interests. Similarly, a developer may fork repositories related to a particular interest than other interests. Thus, a finer way to measure the similarity in developer interests should consider the number of repositories and questions that belong to each interest.

To capture the above mentioned intuition, we propose *cross-platform similarity score*, which is denoted as $Sim^{SO\text{-}GH}(d)$. Given a developer d, we measure d's similarity in interests across Stack Overflow (SO) and GitHub (GH) by computing the *proportion* of d's repositories and questions that fall in d's *common interests* in Stack Overflow and GitHub (i.e., $I^{SO}(d) \bigcap I^{GH}(d)$). By denoting the repositories and questions that are related to d (i.e., d forked, watched, committed, pull-requested, asked, answered, or favorited these repositories or questions) as $d.R$ and $d.Q$, we can mathematically define $Sim^{SO\text{-}GH}(d)$ as follows:

$$CI(d) = I^{SO}(d) \bigcap I^{GH}(d) \tag{1}$$

$$Shared^{Q}(d) = \{q \in d.Q | I(q) \in CI(d)\} \tag{2}$$

$$Shared^{R}(d) = \{r \in d.R | I(r) \in CI(d)\} \tag{3}$$

$$Sim^{SO\text{-}GH}(d) = \frac{|Shared^{R}(d)| + |Shared^{Q}(d)|}{|d.R| + |d.Q|} \tag{4}$$

In Eq. 1, we define the common interests of developer d in both Stack Overflow and GitHub. Equation 2 defines the set of questions that falls into the common interests, while Eq. 3 defines the set of repositories that falls into the common interests. Equation 4 defines $Sim^{SO\text{-}GH}(d)$ as the proportion of repositories and questions of d that falls into the common interests. Please refer to Appendix 1 for an example that illustrate how *cross-platform similarity score* is calculated.

3.2 Developer Interests Similarity Among Co-Participated Developers

To study the similarity of interests among developers who co-participated in GitHub and Stack Overflow activities, we propose *co-participation similarity scores*, each focusing on a platform activity. Given a platform activity and a target developer d, we want to measure the similarity between d and *all other developers* who co-participated in the target activity for *at least one* common GitHub repository or StackOverflow question. For example, considering forking a repository as an activity of interest, we want to find developers who co-fork at least one common GitHub repository with d. Hence, given a developer d, we denote the set of other developers who co-participated in forking at least one common repository or question as $Co^{F}(d)$.

Intuitively, the more repositories or questions of common interests that d share with other developers in $Co^{F}(d)$, the higher the similarities should be. To compute the similarity in interests between d and $Co^{F}(d)$, we measure the average similarity in interests between d and each developer d' in $Co^{F}(d)$; for

each of such pair, we measure their similarity by computing the proportion of d''s forked repositories which share an interest with the interests of d in his/her forked repositories. Mathematically, we define the *co-participation similarity scores* for forking in Eq. 5.

$$Sim^F(d, Co^F(d)) = \frac{\sum_{d' \in Co^F(d)} \frac{|Shared^F(d,d')|}{|d'.RF|}}{|Co^F(d)|} \tag{5}$$

$$Sim^W(d, Co^W(d)) = \frac{\sum_{d' \in Co^W(d)} \frac{|Shared^W(d,d')|}{|d'.RW|}}{|Co^W(d)|} \tag{6}$$

$$Sim^C(d, Co^C(d)) = \frac{\sum_{d' \in Co^C(d)} \frac{|Shared^C(d,d')|}{|d'.RC|}}{|Co^C(d)|} \tag{7}$$

$$Sim^P(d, Co^P(d)) = \frac{\sum_{d' \in Co^P(d)} \frac{|Shared^P(d,d')|}{|d'.RP|}}{|Co^P(d)|} \tag{8}$$

$$Sim^A(d, Co^A(d)) = \frac{\sum_{d' \in Co^A(d)} \frac{|Shared^A(d,d')|}{|d'.QA|}}{|Co^A(d)|} \tag{9}$$

$$Sim^V(d, Co^V(d)) = \frac{\sum_{d' \in Co^V(d)} \frac{|Shared^V(d,d')|}{|d'.QV|}}{|Co^V(d)|} \tag{10}$$

