Temporal Analysis of Inter-Community User Flows in Online Knowledge-Sharing Networks

Anna Guimarães Ana Paula Couto da Silva Jussara M. Almeida
Department of Computer Science
Universidade Federal de Minas Gerais, Brazil
{anna, ana.coutosilva, jussara}@dcc.ufmg.br

ABSTRACT

In this work we provide new insights into the structure of knowledge-sharing networks by approaching them as a dynamic multi-community environment. Focusing on a large dataset collected from Stack Overflow, a popular Q&A site for programming discussions, we analyze how users interact with multiple communities over time, assessing the flow of members across communities. We then employ these estimates of inter-community user flows to identify groups of closely related macro-communities in the network, analyzing how these macro-communities are formed following temporal changes in the flow of users. The insights uncovered by our study about how communities and their relationships evolve over time can be used to improve maintenance of the system and user navigation, as well as to provide a way to track the impact of new topics in existing communities.

1. INTRODUCTION

Online knowledge-sharing networks, such as wikis and question-answering (Q&A) portals (e.g., Stack Overflow[1]), present users with a channel for seeking and discussing information on diverse subjects pertaining to their interests and expertise. Through their collaborative efforts in providing guidance, raising and answering questions, and otherwise engaging in topical and technical discussions, users form interactive communities around topics of shared interests.

Up to now, research into these networks has often neglected their intrinsic community aspect and the dynamics of users interests, in favour of analyzing the content they produce. Moreover, the potential of knowledge-sharing networks (and, indeed, of most online networks in general) as multi-community environments is still largely unexplored [7].

Analyzing the way users organize themselves and their discussions plays an integral part in understanding how these networks function and evolve over time. By gaining insight into how users traverse different topics, we are able to track the impact of a new topic introduced to the network and how it may affect the relationship between existing communities. Such knowledge is also valuable to the design and maintenance of these systems, in order to improve navigation or even guide users towards communities of interest.

In this work, we provide novel insights into the structure of knowledge-sharing networks by approaching them as a dynamic multi-community environment. By focusing on how users interact with multiple and varied communities in the network over time, we establish and investigate the *flow* of users across communities as a measure of their relationship.

We further expand on the idea of inter-community user flows to show that it can be used to identify groups of closely related communities in the network. Discovering macro-communities in this way is a novel form of community detection, as it relies neither on textual attributes to determine semantic similarity between topics in the network, nor on the social graph of user interaction [4], which may be an ineffective portrait of a community when there are hundreds of thousands of active members posting about the same topics but who have a small chance of interacting directly with one another. Instead, our approach focuses on the relationships between users and the varied topics they engage in over time.

Our study focuses on a large dataset collected from Stack Overflow, a popular programming-oriented Q&A site. We analyze the posting activity of 1.7 million users in the 400 most active communities in the network to determine their inter-community user flows. We investigate how these flow values evolve to reflect changes in the communities' central topics and how they influence the development of macro-communities surrounding related topics.

2. RELATED WORK

Due to their potential wealth of information, knowledge-sharing networks have often been investigated for their expert content. In that vein, previous research has looked into assessing question quality [5] and ranking quality answers [1, 6]. Complementarily, the authors of [3] and [8] study contributor profiles in different Q&A networks, in order to understand how user behavior correlates with their importance in the network and the perceived quality of their contributions.

Deviating from those previous work, Zhang et al. [9] propose a probabilistic model to extract clusters of linked users from the social interaction graph of Yahoo! Answers. While the authors recognize that users participate in various communities, they do not investigate this multiple participation or the effect it may have on community evolution. Both of these are key aspects addressed in our study.

[1] http://stackoverflow.com

TAIA'15, August 13, 2015, Santiago, Chile.

In [10], the authors address the issue of overlapping membership across different wikis. Despite finding a positive impact of member overlap on the survival of communities, the authors note that the actual overlap of users was very small, as wikis are separate websites which require individual membership. In contrast, our work analyzes groups of users in a same network, where the interaction between different communities is therefore expected to be more significant.

