A Study of Innovation Diffusion through Link Sharing on Stack Overflow

Carlos Gómez, Brendan Cleary, Leif Singer
University of Victoria
Victoria, BC, Canada
{cgomez,bcleary}@uvic.ca, leif@leif.me

***Abstract*—It is poorly understood how developers discover and adopt software development innovations such as tools, libraries, frameworks, or web sites that support developers. Yet, being aware of and choosing appropriate tools and components can have a significant impact on the outcome of a software project. In our study, we investigate link sharing on Stack Overflow to gain insights into how software developers discover and disseminate innovations.**

We find that link sharing is a significant phenomenon on Stack Overflow, that Stack Overflow is an important resource for software development innovation dissemination and that its part of a larger interconnected network of online resources used and referenced by developers. This knowledge can guide researchers and practitioners who build tools and services that support software developers in the exploration, discovery, and adoption of software development innovations.

I. Introduction

For many software developers, contemporary software development is increasingly less characterized by writing code and more by the need to aggregate, compose, and debug a diverse set of languages, components, services, and code snippets into functioning systems. In this constantly changing software development landscape, using up-to-date tools and techniques can potentially be both a competitive advantage and an important factor in attracting and keeping the most skilled employees.

In our previous work, we explored what we term the *social programmer ecosystem* and how diversity in technical knowledge and a demonstrated ability to quickly learn, adopt, and disseminate new software development innovations are seen as valued traits amongst certain groups of software developers and companies [6]. Developer communities like Stack Overflow are integral components of this ecosystem. They allow developers to share knowledge about specific issues in a just-in-time manner, and enable the expedited expertise-building software developers require to navigate the myriad of options and decisions that characterize contemporary software development.

According to Rogers [5], an innovation is *"an idea, practice, or object that is perceived as new by an individual or other unit of adoption."* Through a *diffusion* process, innovations are *"communicated through certain channels over time among the members of a social system."* Understanding, supporting, or even thwarting the adoption of innovations by individuals and organizations is the subject of a large body of research on the diffusion of innovations.

However, how developers and software companies discover and adopt new software development innovations — e.g., languages, platforms, libraries, or best practices — is a little understood but important part of contemporary software development. In this paper, we investigate link sharing on Stack Overflow to determine if developers disseminate software development innovations by sharing URLs with other developers. We strive to answer the following research questions:

1) What types of links do developers share on Stack Overflow and do these links point to software development innovations?
2) What types of links are *most frequently* shared on Stack Overflow?
3) Which website domains are most frequently referenced by links on Stack Overflow?

By answering these questions, we hope to inform research into innovation diffusion in software development, and guide researchers and practitioners who build tools and services that support software developers in the exploration, discovery, and adoption of software development innovations.

II. Dataset

To address the above research questions, we analyzed posts contained in the August 2012 Stack Overflow dataset provided by the MSR challenge [1]. Using a combination of regular expressions, we mined the text of the posts for URLs (links) posted by Stack Overflow users. This allowed us to generate a list of unique links and the number of times each link was cited in posts from the dataset. We strived to extract only what would be presented to the user as clickable links on the Stack Overflow website. We ignored raw URLs embedded in the text of the post as we do not consider these URLs as an attempt at link sharing. Using this process, we identified 1,999,026 unique links and 4,197,085 total link citations.

III. Types of Links Shared on Stack Overflow

To help validate our hypothesis that developers disseminate software development innovations through link sharing, we characterized the type and content of the links shared on Stack Overflow. To accomplish this, we performed a manual classification of a random sampling of 1000 links from the dataset described above. We used the following coding protocol.

Step 1: Construct the coding schemas- We extracted a random sampling of 100 links from our link dataset and then

978-1-4673-2936-1/13 © 2013 IEEE

all three authors independently coded the random sample, identifying a set of appropriate categories. We then combined our individual codes to form two final coding schemas: one to classify the website type (Blog, Official Documentation, etc.), and one to classify the website content (Reference, Tool Tutorial, etc.) .