In the above formulas, $d'.RF$ denotes the repositories or questions that d' forked. Furthermore, $Shared^F\ (d, d')$ denotes the set of repositories which are forked by d' and share common interests with d's forked repositories. Mathematically, it is defined as:

$$\{r' \in d'.RF | \left[I(r') \bigcap \bigcup_{r \in d.RF} I(r) \right] \neq \emptyset\}$$

In Eq. 5, we define the average similarity in interests between developer d and other developers who had co-forked at least 1 repository with d. The *co-participation similarity scores* for co-watch ($Sim^W(d, Co^W(d))$), co-commit ($Sim^C(d, Co^C(d))$), co-pull-request ($Sim^P(d, Co^P(d))$), co-answer ($Sim^A(d, Co^A(d))$), and co-favorite ($Sim^V(d, Co^V(d))$) are similarly defined in Eqs. 6 to 10. Please refer to Appendix 2 for an example that illustrate how *co-participation similarity score* is calculated.

4 Empirical Study

In this section, we applied the developer interests similarity measures proposed in the previous section on GitHub and Stack Overflow large datasets. We also attempt to answer the two research questions that we have listed earlier in this empirical study **RQ1** and **RQ2**.

4.1 RQ1: Does an Individual Developer Share Similar Interests in His GitHub and Stack Overflow Account?

Figure 1 shows the distribution of the *cross-platform similarity scores* computed for the base developers. On average, the developers have a similarity score of 0.39. This suggests that on average, 39% of the GitHub repositories and Stack Overflow questions that a developer had participated shared similar interests. Also, close to half (49%) of the developers have scored 0.5 or higher, while 26% of the developers have their similarity scores equal to 0, i.e., the interests of these developers are totally different in GitHub and Stack Overflow. This suggests that although most developers do share high similarity in interests in GitHub and Stack Overflow, however, there are a group of developers who have totally different interest in GitHub and Stack Overflow.

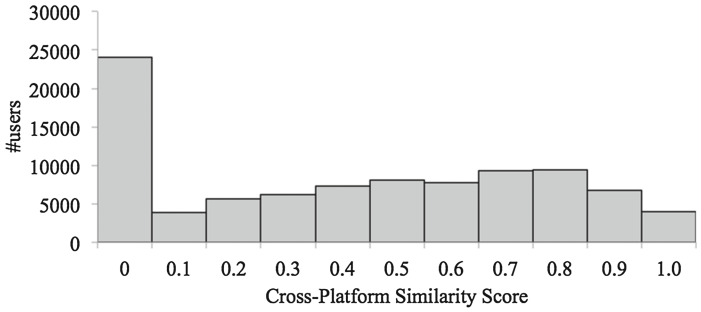

Fig. 1. Distribution of developers' *cross-platform similarity scores* in GitHub and Stack Overow

We further drill down to compare the similarity in developer interests for different types of activity across the two platforms. For example, we measure the similarity in developer's interests by only considering repositories that the developer has forked and questions that the developer has answered. Twelve different combinations capturing different pairs of activities across the two platforms are considered: *Fork-Ask*, *Fork-Answer*, *Fork-Favorite*, *Commit-Ask*, *Commit-Answer*, *Commit-Favorite*, *pull-request-Ask*, *pull-request-Answer*, *pull-request-Favorite*, *Watch-Ask*, *Watch-Answer* and *Watch-Favorite*.

