SMM638: Network Analytics
Final Course Project
(Group 6)

A Case on Stackoverflow: Why do Users Participate in Certain Threads?

Fung, Ka
Laubenthal, Christoph
Vijitratanakit, Pakatorn
Bonaventura, William
Low, Ming
Han, Jiacheng

20 December 2019

Contents

Chapter 1

Introduction

This paper aims to delve into and better understand the topic of link formation in the online setting, specifically looking at the online platform, Stackoverflow, to help answer the research question:

$$RQ: \textit{Why do users participate in certain threads?}$$

Stackoverflow was created in 2008, which allows students and professionals to post queries and to answer questions about programming. Users of the platform can participate and engage with Stackoverflow in several ways; these being posting questions, answering questions, commenting on questions and answers as well as up/downvoting.

Through analysing the SNAP dataset on Stackoverflow, the aim is to infer the motivation for different types of participation from the data. This paper will first go through briefly the literature theories underpinning this research. Following this, the network analytics pipelines that have aided this research will be analysed and key insights and interpretations emerging from the network analyses will be mentioned.

Chapter 2

Literature Review

Three types of link formation in social affiliation networks: triadic, focal and membership closures. In all of these three closure types, a group of three nodes closes ties to each node, forming a triangle. In the timeframe leading up to the final closure, the three nodes only form a line. This previous formation may provide a hint at the reason why the final connection took place. In theory, the already established ties between the three nodes can act as a push for the final connection to form. The differences between triadic, focal, and membership closure can be found in the dimensions of the network, and in the sort of ties that are established at certain times.

Triadic closure is the simplest sort of closure in the sense that it can be observed in one-mode networks. Here, the nodes are thought to be of the same type in a flat, unidimensional space. Triadic closure happens when nodes A, B, and C are connected through node B. As A has already formed a relationship to B, and B has formed a tie to C, it is hypothesised to be likely that A and C also form a tie.

Focal closure and membership closure both only exist in bipartite (two-mode) networks. Here, two modes – or types of nodes – exist, and all the nodes of either one or the other mode are usually connected to nodes of the opposite mode. In the case of Stackoverflow, one mode represents the social space of user-to-user interaction, while the other mode captures the interaction of users with certain topics, which can referred to as the problem space. The ties between the two modes represent the membership of users in certain topic communities.

When node A and node C, both part of the social space, are tied to the same topic community, node B, focal closure describes the the process of A and C forming ties as a result of their shared interest in and interaction within B.

Conversely, when A is connected to C and C is a member of topic community B, membership closure captures the formation of a tie between A and B as a result of C acting as a bridge between A and B and thereby provoking A's interest in B.