Step 2: Coding the links- Two of the authors then applied this combined coding schema to a 1000 link random sample extracted from the link dataset. Each coder worked independently to categorize the websites according to website type and content type. If a coder judged that a website did not fit one of the existing categories, it was assigned to a generic 'Other' category.

Tables I and II present the results of this analysis. *Category* is the classification derived from Step 1. while *Count* is the number of links that were categorized under the corresponding classification by both coders. We used Cohen's kappa to measure inter-rater agreement between the two coders, achieving .82 and .79 for the *Website Type* and *Website Content* classifications respectively. The small number of links where the coders did not agree on the classification are not represented in tables I and II.

A. Website Type Findings

We observed a large number of *404* (page not found) errors when classifying links: more than 18% (185) of the 1000 sampled links were not available. This link rot[1] represents a significant issue with websites like Stack Overflow and is one of the reasons users are encouraged to duplicate content rather than just post links in their answers[2].

Ignoring *404* errors, *Official Documentation* was the most frequently referenced website type. This type also frequently co-occurred with *Reference* and *API* content types. Official documentation is seemingly still important even when the 'crowd documentation' produced by the Stack Overflow community is available.

The *Wiki* classification arose from a high number of links to Wikipedia articles found when we created the coding schema. This trend was repeated in the larger sample, in which approximately 5% of URLs linked to Wikipedia. These frequently co-occured with the *Reference* and *Design/Architecture* content types. Most articles referenced were related to foundational concepts in computer science and software engineering. This is significant in that it suggests that different online resources are being used by developers for communicating different types of information.

B. Website Content Findings

After classifying the content of the analyzed websites, the most frequently cited content type identified was *Reference*. This content type is associated with websites that contained detailed technical information on a specific topic and were usually more general than tutorials or examples. As noted above, these reference documents often form part of a library's or API's official documentation. This finding suggests that the online resources developers use and share are a complex hybrid of crowdsourced Q&A, combined with official vendor documentation.

Tools (IDEs, Databases etc.) were the second most frequently cited link in our sample. This is a significant finding in itself, however, if taken together with the *API* and *Library* classifications — with all three directly relevant to developing software — they represent the most shared link type at over 17% of the total sample. This finding lends strong support to our hypothesis that Stack Overflow users share software development innovations through linking.

TABLE I
WEBSITE TYPE - 1K RANDOM LINKS

Category	Count
Official Documentation	153
Blog Post	135
Product/Project Website	129
Q&A Post	104
Wiki	52
Other	40
Vendor Website	25
Forum Post	15
Code Repo	11
Online Environment	2

TABLE II
WEBSITE CONTENT - 1K RANDOM LINKS

Category	Count
Reference	102
Tool	85
Question	80
Other	76
API	70
Example	49
Tutorial	38
Opinion/Specialty	37
Answer	33
Book	30
Library	22
Design/Architecture	4
Practices	2

Another important category that emerged in the content classification was the *Question* category which primarily represents links to other Stack Overflow questions. If combined with the *Answer* category, these two categories represent a significant proportion of the link sample analyzed. This points to the development of a complex internal linking structure within Stack Overflow. We are not aware of previous research which has demonstrated this finding.

[1] http://en.wikipedia.org/wiki/Link_rot

[2] http://samsaffron.com/archive/2012/06/07/testing-3-million-hyperlinks-lessons-learned

IV. Types of Links Most Frequently Shared

To get a different perspective on the data, in addition to our analysis of 1000 random links we performed a further manual analysis of the top 100 most frequently shared links using the coding protocol described previously.

Tables III and IV present the results of this analysis. *Category* is the website or content type classification, *Count* is the number of links that were categorized under the corresponding classification by both coders, and *Total Citations* is the total citation count for links categorized under the corresponding classification. Again, due to a high level of inter-rater agreement (.79 for the *Website Type* and .8 for *Website Content*) in tables III and IV, we only present results where both coders agreed on the classification.

