Contribution-roles in community-based question answering sites: The case of Mathoverflow

Muffak, A. and Seguel P.

Summary

In this research, we explore the patterns in how users engage in knowledge conversations within a community-based question answering sites (CQAs) through generative-roles based on their contribution patterns. We operationalize this construct in contribution-roles ratios. Using on Mathoverflow data we identify the existence of contribution-roles based on the ways that users contribute to the platform. Based on a network perspective, we built a knowledge collaboration network that represents the flow of information through knowledge sharing practices. Three main contribution-interactions are analyzed: the answer to a question (A2Q), comment to a question (C2Q), comment to an answer (C2A). Based on the analysis of the largest component, and node centralities we build measures of contribution-related roles. We start by analyzing ratios between centralities of different types of relationships to identify different nodes. Then, we explore the possible uses of these metrics and their limitations to explain exchanges within and across communities. Ideas for future research are discussed.

Introduction

Innovations on information systems in the last decades have enabled individual's and organization's knowledge management processes (Faraj, Jarvenpaa, & Majchrzak, 2011; Kane, 2009; Lee, Rui, & Whinston, 2019). Innovations regarding knowledge sharing, in particular, has become an area of major focus within information system community and communication researchers. Researchers are interested in how technology-enabled actions might shape collaboration dynamics. The advances in internet-based technologies enable users to easily create, edit, evaluate, and/or link to content or to other creators of content (Lee et al., 2019; Majchrzak, Faraj, Kane, & Azad, 2013)

The proliferation of electronic networks of practice ENPs has gathered much attention from researchers and practitioners (Sharratt & Usoro, 2003; Wasko & Faraj, 2005). They are defined as self-organizing communities with potentially geographically distributed group of members that use a computer-mediated communication system. These ENPs usually "share some common practice or interests, and user participation is voluntary" (Lee et al., 2019, p. 579). According to Lee et al (2019), a particular ENP that is relevant for technical knowledge production are the community-based question answering sites (CQA). On these sites, users can post questions or provide answers to others' questions. Stack Exchange is a popular example of these platforms.

Platforms Stack Exchange rely on their design to facilitate knowledge exchange among core members. Moreover, it is through the participation and exchange of their members, that these platforms are able to stay in business. Even though much interest has been placed on understanding how individuals engage in ongoing knowledge conversation online (Lee et al., 2019; Majchrzak et al., 2013; Wasko & Faraj, 2005), less is understood regarding the influence of generative role-taking has over CQA dynamics. Generative role-taking is "engaging in the online knowledge conversation by enacting patterned actions and taking on community-sustaining roles in order to maintain a productive dialogue among participants" (Majchrzak et al., 2013, p. 45). Faraj, et al., (2011) used the term to describe non prescribed patterned actions that individuals take with the purpose of facilitating dialogue and exchange within a community. As social media and CQA

platforms enhance dialogue visibility, they facilitate generative-role taking. This research explores ways to identify generative-role taking within a CQA based on contribution patterns between members. We explore possible uses to analyze community dynamics within the CQA.

Goal and approach

We want to analyze patterns in how users engage in knowledge conversations within a community-based question answering sites (CQAs) through generative-roles based on their contribution patterns.

We use a network approach to analyze patterns of interactions among members in a CQA. The network perspective allows us to quantify and analyze the information exchange within the CQA platform. From this perspective, even if not all the actors participate in the same Q&A site, they gather experiences and responses from the interactions with members from the whole CQA. We focus on the interaction patterns that certain members display and how it might relate with their exchanges within communities on the CQA.

Context and Data source

This report focuses on the information exchange in a community-based question answering sites (CQAs) called Mathoverflow. We identify the existence of self emergent roles and groups based on the specific ways that users have to contribute to this platform. Mathoverflow is part of the StackExchange web, and it is defined as “a question and answer site for professional mathematicians” (“MathOverflow,” 2019). Same as sites from Stackexchange, like StackOverflow, many users access to useful technical knowledge based on the contributions and interactions from its members. Therefore, the public and private value that this type of site can provide to users on the web, depends heavily on the quality and frequent interactions of the community members of the site. On this site, we focus on the community of members that participate in knowledge sharing practices within Mathoverflow.