3. COMMUNITIES IN KNOWLEDGE-SHARING NETWORKS

To better understand the structure and the dynamics of knowledge-sharing networks, we here investigate them from a community perspective. We consider that groups of users who routinely display an interest in a given subject do so by engaging in discussions with other users with similar interests and expertise, thus forming communities around particular topics. We further recognize that users may be involved in several topics throughout their stay in the network and that this multiple participation can shape the relationship and the evolution of different topics and their communities.

For this study, we rely on data from the popular programming Q&A site Stack Overflow. As of May 2015, the site boasted over 9.2 million questions, 15.5 million answers, and 4 million registered users[2]. As a sub-domain of the Stack Exchange network, Stack Overflow's full database can be directly accessed and queried via the Stack Exchange Data Explorer[3]. We utilized this interface to collect recent post and user data from the website, starting from 2012. This data was merged with a database collected for a prior study on Stack Overflow's answer quality [1], containing detailed post data from 2008 to 2012.

In place of a fixed category structure, Stack Overflow allows its users to associate up to five tags to their questions, in order to express the subject being addressed. These tags are used to organize content and ease navigation on the site.

As a starting point, we rely on this tag structure to define *topic-based communities.* These communities denote groups of users who contributed to discussions surrounding a same particular topic. Thus, each tag in the network, along with the posts and users associated with it, form a distinct topic-based community. For example, the group of users who have posted about the Java programming language form a "Java-based" community. This definition is flexible enough to allow membership in multiple communities at any moment, while still linking users with specific topics in the network.

We focus our analyses on activity in the 400 most popular tags. This sums up to 19.8 million posts created by 1.7 million users, which corresponds to roughly 90% of the content created in the website from its opening date in August 2008 up to August 31^{st} 2014[4].

4. INTER-COMMUNITY RELATIONSHIPS

As users may have varied interests and expertise in different subject areas, they are not restricted to a single topic or tag in the network. Indeed, we find that the average user in Stack Overflow participates in 17 communities throughout their lifetimes in the system and is involved in 12 communities simultaneously[5]. Furthermore, users are not static in their interests and often change the community set they belong to. This knowledge motivates our study on the flow of users across different communities, dictated by user dynamics in the network.

4.1 Inter-Community User Flows

As we are primarily interested in the relationship between different communities, we approach the network as a connected graph, wherein each node represents a community for a topic of discussion (tag) and edges represent the *flow* of users between communities. We define the flow from a community C_1 to a community C_2 as the rate at which users in C_1 join C_2[6]. Specifically, in a given time window t, the $flow_{c_1,c_2}(t)$ is given by the fraction of users from community C_1 who joined C_2 during that interval, that is:

$$flow_{c1,c2}(t) = \frac{|C_2(t) \cap C_1(t-1)|}{|C_1(t-1)|}. \quad (1)$$

For our following analyses, we fix the time window t at one month. Thus, whenever we refer to the user flow between two given communities, we are discussing the average monthly flow of users between them, computed during a certain period of time.

Figure 1 shows the cummulative distributions of the computed flow values between all possible community pairs in our dataset, as well as between the subset of the top 100 most active communities, over the whole time period of 2008 to 2014. The distributions are skewed towards lower flow values, especially when considering all 400 top communities, with only 10% of all community pairs displaying a flow value above 0.20. In the smaller subset of 100 communities, we find 25% of community pairs with a flow value of over 0.20 and 50% of pairs with a flow value of at least 0.11. These lower figures are nonetheless still significant: an outgoing flow of 0.11 from community A to community B indicates that members from A approximately have an 11% chance of later participating in community B as well.

