Information Security in Software Engineering, Analysis of Developers Communications About Security in Social Q&A Website

Shahab Bayati[1(✉)] and Marzieh Heidary[2]

[1] ISOM Department, Business School, The University of Auckland, Auckland, New Zealand
S.bayati@auckland.ac.nz
[2] IT Department, Spark, Auckland, New Zealand

Abstract. By the growth of Internet based applications, security becomes an important part of software application development. Software developers should apply security modules, frameworks and technologies on their applications to reduce the security risks, bugs and vulnerabilities. This paper focuses on data analysis on the software development social Q&A Website content around security to elaborate the current state and trend of security issues in software engineering. For this purpose Stack Overflow data as the largest Q&A is selected to analyze. A framework is proposed for data collection and analysis from Stack Overflow. The result of analysis is presented in different schematic and tabular views and a brief discussion on each result is illustrated.

Keywords: Software engineering · Stack Overflow · Information security · Social Q&A · Developers community

1 Introduction

Software engineering (SE) is an iterative socio-technical process. Variety of technologies, programming languages, frameworks and toolkits are applied by different software developers in different stages of software development process. A robust software engineering process should consider security as an important part of development [1]. Security in software development life cycle (SDLC) plays an important role in success and failure of the project [2, 3]. Studies on SE bugs show the high impact and priority of security issues in SDLC. This research focuses on importance of security in software engineering by analyzing real data of software developers' communication.

Software engineers and developers use online forums and Q&A Websites to solve their problems [4]. Stack Overflow (SO) is the most famous online communities for software developers to ask their questions and searching for solutions from answered questions. This paper analyzed SO data about information security to know how software developers think about security. By investigation on SO data about security we can understand the real trends about security matters in SE process. This research clarifies who asked more questions related to security; which programming languages are mostly consider security issues; which technologies are highly impacted by security

M. Chau et al. (Eds.): PAISI 2016, LNCS 9650, pp. 193–202, 2016.
DOI: 10.1007/978-3-319-31863-9_14

questions in software engineering. Also it shows the importance of security in software development communications.

Stack Overflow with around 5 Million users by the time of writing this paper and more than 10 Million questions and Alexa world ranking less than 50 is the valuable resource for data analysis around development process. Along with Q&A facilities, it supports variety of features like user reputation management, up-voting and down-voting, badges and tags which helps to enrich the content of Q&A by social participation. SO as a great resource for data mining is considered in many MSR (Mining Software Repositories) researches [4–6]. Table 1 summarizes some basic statistics about SO. SO have valuable data about security in software engineering which is presented in this study.

Table 1. Basic statistics of Stack Overflow Website

Establish year	Users#	Questions#	Answers#	Comments#
2008	4,978,295	10,729,556	17,561,243	44,240,106
Tags#	Answered questions#	Votes#	Accepted answers#	Up-votes#
43,260	9,442,010	94,032,921	5,987,181	66,051,428

The main contribution of this research is showing the real trend of information security in software development based on the Q&A evidences on the largest development forum which is not presented in related previous works.

The rest of this study is structured as following. Next section presents related works. After that data collection, research questions and methodology are presented in Sect. 3. In Sect. 4 research results are illustrated. Discussion about results is presented in Sect. 5. Finally conclusion and future works are presented.

2 Related Works

In this section a brief review of previously published studies on the area of Stack Overflow mining and security in software engineering is presented.

2.1 Data Mining on Stack Overflow

A classification mining research is done on unanswered questions in SO Q&A Website. Normally the questions in SO are answered quickly (around 11 min after posting), but there are plenty of questions which remain unanswered. In mentioned research the main factors which may affect unanswered questions are gathered [6]. Reputation factors of SO users are mined in [7]. In a text analysis research on SO questions, authors used topic modeling analysis (LDA) to identify the relation among questions concepts, types and codes [8]. A paper focused on analyzing the expertise of developer who attends on SO to answer the questions. They checked the user participation in GitHub projects by

technical term analysis and its relation to tags in SO [9]. Stack Overflow is mined to recommend the code example to the JQuery developers [4]. Mobile programming related posts on SO is mined by LDA text analysis approach in [10]. Mentioned studies shows the value of data analysis on SO.

2.2 Information Security in Software Engineering

In a text mining research on bug report categorization, Naive Bayes and TF-IDF are applied on Bugzilla reports. The main goal of this system was classification of reports into security bugs and non-security bugs. In the mention study TF-IDF approach performed better [11]. A dataset of high impact bugs are presented in a MSR research which listed security bugs as the high weight bugs in software development process. They manually tested bug reports of four open source projects. They categorized security bug more related to product bugs than process. In their views security bugs should be resolved with high priority. With the growth of Internet applications are security bugs received more priority [12].