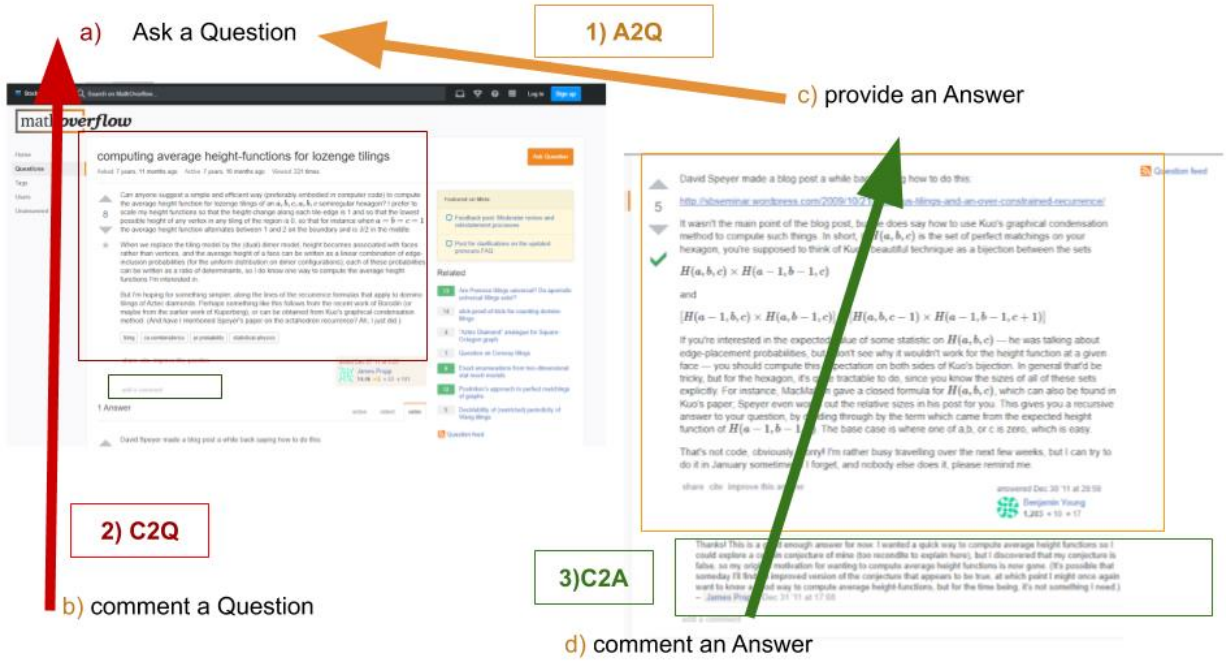


Figure 1. Knowledge practices and interactions in Mathoverflow

Figure 1 shows the core forms in which a user can engage in knowledge sharing conversations. A user can ask a question, provide a response, or comment on either an answer or a question. To engage in any of these practices, users need to create and login to a user account. Responses and comments could be given to any questions available on the website. It should be noticed that questions and answers can be rated and influence the overall ranking of users within the community. Therefore, answering or asking are actions that express a major level of engagement within the community.

From a network perspective, a knowledge collaboration network can be represented from these basic knowledge sharing practices. As can be seen in Figure 1, we can identify three major interactions: the answer to a question (A2Q), comment to a question (C2Q), comment to an answer (C2A).

To study the knowledge collaboration we utilize digital trace data from Mathoveflow. We used public data of interactions within Mathoveflow created by Paranjape, Benson, and Leskovek (2017) and available on the SNAP library ("SNAP: Network datasets: Math Overflow temporal network," n.d.).

The authors derived "*a temporal network by creating an edge (u, v, t) if, at time t, user u: (1) posts an answer to user v's question, (2) comments on user v's question, or (3) comments on user v's answer. We formed the temporal network from the entirety of Stack Overflow's history up to March 6, 2016*" (Paranjape et al., 2017, p. 607). The dataset was a list of edges with the information of the source, target, and the timestamp (in seconds). The data was separated into three datasets named a2q, c2q, c2a. Each node had an exclusive id number, but no identifier for questions or answers were provided. No information regarding content or socio demographics were available on the original dataset. We report network size information below on the results section.