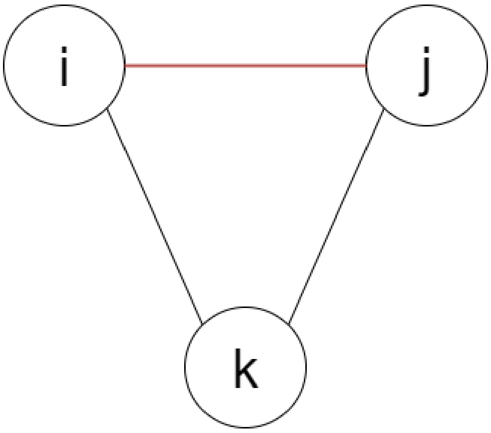

Figure 2.1: Triadic Closure

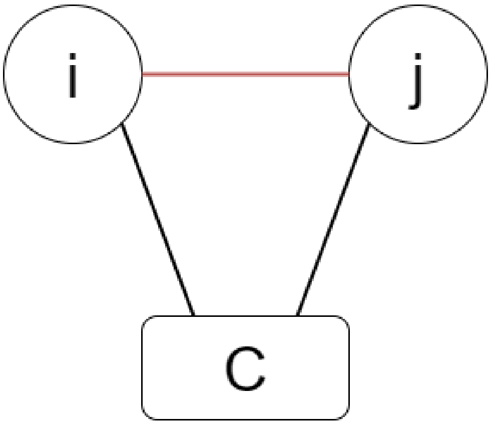

Figure 2.2: Focal Closure

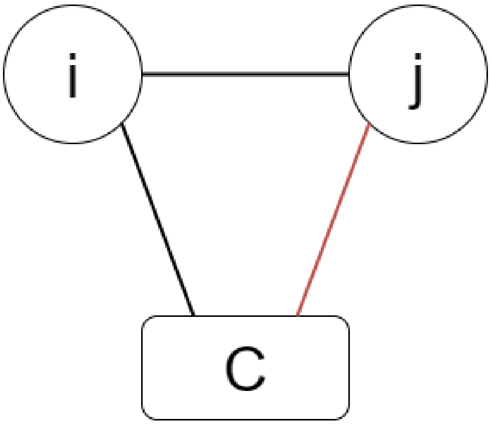

Figure 2.3: Membership Closure

The theory is based on the concept of homophily, which describes the tendency of individuals to form relationships with other individuals, with whom they have certain characteristics in common. In the case of triadic closure, the decisive characteristic is the common alters (or ties) that nodes share, which makes them prone to form new ties to nodes that already have alters in common. Since all nodes are of the same type, there is no differentiation between the effects that certain shared nodes have on tie formation.

On the other hand, focal and membership closure consist of two types of nodes. The decisive characteristic is again the common alters that nodes share, which again determines tie formation. This time, however, either the shared alter between two users is specifically a topic domain (focal closure), or the shared altar between a user and an interest domain is a different user (membership closure). As a result, the type of topic domain present determines the users' motivation in forming ties.

Chapter 3

Methodology (models, metrics, algorithms)

The SNAP dataset has been chosen to undergo network analysis. This dataset includes three elements within the types of participation discussed above – answers to questions and comments to both questions and answers. It reveals a two-mode network projected into a one-mode weighted network. With the two identified attention structures within the Stackoverflow setting; one being the attention in the problem space (user - topic) and the other in the social space (user-user), we use the answers to questions dataset to analyse the interaction with the topics and the dataset with comments to both questions and answers to observe and analyze the social interactions between users.

Through looking at these two spaces, we can observe the formation of triadic, focal and membership closures. First, community detection in the problem space allows identification of individual communities that are hypothesised to be different knowledge domains related to programming, for instance, programming languages like python, SQL or R. Triadic closure can be observed in the social space with the Kossinets-Watts model and after this, we aim to combine the findings from these two spaces, looking at formation of ties over time to then look for potential focal and membership closures.

3.1 Stochastic Block Modelling Algorithm: community detection within the problem space

To begin with, looking at the answers to questions dataset, we identified communities formed in the problem space, made up of users with similar focus on a certain topic.

Applying a community detection onto the answers to questions dataset would allow the detection of different communities to which we hypothesised to be different programming languages. The algorithm, nested stochastic block model was chosen to identify the different communities within the answers to questions dataset over a period of 6 months since the inception of StackOverflow. This algorithm was chosen over the more commonly known algorithm such as the Girvan-Newman algorithm mainly because of its availability in the library Graph-Tool. Furthermore, one limitation of the Girvan-Newman algorithm is that it performs badly on larger graphs (igraph? and Bishop, 2019). Our answers to questions dataset contains over 5000 nodes therefore, this discourages us from using the Girvan-Newman algorithm.

The availability of the stochastic block model algorithm in Graph-Tool was the main reason we decided to apply this algorithm on our dataset. Inside Graph-Tool we found another algorithm, nested stochastic block model. This algorithm has its advantages over the normal stochastic block model in that it allows us to find small communities within the community again. This is done by recursion where stochastic block modelling is applied onto partitions created by the previous stochastic block model (Graph-tool.skewed.de, 2019).

Due to the size of the network we investigated only one community partitioned from the nested stochastic block model algorithm.

3.2 Kossinets-Watts Model: examining triadic closure within the social space

As part of data preparation, the comments to questions and comments to answers datasets were merged for manipulation as they both suggest the presence of social interactions within the user to user space. To attempt to address the triadic closure empirically using network data, the Kossinets-Watts Model is used to understand the probability that links form between two users given they have k common contacts. Through looking at the probability of triadic closure, we can then determine whether such closure is, comparative to membership and focal, the main push for final connections to form. In other words, whether it is the basis to why people participate in certain threads, in particular, commenting on questions and answers.

