CSCI 5352: Network Analysis & Modeling

Upasana Dutta
upasana.dutta@colorado.edu

Sanskar Katiyar
sanskar.katiyar@colorado.edu

1 Proposal

Stack Overflow (SO) is the go-to Question & Answer (QA) website for most programmers: whether its battling through a bunch of incomprehensible errors or looking for a code snippet. Stack Exchange, the parent organization of SO also hosts various QA websites for diverse interests: Science, Business, Technical and Culture. These websites serve as a platform where users can post their questions and other users, who are familiar with the domain, post their answers to those questions. All of these sites have the same design and interface but cater to different audiences.

Even though the website for every field is different, Stack Exchange keeps a unique account ID for every user. This allows a user to participate across these websites without (a barrier of) creating a new account for each website. If a user posts a question or an answer on more than one domain, they will use the same unique account ID while doing this.

A number of studies have been conducted on Stack Exchange, with most of them focussing on the user behaviour and gamification aspect of the QA platform. Stack Exchange encourages users to participate through intermittent reinforcement using badges, reputation, bounty, etc[1][2]. Unlike Quora, another QA platform, Stack Exchange's platform does not inherently represent a network, which is why there have been relatively fewer studies in network science pertaining to dynamic user behaviour within Stack Exchange as compared to general purpose social networks (like Facebook, Twitter, etc.) or Quora [3][4].

We believe that there does exist a network structure in the Stack Exchange platform, although it may not be obvious. Our coarse hypothesis is that there exist latent user communities on the platform which can be discovered by modeling the activity of users on the platform in various ways. We aim at analyzing user activities, study how users post questions and answers in and across domains and try to find answers to some of the following questions:

- **Are users answering questions germane to only specific fields?** Or are users answering questions spread across multiple domains? If a lot of users are answering questions across multiple domains, **what are the domains that have users substantially in common** [6], and can we use this finding to claim which domains are interrelated? **Does this behaviour change with time?**
- Are there users posting questions specific to a given domain, but answering questions specific to a different domain? If yes, can we say that these **users have expertise in one field and inquisitiveness towards a different (probably related) field?**
- **Which are the tags that are significantly common to multiple domains**, and can we claim domains' interrelatedness on the basis of the substantial number of tags they share?
- **Can we identify which tags are similar to each other**, by building a tag co-occurrence network and then running a community detection algorithm on the network? Similarly, can we identify **which users are similar to each other** by building a similarity network based on their activity? [5] Once we have the user communities and the tag communities, **can we claim which user communities exhibit diversity w.r.t answering questions with varied tags?**

  - In tag co-occurrence network, two tags have an edge between them if they appear together in a webpage and the edge weight signifies the number of times they appear together on any webpage. In a user similarity network, two users are connected if both have answered the same question or one has answered the other user's question on any webpage. We then build a bipartite network between users and tags where we have an edge between a user and a tag if the user has posted a question or answered a question that has the tag associated with it. If we find that users belonging to the same community question/answer posts with tags belonging to different communities, it may mean that these user communities exhibit diversity. Or it may also imply an underlying similarity between those tag communities. On the other hand, if the users of same community post questions/answers with tags belonging to a specific community, it can mean that the community comprises of users focused on a particular sub-domain.

- Can we leverage the timestamps at which the users are answering questions? If we find that with time, a lot of questions are being asked and answered specific to a given domain, **can we claim that the domain is trending or gaining importance?**

2 Data

Stack Exchange exports dumps for all its user-attributed content. These dumps are openly available on Archive[1] under the Creative Commons license. Each website has a separate 7z archive for its dump where each dump consists of Posts, Users, Votes, Comments, PostHistory (contains timestamps) and PostLinks in XML format[2].

The type, size of these dumps poses multiple computational challenges. First, we will need to write a custom parser in order to build the required network representation from these dumps. Secondly, we will have to use a distributed framework like Apache Spark to do our analysis on this large-scale network. Initially, we will be excluding Stack Overflow and Mathematics forums from our analysis. This is done due to computational constraints as well as ease of exploratory analysis. We believe that these communities will also skew the analysis due to their exceptionally large size.

References

[1] Andrew Marder. 2015. Stack overflow badges and user behavior: an econometric approach. In Proceedings of the 12th Working Conference on Mining Software Repositories (MSR '15). IEEE Press, Piscataway, NJ, USA, 450-453.
[2] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste and C. Faloutsos, "Analysis of the reputation system and user contributions on a question answering website: StackOverflow," 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara Falls, ON, 2013, pp. 886-893.
[3] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. 2013. Wisdom in the social crowd: an analysis of quora. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). ACM, New York, NY, USA, 1341-1352. DOI: https://doi.org/10.1145/2488388.2488506
[4] Burghardt, K., Alsina, E. F., Girvan, M., Rand, W., & Lerman, K. (2017). The myopia of crowds: Cognitive load and collective evaluation of answers on Stack Exchange. PloS one, 12(3), e0173610. doi:10.1371/journal.pone.0173610
[5] Tanveer Ahmed and Abhishek Srivastava. 2017. Understanding and evaluating the behavior of technical users. A study of developer interaction at StackOverflow. Hum.-centric Comput. Inf. Sci. 7, 1, Article 91 (December 2017)
[6] Yunxiang Xiong, Zhangyuan Meng, Beijun Shen, Wei Yin. 2017. Developer Identity Linkage and Behavior Mining Across GitHub and StackOverflow, International Journal of Software Engineering and Knowledge Engineering

[1] `https://archive.org/details/stackexchange`
[2] `https://ia800107.us.archive.org/27/items/stackexchange/readme.txt`