Figure 2 shows the boxplots of *cross-platform similarity scores* for the 12 different activity pairs. The platform activity pairs have average similarity scores between 0.27 to 0.38, slightly lower than the overall average of 0.39. All the platform activity pairs also have significantly higher number of developers with scores of 0. This is as expected since by combining all platform activity pairs we have a larger pool of common interests. Among the 12 activity pairs, *pull-request-Answer* pair has the highest average similarity score. A possible explanation for this observation could be attributed to the nature of the platform activity; *pull-request* and *answer* not only reveal the interests of the developers but also

demand the developers to have a certain expertise on the topics or programming languages of the participated repositories and questions. For example, a developer who is proficient in Java programming language would only *answer* Java programming related questions and submit *pull-request* for Java repositories but he could *watch* other programming language repositories or *favorite* questions from other topics for learning purposes.

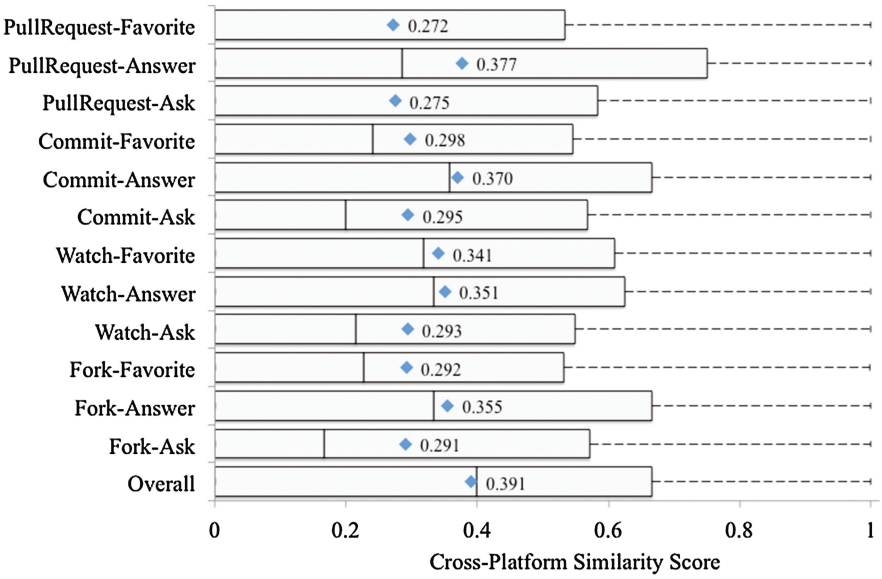

Fig. 2. Boxplots of interest similarity for different activity pairs

4.2 RQ2: Does an Individual Developer Share Similar Interests with Other Developers Who Co-Participated Activities in GitHub and Stack Overflow?

Figure 3 shows the boxplots of *co-participation similarity scores* of the base developers. We observe that an individual developer has average similarity scores between 0.45 to 0.86 with other developers who participated in at least one common platform activity. This means that given two developers who participated in a common platform activity, on average 45–86% of all repositories and questions that they participated in that platform activity shared common interests. Interestingly, we also observed that *commit*, *pull-request* and *answer* have higher average similarity score compare to the rest of the platform activities (0.81, 0.86 and 0.78 respectively). A possible reason for this observation could again be related to the expertise of the developers. We would expect that the expertise of the developers to be more specialized and less diverse than developers' interests, thus resulting in higher similarity scores for developers sharing a common *commit*, *pull-request* and *answer*.

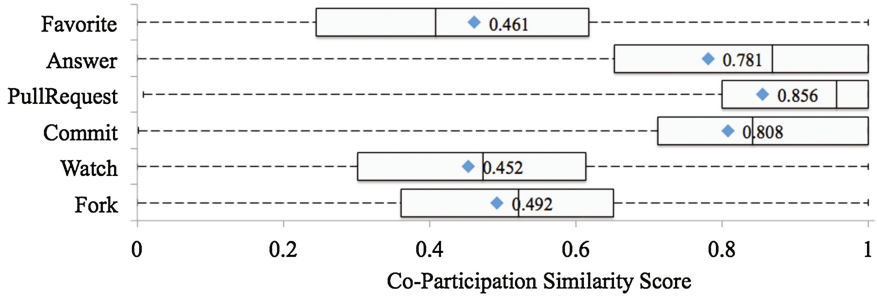

Fig. 3. Boxplots of *Co-Participation Similarity* scores for different activities