We also investigate how these figures change over time by observing the inter-community flows during individual one-year intervals. Figure 2 shows the mean flow values and standard deviations computed for community pairs on each year, on both our sets. Over time, we find an increasing number of community pairs with lower flow values. These start at a mean value of 0.21 in 2008-2009 and steadily decrease to 0.08 in 2013-2014. At the same time, the variability (estimated by the coefficient of variation (CV)[7]) of these values increase, with a CV of 0.73 in 2008-2009 and 1.23 at 2013-2014. Thus, over time, community pairs tend to present more distinct relationship levels. This evolution could be partly attributed to the overall popularity growth of the website. As new users join the network with specific intents, they aid in building up distinct community member bases.

When focusing on specific community pairs, we find diverse evolution patterns of user flow, which are coherent with changes in the relationships between the topics they refer to. As an example, we look at the evolution of the user flow between the Javascript and CSS communities, which correspond to two of the most active topics in the site. In 2008,

[2] Data from Stack Exchange, updated weekly.

[3] http://data.stackexchange.com/

[4] Dataset available upon request.

[5] Considering monthly time windows.

[6] Users may still participate in C_1 even after joining C_2

[7] Ratio of the standard deviation to the mean

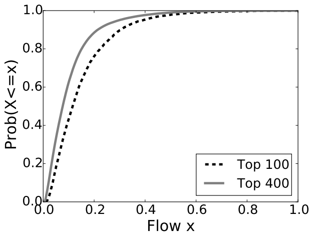


Figure 1: Distribution of flow values between community pairs.

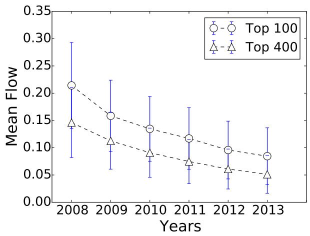


Figure 2: Mean flow value and standard deviation over time.

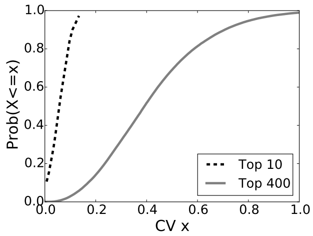


Figure 3: Distribution of CV of user flow between community pairs.

the outgoing flow from CSS to Javascript was 0.75 and it was met with an incoming flow of 0.55. Six years later, in 2014, the outgoing flow from CSS to Javascript only marginally increased to 0.76, while the flow to CSS from Javascript remained stable at 0.55. Other community pairs, such as Flash and HTML, grew more distant over time, with a flow of 0.43 from Flash to HTML in 2009 and a lower flow value of only 0.29 in 2013.

In some cases, rather than a natural distancing between topics (as with HTML and Flash), we can observe the disruptive effect that one community may have on existing relationships. For example, the Ruby on Rails 3 community starts out with an incoming flow of 0.61 from the Active Record community in 2010. Shortly after the launch of Ruby on Rails 4 and its introduction as a topic in the network in late 2012, this flow value drops to 0.41 in 2013. During this same period, the outgoing flow from Active Record to Ruby on Rails 4 was 0.45. This is a good illustration of how users quickly adapt to the evolution of topics in the network and how the emergence of new technologies (and their respective communities) can impact previously well-established relationships between existing communities.

To summarize these distinct evolution patterns, we estimate the variability in the user flow for each community pair over time by computing the coefficient of variation (CV) of the user flows measured for the pair[8] in all six years covered by our dataset. The higher the CV computed for a given pair (C_1, C_2), the greater variability observed in how the flow from C_1 to C_2 evolved over time. We summarize these results in Figure 3, which shows the cumulative distributions of the CV values for all community pairs in our dataset.

Overall, roughly 70% of all pairs had a CV below 0.5 and less than 1% had a CV over 1.0. Thus, most inter-community flows in the network tend to remain roughly stable over time, suffering from moderate to little variation in consecutive time windows. As a more pronounced example of this, we also single out the top 10 communities which presented the highest flow values in 2008, also shown in Figure 3. Only two of these community pairs had a CV above 0.1, an already low value in itself. We note that this set of communities mainly refer to broader topics (such as "Ruby", as opposed to "Ruby on Rails 3") and therefore may be more robust to temporal changes in the network.