Despite the limited information on the dataset, we decided to use it to explore the creation of contribution-roles based only on structural data. We believe that our key ideas and methods could be explored in a richer dataset. Furthermore, the size and sparsity of the data qualified as a Complex Network. This aspect included challenges that were important for the learning purposes of the authors.

We transformed this data using networkX as a directed, multilayer, weighted network. Some key transformations included:

- We analyzed the full network of interactions including all three types of interactions. We add each type of interaction as an attribute of the edges.
- We interpreted the data as a static network. Therefore, edges with the same source and target were summed into an edge's attribute called "weight".

Based on this context and dataset, we define these key objectives:

1. Identify roles within the community based on contributions regarding answer to question (a2q), comment to question (c2q), comment to answer (c2a).
2. Identify the presence of these roles within and across communities.

Methods and key measurements

All the project's code and the curated dataset are is available on this team GitHub repository: https://github.com/ppseguel/network_math

Analysis

We used networkX package from Python for all the analyses on this research. Our analyses included global, node and cluster descriptor analyses. We also created new node attributes based on structural data. First, we started by analyzing the connectedness of the overall network based on the three interaction types mentioned above. Then, we analyze different centralities measures based on node position on the network and combined these centralities to identify generative-roles based on their contribution. Also, we analyze clusters based on common modularity modules available in Python networkX package.

Contribution-role metrics

We created new metrics to identify roles within the community based on the flows of information of knowledge contribution practices. Our main interest was to identify recurrent patterns of contribution that could describe a generative-role. We focused on how an individual's participation or relevance in the network might balance different forms of participation or tend to specify one form of contribution.

Therefore, we created rations based on the centrality measures for specific types of interactions within the community. The key idea is that a balanced ratio would represent a balanced or general style of relevant interaction of an individual, while an imbalanced one would represent a more specific form contribution. By the distinction of balanced and unbalanced forms of relevant contribution, we can infer role taking practices within the community.

Our ratios are based on centrality measures of nodes. Centrality measures are node descriptor metrics that allow us to explore the relevance that a node might have in the overall exchange of information.

For this explorative study we focused on outdegree and indegree measures to identify relevant forms of contributions. First, outdegree centrality measures represent the relevance of the volume of contributions for a specific type of interaction that a node makes in relation to other nodes. We analyze the outdegree measures for all types of relationships A2Q, C2Q and C2A. Second, we used indegree centrality just for A2Q relationship. As we didn't have a direct information regarding the contribution behavior of individual in the platform related to posting question, we use the interaction that an individual receives to its question-related behavior as a proxy for this form of contribution.

It should be noticed, that degree centrality in networkx related measures considered the weight of edges, as well are normalized by occurrence. Therefore, they can be used for comparisons among different network types. Also, the measures aren't sensible to disconnected networks, as is based on the direct relations of a node with its neighbors. So, these measures have a theoretical and methodological fit for our purposes.

Our main formulas are the following:

$$a)\ Questioner - answerer\ ratio = \frac{indegree(A2Q) + 0.005}{outdegree\ (A2Q) + 0.005}$$

$$b)\ Answerer - commenter\ in\ questions\ ratio\ = \frac{outdegree(A2Q) + 0.005}{outdegree\ (C2Q) + 0.005}$$

$$c)\ Answerer - commenter\ in\ answers\ ratio = \frac{outdegree(A2Q)\ + 0.005}{outdegree\ (C2A)\ +\ 0.005}$$

We included a constant due to the possibility of having values with value 0.

Measure a) allows us to compare question and answer network behavior, which are the key elements of Mathoverflow. Measures b) and c) allow us to compare overall answerer and commenter behavior.

Our interpretation of this measures is the following:

- If the ratio is close to 1 that is interpreted as a balanced or general role in relation to the interactions of focus.
- If ratio is closer 20, then the role is defined by a specification in relation to the numerator.
- If ratio is closer 0.05, then the role is defined by a specification in relation to the denominator.

We focused our analysis on the 10 cases that presented most balanced and most unbalanced forms of relevant contribution for the three ratios of interest. Also, to ensure the comparison of relevant participation we included the following conditions:

- For the unbalanced, we consider cases that have at least one degree centrality measure above average.
- For the balanced, we consider cases that have both degree centrality measure above average.

These conditions allow us to remove nodes with very small centralities that can be interpreted as individuals who participate only once or very few times.