4.3 Discussion

Our empirical study has validated that developers do display some similarity in interests in their GitHub and Stack Overflow accounts (**RQ1**) and developers do share common interests with other co-participating developers in the platforms (**RQ2**). Furthermore, we were able to quantify the level of similarity in developer interests across different social collaborative platforms; we found that on average, 39% of the GitHub repositories and Stack Overflow questions that a developer had participated fall in the common interests. The findings in this research could also spark more inter-platforms software engineering research. For instance, when studying the evolution of developer interests, one could take a different perspective and investigate the differences in developer interests in multiple social collaborative platforms over time to observe how developers learn and pick up new interests (e.g., a new programming language).

The findings from our empirical study could be extended to build predictive analytics and recommendation application. As we learned that developers do share interest similarity across platforms (**RQ1**), intuitively we could predict a developer's activities in one platform using his or her interests displayed on another platform. For example, if we learn that a developer answer Java related questions in Stack Overflow, and he displays high similarity in interests across platforms, we can predict that the developer is likely to participate in Java related repositories in GitHub. Likewise in our empirical study, we found that developers do share similar interests with other developers who co-participated activities in GitHub and Stack Overflow (**RQ2**). With this insights, we could predict a developer' activities in a platform using the interests of other developers who had co-participated with him or her in the platform. For example, if we learn that a developer answers a Java related question in Stack Overflow, and we learn that other developers who answered the same questions also display strong interests in Android related questions, we can predict that the developer too, is likely to participate in Android related questions in Stack Overflow. We will look into extending our empirical study predictive analytics and recommendation systems in future works.

5 Related Work

There have been few existing inter-network studies on GitHub and Stack Overflow. These works did deepen our understanding of developer behaviors in the two social-collaborative platforms. Vasilescu et al. performed a study on developers' involvement and productivity in Stack Overflow and GitHub [13]. They found that developers who are more active on GitHub (in terms of GitHub commits), tend to ask and answer more questions on Stack Overflow. Badashian et al. [1] did an empirical study on the correlation between different types of developer activities in the two platforms. Their findings supported the findings of the earlier work by Vasilescu et al., that is: developers who actively contributed to GitHub, also actively answered questions in Stack Overflow. They observed overall weak correlation between the activity metrics of the two networks and concluded that developer activities in one network are not strong predictors for activities on another network. Both the works, however, did not consider intrinsic interests of the developers, although Vasilescu et al. did mention the possibility of extending their work to consider topic interests of the developers.

Stack Overflow and GitHub have also been studied for empirical works on developer interests. For example, there were research works that focused on analyzing topics asked by developers in Stack Overflow [2,3,10,15–17]. Similarly, there were also works on analyzing programming languages used by developers in GitHub and their relationships to GitHub contributions [4,6–9,11,14]. There are also studies characterizing social network properties of GitHub and Stack Overflow [12,15]. Our work extends this group of research by comparing developer interests in the two social collaborative platforms. To our best of knowledge, our work is the first inter-network study that examines cross-site developer interests in GitHub and Stack Overflow.

6 Conclusion and Future Work

In this paper, we studied the similarity in developer interests within and across GitHub and Stack Overflow. Our findings were based on data for 92,427 users who were active in GitHub and Stack Overflow. We first proposed similarity scores to measure similarity in developers' interests within and across social collaborative platforms. Next, we applied our proposed similarity scores in an empirical study on GitHub and Stack Overflow. We observed that on average, 39% of the GitHub repositories and Stack Overflow questions that a developer had participated fall in the common interests. The developers also do share common interests with other developers who co-participated activities in the platforms. For future works, we intend to we conducted experiments to predict the GitHub and Stack Overflow activities of developers using the insights gathered from our empirical analysis. For example, we can predict developer's GitHub activities using the interests learnt from his or her Stack Overflow activities, and vice versa. We also plan to conduct empirical studies to separate the expertises and interests of developers.