3.3 Examining focal closure

We defined observations of focal closure as instances, in which new ties formed between users in the social space, given that both users had already been a member of the same topic community before. They presumably connected because of their previous interest in the same community and therefore ended up directly interacting with each other in the social space.

Projected onto the case of Stackoverflow, the instances of focal closure we observe describe the interaction between users in the same comment thread, given that both users had previously posted or answered a question in the same topic community on the platform. More generally, the analysis of focal closure in this context can explain the degree to which the membership in certain topic communities predicts the commenting behaviour of nodes in a specific community.

For each timeframe, the amount of newly formed ties in the social space between nodes that were both previously members of the same community is compared to the total number of newly formed ties:

$$\begin{aligned}
&ties_{uf} = ties\ to\ nodes\ in\ focus\ in\ the\ user\ space\\
&ties_{ua} = ties\ to\ all\ nodes\ in\ the\ user\ space\\
&ratio = \frac{ties_{uf}}{ties_{ua}}
\end{aligned}$$

The timeframe of our analysis includes the first six months since the inception of the Stackoverflow platform in August 2008. For the analysis, five timeframes are relevant since the first month – t0 – sets the baseline in terms of the number of nodes that first became members of the community in the knowledge domain. For each of the following months, the number of newly formed ties is then split into ties whose nodes were either already part of the community in the previous time frame – indicating focal closure – or not. This relationship is then expressed as the aforementioned ratio between ties indicating focal closure and the total amount of ties.

3.4 Examining membership closure

Membership closure is defined as a node being involved in a particular focus as a function of the number of friends already involved in that focus. We investigated this by looking at the two spaces in the SNAP dataset.

The problem space is hypothesised to be the knowledge domain relating to different programming languages or in this case, the focus. The social space is therefore, the interactions between the nodes or more generally the friendship ties between each user. Thus membership closure in the domain of the SNAP dataset is a user being involved in a programming language as a function of how many users in the social space are already involved in that programming language. This can be defined as a function of:

$$ties_{uf} = ties\ between\ nodes\ that\ were\ previously\ members\ of\ the\ same\ focus\ group$$
$$ties_{ua} = all\ newly\ formed\ ties\ in\ the\ user\ space$$
$$ratio = \frac{ties_{uf}}{ties_{ua}}$$

We compared between different time periods in order to see the ties to the focus forming over time. Specifically we looked at the snapshot of the network at the current time in both the problem space and social space at time t and time $t-1$. In order to construct the networks for both problem space and social space at different times we took the largest community detected from the community detection algorithm as the ground truth. We only looked at the nodes inside that community and created a problem space where nodes in that community are involved, as well as all ties that involve those nodes. Hence, doing so we know that all nodes within the data share the same focus. We do the same to construct the data for the social space looking at the nodes within the ground-truth community and all ties that involve those nodes. Therefore, both the problem space and social space also includes nodes that are not in the ground-truth community. This allows us to investigate the question of "what percentage of ties a node has that is already involved in the focus causes the node to become involved in the focus."

We started off with constructing the problem space and social space network at t_0 and look at all the nodes in the social space and see whether they are involved in the focus by examining the problem space and see if they are involved. We then construct the network again for the problem space and social space at time t_1. We do the same as previously where we examined nodes that are and are not involved in the focus. Next we compare nodes that are not in the focus at t_1 to nodes that are not in the focus at t_0. We do this because any node that is not in the focus at t_0 and does not appear in the list of nodes not in the focus at t_1 means that, that node has become involved in the focus. Consequently we look at that portion of nodes which has entered the focus and look at how many ties to other nodes it has in the user space and what percentage of those ties are ties with nodes already involved in the focus.

This workflow can be generalized to look at other time periods as well where we only need to initialize the base case t_0. We would then compare t with $t-1$, or the previous time.