A. Website Type Findings

Compared to the results from the random sample of 1000 links, the website types of the top 100 most frequently cited links are much more homogeneous. *Official Documentation* accounts for the vast majority of links analyzed. This again supports the hypothesis that Stack Overflow is part of a larger ecosystem of interlinked online resources and documentation used and referenced by developers.

Next to *Official Documentation*, the *Project/Product Website* category accounts for the majority of the remaining links. This shows that developers refer others to specific products and projects. These links frequently co-occurred with the *Tool* and *Library* content types, again supporting our hypothesis that Stack Overflow users share software development innovations through linking.

Of the remaining significant categories, the *Wiki* category accounts for about 5% of the 100 most frequently referenced links. These URLs all reference specific Wikipedia articles on architectural, security, or computer science topics. This reiterates the findings from the analysis of the 1000 random links and suggests that Stack Overflow users are actively directing developers to Wikipedia for certain foundational topics. This might indicate that Stack Overflow users either consider the content on Wikipedia to be more authoritative or that the Stack Overflow format is not as appropriate to this type of content.

B. Website Content Findings

Compared to the results from the random sample of 1000 links, the website content of the most frequently cited 100 links differs significantly. This time the role of innovation diffusion through link sharing is clearly highlighted by the dominance of the *API*, *Tool*, and *Library* categories. Taken together these three categories represent 62% of the top 100 most frequently cited links on Stack Overflow. We consider this as very strong support for our hypothesis that developers on Stack Overflow share innovations through linking.

V. Domains Most Frequently Referenced

Through our manual analysis of the links posted on Stack Overflow, we identified the types of resources that developers most frequently choose to share with other developers. We verified that innovation sharing is an important part of developer link sharing, but also that Stack Overflow is part of a much larger ecosystem of online resources referenced by developers. To paint a more detailed picture of this ecosystem, we performed an analysis of the most frequently referenced domains contained in our dataset. The results of this analysis are presented in Table V. *Domain* refers to the domain name of the URL, *Unique* is the number of unique links for this domain, *Total* is the total number of citations, and *%* is the percentage of unique links in terms of the total number of citations for that domain.

TABLE III
Website Type - Top 100 Links

Category	Count	Total Citations
Official Documentation	43	35,083
Product/Project Website	29	22,963
Blog Post	6	3,963
Wiki	5	3,635
Code Repo	1	859
Other	1	630
Q&A Post	1	579
Forum Post	0	0
Online Environment	0	0
Vendor Website	0	0

TABLE IV
Website Content - Top 100 Links

Category	Count	Total Citations
API	27	24,178
Tool	18	17,148
Library	17	10,501
Reference	13	9,264
Example	2	1,412
Other	2	1,620
Tutorial	2	1,403
Answer	1	579
Design/Architecture	1	624
Opinion/Specialty	1	565
Book	0	0
Practices	0	0
Question	0	0

A. Most Frequent Domains Findings

Looking at the data in Table V, the number of citations referring to the stackoverflow.com domain immediately stands out. While it is to be expected that developers answering questions on Stack Overflow may refer users to other questions on the site, the large number of these internal links suggests that Stack Overflow is developing a complex internal linked knowledge graph.

The prevalence of platform vendor domains like microsoft.com, oracle.com, apple.com, and python.org largely reflects the popularity of their respective programming languages or platforms on Stack Overflow. However, the ratio of unique to total link citations provides some additional insight into how these domains are referenced by developers. For example, the jquery.com domain has a very high total citation count but a relatively small number of unique links, reflecting a comparatively small API with a very large number of users repeatedly referring to it.

Also of note is the large number of citations for the jsfiddle.net domain. JSFiddle is an online environment where users can post and execute JavaScript examples, and share these examples with others using a unique URL. JSFiddle has become a popular method for developers to try out JavaScript examples and debug JavaScript problems. Its popularity in domains referenced from Stack Overflow provides an insight into software developers' reliance on this tool, and how developers create ad hoc workflows by combining different types of online resources and tools.