Online communities are usually characterized by a high number of individuals that contribute for a short period of time or in a low volume. Even though their participation and retention is important, sustainability of online communities is usually attributed to the retention of a core groups of few members (Kane, 2009). Therefore, our analysis of the 10 nodes with most relevant contribution ratios could apply for the studies that interest on the analysis of core members or contributors within an online community.

Analysis

Global descriptors

Table 1. Largest component

	Full	A2Q	C2Q	C2A
Largest_nodes	13,095 (52.77%)	12,554 (57.88%)	9,405 (55.86%)	12,837 (92.75%)
Largest_edges	205,519 (85.64%)	55,128 (54.40%)	70,669 (87.12%)	79,722 (88.10%)
Av. outdegree		0.0003498183473936052	0.0007990197707899023	0.0004838212462014734
Density	0.0011	-	-	-
Transitivity	0.0388	-	-	-

First we analyzed global properties of our network based on the full network types (combination of all types of interactions) and for each specific type. The network proved to be disconnected for checks available in networkX for strongly and weakly connected network. Therefore, we analyzed the largest component of the network, which presented 13,095 nodes and 205,519 static edges. This corresponds to more than half of nodes (52.77%) and most of the statics edges (85.64%). Furthermore, it presented a low level of density (0.0011) and transitivity (0.0388). Since the largest component presented a huge number of nodes and edges, as well as a low level of density among nodes, we described the network as representative of complex networks.

Also, we analyze the distribution of nodes and edges for each interaction type within the largest component. For the case of nodes and edges, the largest component includes more than half of the network size for all the interaction types. It should be noted that with respect to the number of edges, the percentages are above 87% for the interactions related to commenting on questions or answers. Therefore, we can expect that the act of commenting might play a major role in connecting nodes within the largest component.

Furthermore, in the case of nodes of C2A, their number represents 92.75% of the nodes of the full size network. Therefore, it might play a major role.
In conclusion, the largest component seems to be an adequate representation of the full size network for nodes and edges. In particular, it represents the C2A network type.

Community identification

We used the Louvain algorithm for best partition to identify communities within the largest component of the full network types. Based on this approach, we identified 16 communities within the largest component. We discard three communities with a number of nodes that were less than 0.5% of the largest component. Table 2 shows the values of nodes for each partition.

Table 2. Communities

Community	nº of nodes	% of largest component
0	1508	11.52%
1	1222	9.33%
2	945	7.22%
3	695	5.31%
4	2183	16.67%
5	1770	13.52%
6	100	0.76%
7	879	6.71%
8	711	5.43%
9	745	5.69%
10	1103	8.42%
11	1003	7.66%
12	226	1.73%
13	1	0.01%
14	2	0.02%
15	2	0.02%

Contribution-roles analysis and node selection

We report and analyze the results for the 10 nodes that presented more balanced and unbalanced ratios for each type of interaction pairs. We included the values of each centrality measure and their community. Moreover, we include appendices with graph distributions for ratios across communities.

We discuss the possibility to use the contribution-role metrics to analyze community dynamics. We explore these ideas for analysis within and across communities and a road map for future implementation is provided.

First, we explore the option to analyze contribution-roles within communities. Later, we explore the possibilities to find patterns across communities. As values of the metrics might vary across communities, but with systematic or specific distributions

We can observe that the top and bottom values help to identify a core group of members with a marked style of contribution. We observe than in some cases the unbalanced nodes seem to correspond to a specific community. For instance, table 3 shows that unbalanced roles (Questioner-answerer ratio), nodes that could be described as “questioners” tend to be from group 0 and 13. As making questions is one of the core drivers for knowledge exchange within CQAs, a deeper analysis to group 0 could be of interest for Mathoverflow. Furthermore, as Mathoverflow is a community of professional mathematicians, it should be noted strong formulation of math problems is one of the main drivers of the discipline. And people that build strong and interesting questions that attracts others to solve them, could require a high level of skill or even intuition. Questioners in these communities could represent individual with high status and ability.