Acknowledgments. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative.

Appendix 1: Example for Cross-Platform Similarity Score Calculation

Figure 4 shows an example for the calculation of *cross-platform similarity score* $Sim^{SO\text{-}GH}(d)$. Consider developer d who has participated activities in GitHub and Stack Overflow. d has forked 2 repositories; *Repository A* which description contains the tag set $\{Java, Android\}$, and *Repository B* which description contains the tag set $\{Java\}$, and watched *Repository C* which description contains the tag set $\{C\#\}$. d also favorited 2 Stack Overflow questions; *Question D* which are tagged with $\{Android\}$, and *Question F* which are tagged with $\{iOS\}$, and answered *Question E* which are tagged with $\{Java\}$. We can estimate d's interests in GitHub (i.e. $I^{GH}(d)$) as $\{Java, Android, C\#\}$ and d's interests in Stack Overflow (i.e., $I^{SO}(d)$) as $\{Android, iOS\}$. The common interests of d (i.e., $CI(d)$) would be $\{Java, Android\}$. Therefore, $Shared^{R}(d)$ would include repositories A and B, while $Shared^{Q}(d)$ would include questions D and E. Thus, $Sim^{SO\text{-}GH}(d) = \frac{|2|+|2)|}{|3|+|3|}$.

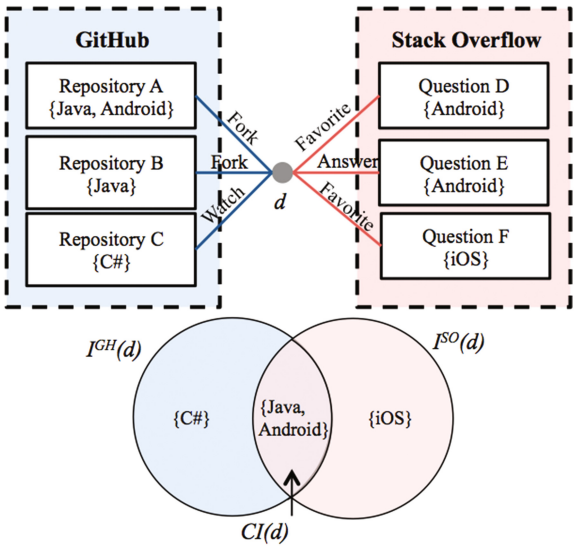

Fig. 4. Example of *cross-platform similarity score* calculation

Appendix 2: Example for Co-Participation Similarity Score Calculation

Figure 5 shows an example for the calculation of *co-participation similarity score* for watch activity $Sim^W(d, co^W(d))$ for developer d. Let us consider two developers d and d' and assume that there are no other developers. Developer d watched repositories A and B. Developer d' co-watched B with d. Thus, $co^W(d)$ is {d'}. In addition to B, developer d' also watched repositories C and D. $Shared^W(d, d')$ would then include B and C as both of the repositories share common interests with the repositories that d watched. $Sim^W(d) = \left[\sum_{d' \in Co^W(d)} \frac{|2|}{|3|}\right] / |1| = 0.67$.