In another MSR research characteristics of security bugs in Firefox project compare to other performance bugs. They mentioned that security bugs are more critical issues and should be treated faster. Their investigation shows that security bugs are assigned and fixed faster and also re-opened more frequently than others. Security bugs involve more files and developers [13].Importance of security consideration on Android application development to overcome SSL and MITM attacks is mentioned in [14].

Stack Overflow data is analyzed to find the role of security permission in Android applications. They found as the popularity of permission type grows the number of questions on SO also grows which results to better understanding in developers and less misuses [5]. Using online source code repositories like GitHub may lead to leakage of API secret keys in source codes which may happen to all collaborative development environments. A heuristic approach is used to detect and prevent this security problem in [15].

In a MSR research, the effect of pre-release security bugs on post-release vulnerabilities are discussed [16]. In another MSR research text mining applied on bug report for labeling them as security related or non security related. This is done by the reason of security bugs priority and mislabeled bugs [2]. All of these studies show the importance of security data analysis in software engineering.

3 Data Collection and Methodology

In this section data collection process from Stack Overflow social Q&A Website is introduced. SO publicly publishes its data in different formats. In this research SEDE (Stack Exchange Data Explorer) is used which provides an interface for running SQL queries. This dataset contains data about SO posts, comments, users, votes, badges, tags, feedbacks and etc. in a relational format. The result of queries can be downloaded in CSV format for further data analysis. SEDE originally is an open source project accessible from GitHub which uses Ms-Sql Server in data access layer. For the purpose of

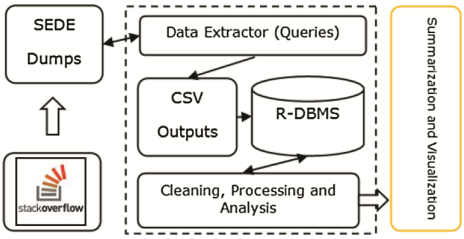

Fig. 1. Data collection and analysis framework.

this research the downloaded CSV files are cleaned and imported to R-DBMS for summarization and analysis. Figure 1 illustrates this process schematically.

Based on literature [17], to retrieve data about information security from SO we use SO security related tags and a dictionary of information security terms [18] along with an ontology of software security [19]. These collected terms are matched with similar SO tags in data extractor queries. Main terms in our queries include Security, Encryption, Authentication, SSL, Decryption, Sql-Injection, Script-injection, privacy, audit, policy, permission, access-control, hack, penetration, confidentiality and more other words which come out from this term list and correlate with them.

This research wants to answer the following questions on the gathered data from proposed abovementioned framework. 1. Which country users ask more about security issues of software development? 2. Which concepts in information security are asked more? 3. Which Programming Languages (PL), Operating Systems (OS), Mobile Technologies (MT) and Web Technologies (WT) have more questions about security? 4. How is the trend of information security related concepts in software engineering community? 5. How is different the answer rate of security related conversation in SE community with general questions? The answers of these questions can show the position of security in SDLC and the portion and priority of security concepts in SE. Based on our best of knowledge there is not any similar data analysis on software communities and repositories and this is the first work in this area which provide variety of future researches.

4 Research Results

This section provides the results of data analysis on different entities of SO. The first analysis aims to answer the question about which country developers are more active in security posts. To reach to this results top 500 users are gathered based on votes and activities, 121 users of top users do not fill their location positions and 10 users have inappropriate values for their locations. Many of the users just mentioned their states and city and countries together which are processed with a query to realize their main country. The pie charts in Fig. 2 show summary of this part of study. Table 2, shows more details. For data analysis and visualization we used R, Excel and D3.js.

Fig. 2. Pie charts for the portion analysis of countries participation. Left up-votes, right activities

Table 2. Tabular view of each country participants.

		1	2	3	4	5	6	7	8	9	10
Up-votes	Country	***US***	***UK***	***SW***	***DE***	***CA***	***AU***	***NL***	***IN***	***DK***	***PL***
	Value	32551	12548	5752	4613	3642	3543	2482	2373	1688	1048
Activities	Country	***US***	***UK***	***DE***	***CA***	***AU***	***NL***	***IN***	***SW***	***CH***	***NZ***
	Value	350	104	46	37	32	24	23	13	13	12
Ratio	Country	***SW***	***SK***	***PK***	***BG***	***ES***	***PL***	***DK***	***MT***	***IL***	***BD***
	Value	442	283	222	188	157	150	141	136	137	122

Another area is about the response time to security questions. From the starting date of SO work in 2008 till December 2015 it is around 385 weeks and the average response time is about 61 min for security related posts which is about 20 min generally. In Fig. 3 the response time is presented over the time.