4.2 Macro-Communities

As communities and their relationships, as dictated by user flows, become more well-established in the network, it is possible to find patterns between them, such as a common theme. Thus, by observing users' behaviour and trajectory in the network, we can recognize groups of related communities purely based on inter-community flow dynamics, rather than having to rely on user interactions or semantic similarity [4].

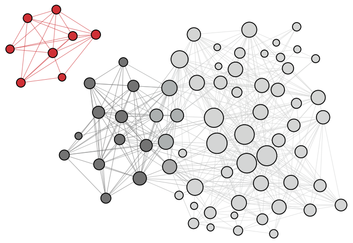


Figure 4: Macro-Communities in the top-100 set.

In order to identify these macro-communities, we apply the Clique Percolation Method (CPM)[9] [2] over the community graph. For this task, we consider only the top 10% of edges with the highest flow values, which denote a more significant relationship between the linked communities. The CPM additionally discards a small number of communities which only appear in isolated small cliques (with 3 nodes or less). This corresponds to 13% of nodes in our top-100 set and 9% of nodes in the top-400 set.

Figure 4 illustrates the macro-communities we found on the top-100 set, considering an aggregated view of flows over the whole 6-year period of our dataset. Overall, we find a relatively small number of community clusters, with at most 5 clusters in the top-400 community set and 3 in the top-100 set. Both cases feature one larger cluster, containing over half of the communities (259 in the top-400 set and 51 in the top-100 set). These clusters include very popular communities (with the greatest number of posts), to which several smaller "satellite" communities are connected with high incoming flows. As an example, in the top-100 subset, the "Javascript" community was connected to 88 communities. This points towards popular communities acting as hubs in the network, as they gather and redistribute users from and to smaller communities.

These popular communities have another interesting effect in macro-community composition. As macro-communities describe groups of densely connected communities, we ex-

[8] Each pair appears twice, once for each direction (C_1, C_2) and (C_2, C_1).

[9] Implementation available at
http://github.com/michelboaventura/rcpm

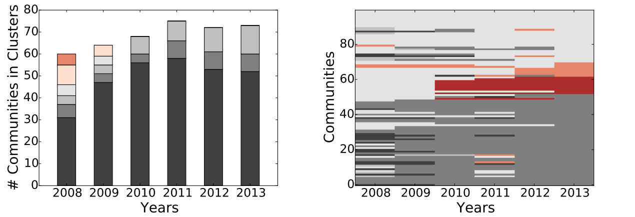
pect the average flow between communities in a same cluster to be greater than the flow between distinct clusters. However, because the CPM allows communities to simultaneously belong to different clusters, exceptions do occur. In both our sets, two distinct clusters featured the same four very popular communities (namely, those surrounding the ".net", "c#", "windows", and "asp.net" tags). This makes it so that their overall shared member base is similar, which results in a high flow across clusters. In the top-100 set, we found both an incoming and outgoing flow of 0.31 between the two macro-communities, which stands above the average flow of 0.28 across different macro-communities.

The macro-communities we discovered in our aggregate analysis are nonetheless internally cohesive, both in terms of the user flow across their communities and in terms of their underlying common topic. Among the three macro-communities in our top-100 set, one seems related to general programming discussions, including a majority of communities surrounding programming languages and operating systems. Another cluster refers to Windows and technologies commonly associated with it, such as Visual Basic and the .net framework. The third cluster, which interestingly featured no community intersection with the other two clusters, encompasses discussions strictly related to Apple products and associated technologies, such as iPad, iPhone, OSX and Objective-C. In addition to these three, the top-400 set also features one macro-community surrounding programming IDEs and their many extensions, such as Eclipse, NetBeans and jUnit, and another macro-community exclusively about Ruby and Ruby-on-Rails technologies.