Chapter 4

Results and Key Insights

4.1 Community Detection

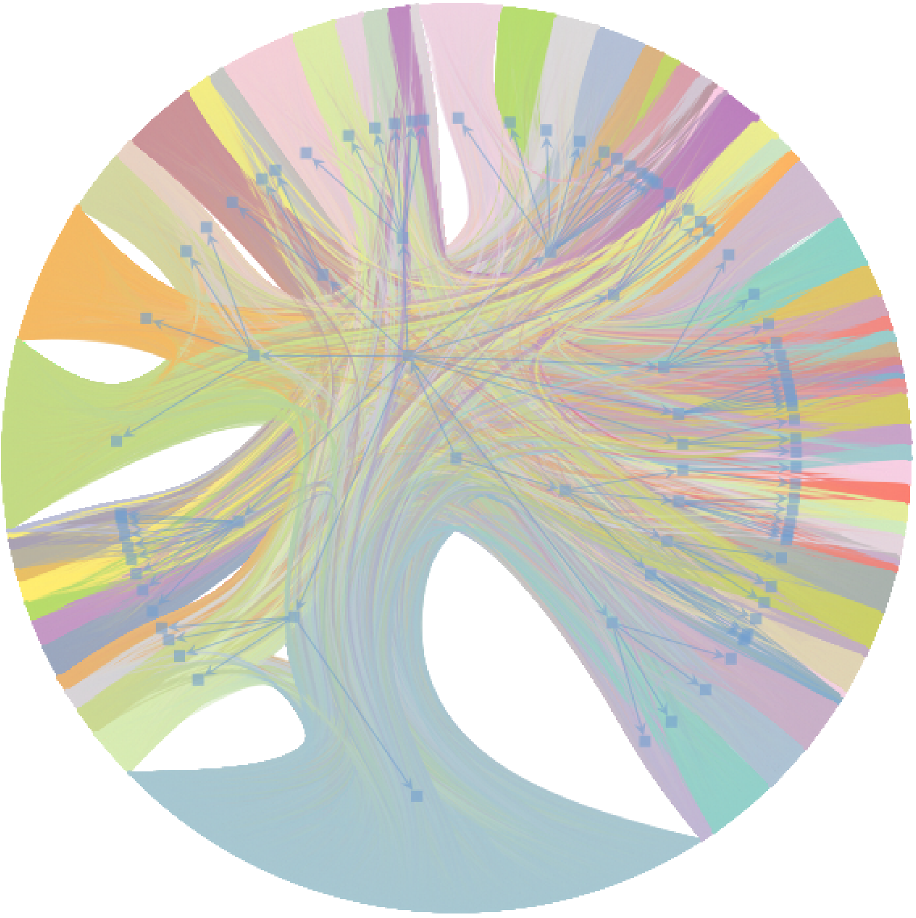

Figure 4.1: Community Detection

The figure above shows the result of applying the nested stochastic block modelling algorithm to the answers to questions dataset. We can see multiple levels forming as a result of the recursion. These are smaller communities within communities. In total there are 4 levels of communities. There are a total of 76 small communities forming at the lowest level with the largest community having 4000 nodes. At the next level there are 15 communities forming with the largest community having 10000 nodes. The community in the second level with 10000 nodes contains too many nodes to do analysis on it therefore, we conducted our analysis on the lowest level community with 4000 nodes.

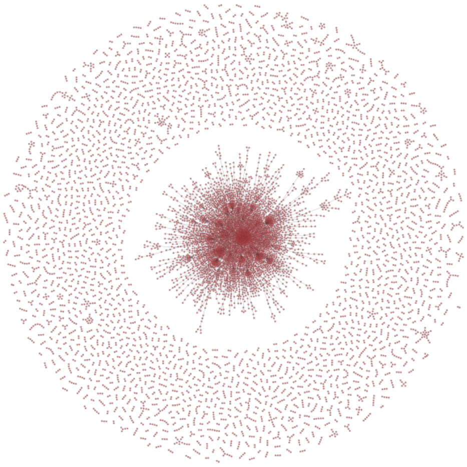

Figure 4.2: Network Structure of One Community

Figure 4.6 illustrates the network structure of one community with many sub-communities within the Stackoverflow answers to questions participation setting. We can assume that this community represents one particular knowledge domain of programming and sub-communities are highly likely to be different threads or "problems" regarding a similar topic area.

The community has a clear core-periphery structure, with peripheral nodes being completely detached from the core nodes. At the same time, such peripheral nodes are very sparsely connected to each other. Such peripheral structure suggests that many users have either had a one-off engagement with a particular thread or have engaged very few times. To add, the majority of the nodes in the core network have very few connections and only very few nodes have a large amount of ties (470 degrees). From this, we can deduce that such nodes with high degrees could be seen as 'bridges', linking two or more communities and in the Stackoverflow setting, we can identify them as users that are active in multiple threads across different knowledge domains while others stay within one.