TABLE V
REFERENCED DOMAINS - TOP 20 SORTED BY TOTAL CITATIONS

Domain	Unique	Total	%
stackoverflow.com	301,134	449,365	0.67
microsoft.com	128,875	364,315	0.35
wikipedia.org	40,989	178,578	0.23
google.com	67,274	140,190	0.48
github.com	71,761	132,037	0.54
jsfiddle.net	119,330	122,964	0.97
php.net	27,922	110,790	0.25
oracle.com	31,438	78,467	0.40
jquery.com	4,534	76,415	0.06
apple.com	31,270	66,456	0.47
android.com	11,073	64,814	0.17
sourceforge.net	19,328	56,758	0.34
apache.org	18,148	56,081	0.32
python.org	12,510	49,460	0.25
sun.com	13,640	38,555	0.35
blogspot.com	20,049	35,461	0.57
msdn.com	16,170	33,814	0.48
codeplex.com	9,414	33,643	0.28
mozilla.org	7,771	33,464	0.23
mysql.com	6,201	32,262	0.19

VI. RELATED WORK

Since its establishment in 2008, users on Stack Overflow have asked over 4 million questions and provided more than 8 million answers[3]. A question asked on Stack Overflow has a median answer time of 11 minutes [3]. This, combined with the more than 21 million unique visitors and 350 million page views[4] each month, makes Stack Overflow one of the most popular programming Q&A sites on the Web.

Several authors have studied what developers discuss on Stack Overflow. Barua et al. [2] discuss the application of an automatic topic analysis technique to Stack Overflow posts to identify topics in software developers' discussions. The authors analyzed which topics were discussed most frequently and how the frequency of those topics evolve over time, showing trends in the popularity of particular developer technologies. Parnin et al. [4] discuss crowdsourced developer documentation on the Web and the role of different kinds of online developer resources, such as Stack Overflow. By analyzing Google search results for the jQuery library, they found at least one Stack Overflow question on the first page of the search results for 84% of the library's API, showing that contributions of Stack Overflow users can reach a high level of documentation coverage for certain programming topics.

This paper concentrated on the tools and resources external to Stack Overflow that developers choose to refer other developers to. To our knowledge, we are the first to focus explicitly on link sharing on Stack Overflow and the first to utilize Stack Overflow for the study of the diffusion of software development innovations.

VII. CONCLUSIONS

Link sharing is an important activity on Stack Overflow. Our study shows that a significant proportion of links shared on Stack Overflow are referring readers to software development innovations like libraries and tools. Our results also show that Stack Overflow is part of a much larger ecosystem of online resources used and depended on by contemporary software developers. By mining links from Stack Overflow, we have started the process of charting this ecosystem. Our findings will aid researchers in developing a model of innovation diffusion in software development. Future work will investigate methods for assisting developers discover software development innovations by mining innovation sharing activity.

REFERENCES

[1] Alberto Bacchelli. Mining challenge 2013: Stack overflow. In *The 10th Working Conference on Mining Software Repositories*, page to appear, 2013.

[2] Anton Barua, Stephen W. Thomas, and Ahmed E. Hassan. What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, pages 1–36, 2012.

[3] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2857–2866, New York, NY, USA, 2011. ACM.

[4] Chris Parnin and Christoph Treude. Measuring api documentation on the web. In *Proceedings of the 2nd International Workshop on Web 2.0 for Software Engineering*, Web2SE '11, pages 25–30, New York, NY, USA, 2011. ACM.

[5] Everett M. Rogers. *Diffusion of Innovations*. Free Press, 5th edition, 2003.

[6] Leif Singer, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret-Anne Storey, and Kurt Schneider. Mutual assessment in the social programmer ecosystem: an empirical investigation of developer profile aggregators. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, pages 103–116, New York, NY, USA, 2013. ACM.

[3] http://data.stackexchange.com/ (Mar, 2013)

[4] http://www.quantcast.com/stackoverflow.com (Mar, 2013)