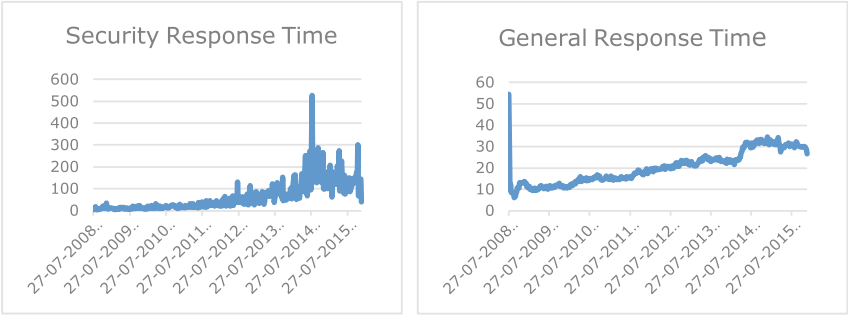

Fig. 3. Response Time (left: security related response time, right: general response time)

Amount of security related posts based on years is mentioned in Table 3. It shows growth of question amount year by year and compared to general questions. These amounts of post are added to previously asked questions.

Table 3. Number of asked questions in each year.

Year	2008	2009	2010	2011	2012	2013	2014	2015	Total
Security	1360	6612	11202	17663	20881	22854	23983	30163	135168
General	58219	343994	699773	1209317	1659879	2069137	2178788	2472763	10729556

To know which tags are most popular tags within security domain we listed them in the Table 4. It also compares security tags position within other general tags.

Table 4. Tags ranks by year.

	2008	2009	2010	2011	2012	2013	2014	2015	Total
C#	1	1	1	1	2	3	5	5	3
Java	3	2	2	2	1	2	2	2	2
Javascript	6	6	4	4	3	1	1	1	1
Android	587	66	10	5	5	5	4	3	5
PhP	7	4	3	3	4	4	3	4	4
.net	2	3	8	13	18	18	27	31	17
MySql	16	12	12	12	11	11	12	12	11
Html	12	13	13	11	9	7	7	8	8
Jquery	22	8	5	6	6	6	6	7	6
Security	35	52	58	77	102	125	151	152	102
Authentication	115	107	137	138	170	192	187	143	151
SSL	166	217	247	242	230	242	165	133	184

Based on available data on SEDE it is applicable to find the trend of security questions in SE. Figure 4 presents this trend for some of the security tags. The same process for more general tags shows the exponentially growth on them rather than linear growth on security issues.

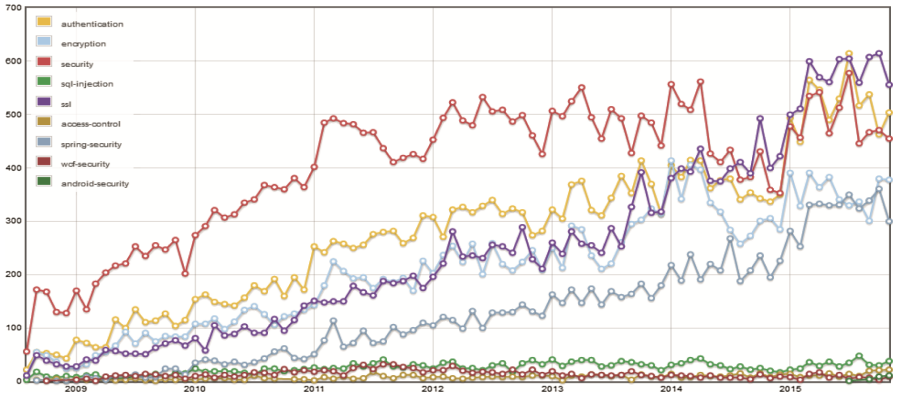

Fig. 4. Trends of some security related tags (Authentication, encryption, security, sql-injection, SSL, access control, spring-security, wcf-security and android-security)

Active users who answered to questions about security in SO are collected for a length of 7 years per week. Results are graphically shown in Fig. 5. As it is clear from this chart by the growth of SO members the active users in security area not only does not raise but also reduce. This problem should be investigated in future research.