In an effort to better understand the dynamic relationships between the communities in the network, we again analyze their evolution in individual one-year periods.

Figure 5(a) illustrates the clusters/macro-communities we identified in the top-100 set in each year, along with the number of communities that belonged to each cluster in that year (each bar is divided into a number of sectors corresponding to the number of clusters identified in the year). As in our aggregate analysis, every year sees the presence of one large cluster, involving a majority of the communities, accompanied by a few smaller clusters. These are more fragmented during the early years of the network and display varied compositions. For instance, 6 macro-communities were identified in 2008. In later years, as communities and their relationships grow more established, these clusters also become more distinguished (converging to 3 clusters in 2010) and begin to feature a similar core of communities over time. The larger cluster continuously features a same group of core communities (the same 24 communities were present in the cluster in every year) while also drawing in new communities. As new topics appear in the network over time, communities arise that may help bridge existing communities, thus influencing the make-up of clusters at different periods.

This variability in cluster composition is displayed in Figure 5(b), which relates each community to its containing cluster (a different color is used to represent each cluster[10]) in each year. We see how individual communities can belong to different macro-communities in different periods, which translates how inter-community flow dynamics vary over time. Nonetheless, we see a tendency towards communities settling in to specific clusters as time goes on, which points to their growing maturity as topics in the network. The few communities which continue to feature in different clusters are often those that show up in the overlap between macro-communities and that thus relate to multiple disciplines (e.g., discussions about Microsoft's C# programming language may appear in the programming macro-community or in the smaller Windows-based macro-community).



(a) Number of communities per cluster

(b) Community membership in clusters

Figure 5: Temporal evolution of macro-communities.

5. CONCLUSIONS

In this work, we have presented a study on the dynamics of online knowledge-sharing networks as multi-community environments, wherein users can transition between communities devoted to different topics of discussions. Through our analyses of the inter-community user flows, we have found interesting patterns in the evolution of the relationship between communities in the network.

This knowledge can be valuable to drive ad placement and recommendation mechanisms, in order to guide users to communities of interest, promote new communities or encourage participation in communities that are losing popularity. It could also help track the impact of different topics in the evolution of existing community relationships.

6. REFERENCES

[1] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. Exploiting User Feedback to Learn to Rank Answers in Q&A Forums: A Case Study with Stack Overflow. In *Proc. ACM SIGIR*, 2013.

[2] I. Derényi, G. Palla, and T. Vicsek. Clique percolation in random networks. *Physical review letters*, 2005.

[3] A. Furtado, N. Andrade, N. Oliveira, and F. Brasileiro. Contributor Profiles, Their Dynamics, and Their Importance in Five Q&A Sites. In *Proc. CSCW*, 2013.

[4] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3), 2012.

[5] S. Ravi, B. Pang, V. Rastogi, and R. Kumar. Great Question! Question Quality in Community Q&A. In *Proc. ICWSM*, 2014.

[6] T. Sakai, D. Ishikawa, N. Kando, Y. Seki, K. Kuriyama, and C.-Y. Lin. Using graded-relevance metrics for evaluating community qa answer selection. In *Proc. ACM WSDM*, 2011.

[7] C. Tan and L. Lee. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proc. WWW*, 2015.

[8] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the Social Crowd: an Analysis of Quora. In *Proc. WWW*, 2013.

[9] Z. Zhang, Q. Li, D. Zeng, and H. Gao. Extracting Evolutionary Communities in Community Question Answering. *Journal of the Association for Information Science and Technology*, 65(6), 2014.

[10] H. Zhu, R. E. Kraut, and A. Kittur. The impact of membership overlap on the survival of online communities. In *Proc. ACM SIGCHI*, 2014.

[10] Light-gray is used for communities that did not belong to any cluster.