4.2 Kossinetts-Watts Model

Initially, in the trial and error phase, the Kossinetts-Watts Model was first implemented on the full dataset with a time frame of 5 months from day of inception of Stackoverflow. A total of 2177K nodes were included in the analysis and the output could be seen below.

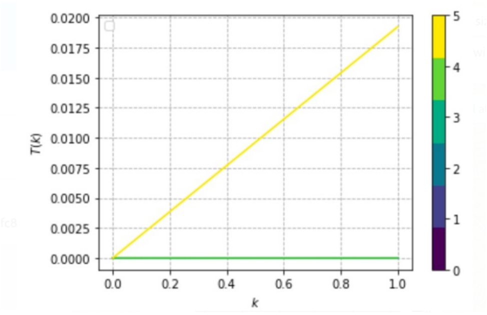

Figure 4.3: Network Structure of One Community

It is clear the highest number of shared contacts between any two nodes (hereafter, users) is 1 and from month 0 up to month 5, the probability that a link will form between two users with k shared contact is 0 for all values of k. In month 6, the probability of link formation increases steadily as the number of k increases but such result is nonetheless not insightful. There is, in general, no clear evidence for triadic closure.

In hope to find better insights to triadic closures within the user-to-user social interaction space, a reduced dataset was then used. As communities were detected in the previous section based on the questions to answers dataset, users from one particular community in the problem space were singled out and were examined in the social space (comments to questions and answers) instead and the results from this is shown below.

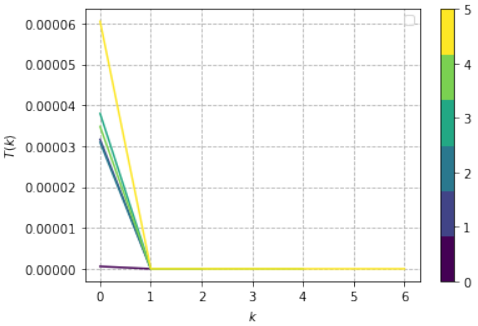

Figure 4.4: The Effect of Common Contacts (k) on the Formation of Links (T(k)) – based on one community

The results suggest high randomness of the data with nodes that are not well-connected and hence doesn't fit the model well; the highest number of shared contacts between any two nodes, again, is 1. The results from the Kossinets-Watts model, therefore, show that triadic closure is not prominent in the Stackoverflow setting as the significance of shared common connections between users in the community we investigated is low. Therefore, to answer the question regarding motivations behind participation, we need to look at focal and membership closure for better understanding.

4.3 Focal Closure

Time	**Ratio**
t_1	0.830
t_2	0.822
t_3	0.804
t_4	0.807
t_5	0.800

Results show that the percentage of newly formed ties in the social space that can be attributed to focal closure remains pretty constant with time at around 80%. This means that about 80% of the nodes that formed ties in each time period already shared the same topic domain in the previous time frame. However, one can also observe a declining trend, as the ratio reaches its maximum (83.0%) at t1 and its minimum (80.0%) at t5.

Since the vast majority of newly formed ties in the social space can be attributed to focal closure, the nodes in the social space appear to establish new connections largely within the community they're associated with, and much less to nodes outside the community. The results are aligned with the concept of homophily, as the likelihood for nodes to form new ties is much higher if both nodes were previously already part of the same community.

In the context of Stackoverflow, the results indicate that users comment mostly on questions or answers where other members of their community interact as well. As a consequence, the users' commenting behaviour is far from random; in fact, it appears to be directly connected to the topic community where those same users also pose and answer questions. This observation is pretty intuitive, however the degree to which it manifests itself in the analysis of focal closure here is quite stark and can certainly provide value in the quest for a better understanding of the Stackoverflow network.