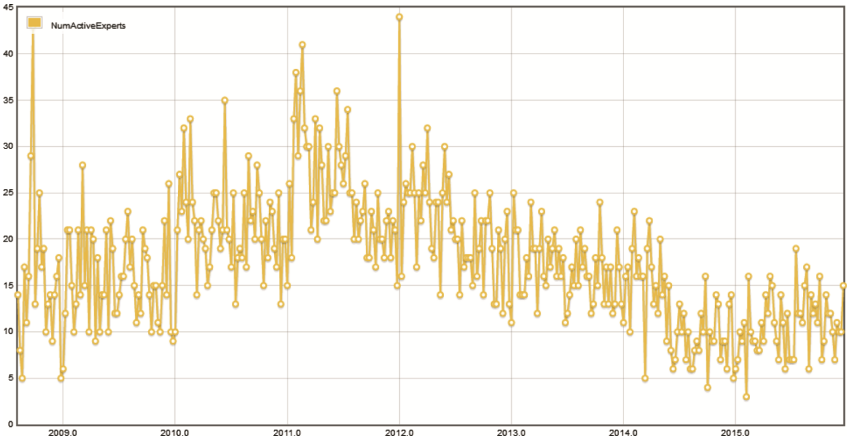

Fig. 5. Trend of active users for security questions

The maximum amount of tags for a post in SO is five. Security related tags can be used together in a post. Table 5 listed the number of times security tags co-occurred in SO posts with their frequencies. Only one question has five security tags together. Figure 6 shows this post.

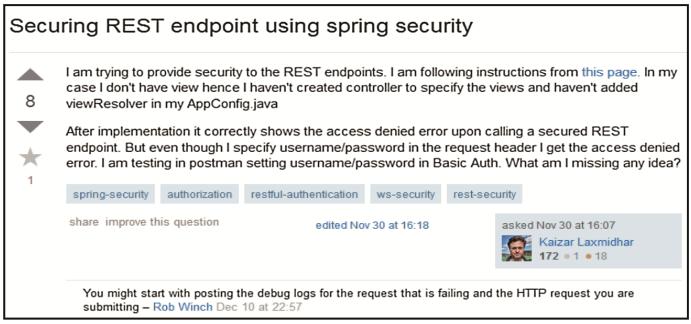

Fig. 6. The only question with five security related tags.

Table 5. co-occurred security tags. (only one question with five security tag)

Co-occurred	5	4	3	2
Post count	1	44	999	13431

Our next analysis counts for different software development technologies like NoSQL databases, Relational DBMSs, Web technologies and operating systems. The number of posts with the frequency of security related tags is presented in these charts. We have selected most trending items for each part. Figure 7 shows the details.

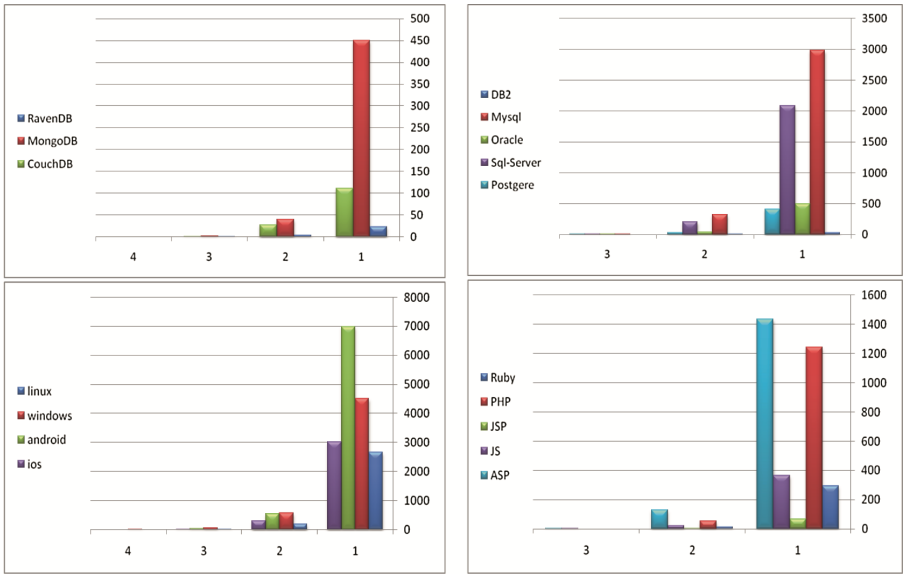

Fig. 7. Security related questions in different technologies, frameworks, OSs, DBMSs with frequency of security tags.

5 Discussion and Implication

By presenting this information about information security issues in software engineering we can gain valuable, actionable and practical understanding of this domain. Although literature shows the importance of security in software development but our analysis illustrated that software engineers do not take care of this important domain. The great need for experts in this area based on different trending technologies and programming languages felt inside the developer community to reduce the response time and accuracy of answers. Internet based frameworks and mobile technologies are top trending tags in security and the security management process is important for them. Most of the experts come from developed countries and big portion of them are from USA. It shows the lack of security efforts in other countries or may be lack of participation of other countries in Q&A knowledge sharing environments.