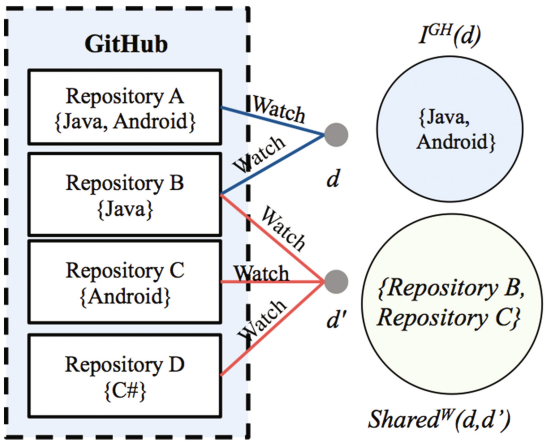

Fig. 5. Example of *co-participation similarity score* calculation for *watch* activity

It is important to note that the *co-participation similarity scores* only consider the similarity in interests between pairs of developers who have co-participated in at least one common repository or question with each other but the developers may have participated in many other repositories and questions different from each other. For example, developers d and d' only watched one common repository but they had watched many other repositories which were different from each other. Also, when computing the co-participation similarity measure between developers who participated in a particular activity, we only consider the interests of the developers in that target activity. For instance, when computing $Sim^W(d)$, we consider how similar are the interests between developers based only on the *watch* activities, i.e., we do not consider repositories forked by the developers or questions answered and favorited by the developers.

References

1. Badashian, A.S., Esteki, A., Gholipour, A., Hindle, A., Stroulia, E.: Involvement, contribution and influence in GitHub and stack overflow. In: CSSE (2014)
2. Bajaj, K., Pattabiraman, K., Mesbah, A.: Mining questions asked by web developers. In: MSR (2014)
3. Barua, A., Thomas, S.W., Hassan, A.E.: What are developers talking about? an analysis of topics and trends in stack overflow. Empir. Softw. Eng. **19**(3), 619–651 (2014)
4. Bissyandé, T.F., Lo, D., Jiang, L., Réveillere, L., Klein, J., Traon, Y.L.: Got issues? who cares about it? a large scale investigation of issue trackers from GitHub. In: ISSRE (2013)
5. Gousios, G.: The GHTorrent dataset and tool suite. In: MSR (2013)
6. Jiang, J., Lo, D., He, J., Xia, X., Kochhar, P.S., Zhang, L.: Why and how developers fork what from whom in GitHub. Empir. Softw. Eng. **22**(1), 547–578 (2017)
7. Kochhar, P.S., Lo, D.: Revisiting assert use in GitHub projects. In: EASE (2017)
8. Rahman, M.M., Roy, C.K.: An insight into the pull requests of GitHub. In: MSR (2014)
9. Ray, B., Posnett, D., Filkov, V., Devanbu, P.: A large scale study of programming languages and code quality in GitHub. In: FSE (2014)
10. Rosen, C., Shihab, E.: What are mobile developers asking about? a large scale study using stack overflow. Empir. Softw. Eng. **21**(3), 1192–1223 (2015)
11. Sheoran, J., Blincoe, K., Kalliamvakou, E., Damian, D., Ell, J.: Understanding "watchers" on GitHub. In: MSR (2014)
12. Thung, F., Bissyandé, T.F., Lo, D., Jiang, L.: Network structure of social coding in GitHub. In: CSMR (2013)
13. Vasilescu, B., Filkov, V., Serebrenik, A.: StackOverflow and GitHub: associations between software development and crowdsourced knowledge. In: SocialCom (2013)
14. Vasilescu, B., Yu, Y., Wang, H., Devanbu, P., Filkov, V.: Quality and productivity outcomes relating to continuous integration in GitHub. In: FSE (2015)
15. Wang, S., Lo, D., Jiang, L.: An empirical study on developer interactions in StackOverflow. In: SAC (2013)
16. Yang, X.-L., Lo, D., Xia, X., Wan, Z.-Y., Sun, J.-L.: What security questions do developers ask? a large-scale study of stack overflow posts. J. Comput. Sci. Technol. **31**(5), 910–924 (2016)
17. Zou, J., Xu, L., Guo, W., Yan, M., Yang, D., Zhang, X.: Which non-functional requirements do developers focus on? an empirical study on stack overflow using topic analysis. In: MSR (2015)