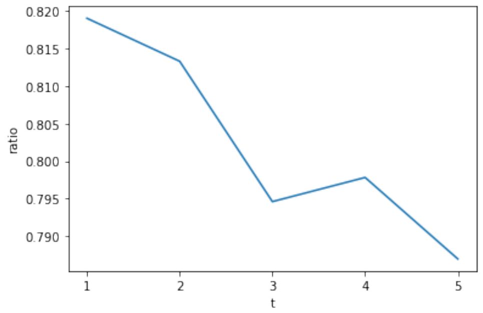

Figure 4.5

4.4 Membership Closure

Time	Ratio
t_1	1.000
t_2	0.871
t_3	0.944
t_4	0.982
t_5	0.954

Looking at the results in the table we can see that from t_0 to t_1 all the nodes that have become involved in the focus between then, 100% of the ties of those nodes are with nodes that are already participating in that focus.

In general we can see a large proportion of ties are with nodes already involved in the focus hence this is intuitive in that we expect nodes to more likely become involved in the focus if they have many friends which are already involved in that focus.

Overall, in the context of Stackoverflow, the more ties a user has with other users who are already participating in a programming language the more likely that user participates in that programming language as well. This could be due to the same knowledge domain being required and it would be counter-intuitive for a user to participate with other users in one language but participate in another language when it comes to questions and answers, knowing that the knowledge required for 2 programming languages is not exactly identical.

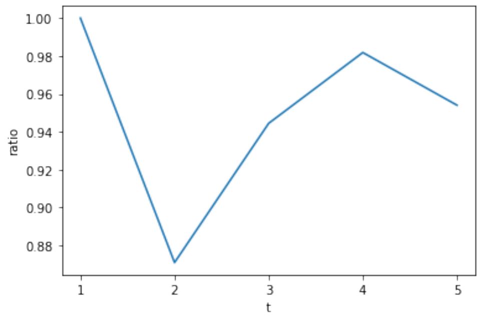

Figure 4.6

Chapter 5

Limitations and Conclusion

In conclusion, our analyses provide new and insightful answers to the motivations behind user interaction on Stackoverflow. Focusing on a small community of about 1,000 nodes, which we detected in the knowledge domain using block-modeling community detection, we first observed that those nodes were only sparsely interconnected in the social space, which we defined as user interactions in comment threads. The majority of nodes present in the social space had no connection to the inner core of the community and formed small isolated communities in the outer rings of the network. The larger communitites seemed to be connected through a few very active users with high in-betweenness scores. Thus, a tiny minority of nodes was active in multiple comment threads, while the majority only interacted with at most two comment threads per user.

As a result, our analysis of triadic closure using the Kossinetts-Watts model yielded results of limited insight. Extremely few nodes had more than one connection, and therefore the probability of tie formation given multiple shared alters declines to zero, or was so small that it was basically equivalent to zero.

The results for focal and membership closure were much more insightful. Using the same community, we compared tie formation for both types of closure during the first six months since the inception of Stackoverflow. Our results show that both focal and membership closure can be used to explain tie formation over the six months for this community. In comparison, membership closure is seen to contribute to on average 95% of newly formed ties in the social space, while focal closure is observed to contribute to an average of 81% of newly formed ties in the social space.

Future research may take a closer look at the difference between communities on Stackoverflow, which is certainly a limitation of our analysis. Furthermore, the computation of focal closure and membership closure ratios is based on the data from one community, which may skew the results and limit the comparability to similar or future work. Finally, future research should aim to find more insightful results from the analysis of triadic closure. To maximise the chances of success, we recommend to focus on a different community within Stackoverflow and possibly to include more factors in the data used.

Bibliography

[1] Easley, D. & Kleinberg, J. 2010. Networks, crowds, and markets: reasoning about a highly connected world, Cambridge University Press, Cambridge.

[2] Graph-tool.skewed.de. (2019). Inferring modular network structure — graph-tool 2.29 documentation. [online] Available at: https://graph-tool.skewed.de/static/doc/demos/inference/inference.html [Accessed 20 Dec. 2019].

[3] igraph?, W. and Bishop, M. (2019). What are the differences between community detection algorithms in igraph?. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/9471906/what-are-the-differences-between-community-detection-algorithms-in-igraph/9478989#9478989 [Accessed 20 Dec. 2019].

[4] Stack Overflow. (2019). Stack Overflow - Where Developers Learn, Share, & Build Careers. [online] Available at: https://stackoverflow.com/ [Accessed 10 Dec. 2019].