From response time analysis it can be argued that as we go further the response time to questions gets higher which may be affected by the quality of questions. Normally easy and general questions are asked in the beginning and more difficult ones are coming year by year. Another reason could be the lack of amount of active experts in the area and the growth in the number of questions on the other side.

6 Conclusion and Future Works

This study focuses on information security related posts and its surrounding data on the largest social Q&A community of software developers with different skills and expertise level. Stack Overflow data is used to analyze and response to research questions in this area. It shows that most of the security related questions are asked by US's developers and developed country users. Security, Authentication and SSL are the most usable security tags in questions. The response time for security questions are 3 times higher than the average. A trend analysis on most usable security tags on posts shows the linear growth of security questions compare to exponential growth in general questions amount. Also this study shows most trending technologies and frameworks have more questions about security.

For future works in this area, further analysis on security related data can be done. Furthermore, theoretical research studies can be applied on SO data to find which factors correlate with each other on some of the results shown in this study. Also expert reputation analysis in software security area is another research area for future. Moreover, text mining techniques like topic modeling and sentiment analysis can be applied in future on security related posts.

References

1. Van Wyk, K.R., McGraw, G.: Bridging the gap between software development and information security. IEEE Secur. Priv. **3**(5), 75–79 (2005)
2. Gegick, M., Rotella, P., Xie, T.: Identifying security bug reports via text mining: an industrial case study. In: 2010 7th IEEE Working Conference on Mining Software Repositories (MSR). IEEE (2010)
3. Tsipenyuk, K., Chess, B., McGraw, G.: Seven pernicious kingdoms: a taxonomy of software security errors. IEEE Secur. Priv. **3**(6), 81–84 (2005)
4. Zagalsky, A., Barzilay, O., Yehudai, A.: Example overflow: Using social media for code recommendation. In: Proceedings of the Third International Workshop on Recommendation Systems for Software Engineering. IEEE Press (2012)
5. Stevens, R., et al.: Asking for (and about) permissions used by android apps. In: Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE Press (2013)
6. Asaduzzaman, M., et al.: Answering questions about unanswered questions of stack overflow. In: Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE Press (2013)
7. Bosu, A., et al.: Building reputation in stackoverflow: an empirical investigation. In: Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE Press (2013)
8. Allamanis, M., Sutton, C.: Why, when, and what: analyzing stack overflow questions by topic, type, and code. In: Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE Press (2013)
9. Venkataramani, R., et al.: Discovery of technical expertise from open source code repositories. In: Proceedings of the 22nd International Conference on World Wide Web companion. International World Wide Web Conferences Steering Committee (2013)

10. Linares-Vásquez, M., Dit, B., Poshyvanyk, D.: An exploratory analysis of mobile development issues using stack overflow. In: Proceedings of the 10th Working Conference on Mining Software Repositories. IEEE Press (2013)
11. Behl, D., Handa, S., Arora. A.: A bug Mining tool to identify and analyze security bugs using Naive Bayes and TF-IDF. In: 2014 International Conference on Optimization, Reliabilty, and Information Technology (ICROIT). IEEE (2014)
12. Ohira, M., et al.: A dataset of high impact bugs: manually-classified issue reports (2014)
13. Zaman, S., Adams, B., Hassan, A.E.: Security versus performance bugs: a case study on firefox. In: Proceedings of the 8th Working Conference on Mining Software Repositories. ACM (2011)
14. Zhao, Y., et al.: A new strategy to defense against SSLStrip for Android. In: 2013 15th IEEE International Conference on Communication Technology (ICCT). IEEE (2013)
15. Sinha, V.S., et al.: Detecting and mitigating secret-key leaks in source code repositories. In: Proceedings of the 12th Working Conference on Mining Software Repositories. IEEE Press (2015)
16. Camilo, F., Meneely, A., Nagappan, M.: Do bugs foreshadow vulnerabilities? a study of the chromium project. In: 2015 IEEE/ACM 12th Working Conference on Mining Software Repositories (MSR). IEEE (2015)
17. Pletea, D., Vasilescu, B., Serebrenik, A.: Security and emotion: sentiment analysis of security discussions on GitHub. In: Proceedings of the 11th Working Conference on Mining Software Repositories. ACM (2014)
18. Kissel, R.: Glossary of key information security terms. DIANE Publishing, Collingdale (2011)
19. Raskin, V., et al.: Ontology in information security: a useful theoretical foundation and methodological tool. In: Proceedings of the 2001 workshop on New security paradigms. ACM (2011)