Table 3. Unbalanced roles (a) Questioner-answerer ratio

	Nodes	in_A2Q/out_A2Q	in_A2Q	out_A2Q	Community
	812	3.0339	0.0126	0.0008	0
	416	2.9184	0.0108	0.0004	0
	4721	2.5096	0.0127	0.0021	1
	2672	2.4234	0.0092	0.0009	13
	10909	2.3916	0.0102	0.0014	13
	1459	2.3892	0.0210	0.0059	5
	942	2.3780	0.0076	0.0003	0
	1	2.2353	0.0126	0.0029	0
	1047	2.2217	0.0104	0.0019	0
Top 10	3621	2.1982	0.0067	0.0003	13
	13650	0.1687	0.0004	0.0270	9
	11260	0.1810	0.0002	0.0235	9
	6794	0.2094	0.0000	0.0189	13
	26935	0.2118	0.0002	0.0194	12
	25510	0.2150	0.0010	0.0227	8
	18060	0.2423	0.0018	0.0229	1
	11142	0.2544	0.0080	0.0460	9
	10076	0.2593	0.0002	0.0149	1
	7460	0.2610	0.0003	0.0154	1
Bottom 10	8008	0.2694	0.0009	0.0168	5

Table 4. Unbalanced roles (b) Answerer-commenter in questions ratio

	Nodes	out_A2Q/out_C2Q	out_A2Q	out_C2Q	community
	26935	2.4625	0.0194	0.0049	12
	21907	2.3082	0.0235	0.0012	12
	32389	2.3008	0.0270	0.0005	5
	1098	2.1890	0.0227	0.0009	0
	51	2.0066	0.0129	0.0014	0
	2968	1.9884	0.0177	0.0003	6
	12205	1.9552	0.0154	0.0005	12
	4600	1.9470	0.0124	0.0038	13
	1059	1.9373	0.0092	0.0002	0
Top 10	6129	1.9324	0.0077	0.0006	6
	763	0.0852	0.0029	0.0883	12
	290	0.1777	0.0124	0.0930	9
	14094	0.1892	0.0030	0.0374	9
	1465	0.1908	0.0069	0.0575	8
	1409	0.2072	0.0093	0.0641	4
	4177	0.2084	0.0054	0.0450	12
	2841	0.2391	0.0056	0.0392	1
	1353	0.2412	0.0016	0.0223	12
	1384	0.2452	0.0033	0.0290	12
Bottom 10	78	0.2470	0.0064	0.0410	6

Table 5. Unbalanced roles (c) Answerer-commenter in answers ratio

	Nodes	out_A2Q/out_C2A	out_A2Q	out_C2A	Community
	26935	3.8514	0.0194	0.0013	12
	11260	3.3203	0.0235	0.0036	9
	13650	2.7913	0.0270	0.0065	9
	25510	2.7743	0.0227	0.0050	8
	6153	2.4404	0.0129	0.0023	4
	20302	2.2541	0.0177	0.0051	12
	7460	2.2511	0.0154	0.0041	1
	12120	2.2329	0.0124	0.0028	12
	21907	2.0968	0.0092	0.0018	12
Top 10	8588	2.0632	0.0077	0.0012	6
	763	0.1821	0.0029	0.0386	10
	1353	0.2677	0.0016	0.0196	0
	2530	0.2769	0.0041	0.0277	7
	2926	0.3012	0.0091	0.0418	2
	2383	0.3084	0.0010	0.0146	11
	2841	0.3193	0.0056	0.0281	2
	5740	0.3319	0.0048	0.0245	0
	3546	0.3364	0.0028	0.0182	0
	4177	0.3445	0.0054	0.0252	2
Bottom 10	1409	0.3461	0.0093	0.0364	0

Sorting unbalanced cases by ratios allowed us to construct analytical threshold to analyzed the data. This threshold was based on the minimum value of the bottom 100 and top 100 unbalanced cases. For the Lower bound we have a threshold of 1.25 under which, we don't consider those nodes as imbalanced. So we are only looking at the nodes with ratios above 1.25. For the upper bound, we have a threshold at 0.5. Only nod with ratios smaller than 0.5 are considered unbalanced. The rest of nodes can count as balanced or nodes that don't participate.

Ratio variation across communities

Figures 1, 2 and 3 in the appendix show that some communities tend to be more balanced and others have individual with more unbalanced roles. For example communities 0 and 1 tend to be more unbalanced when it comes to answering questions. Individuals in those two communities take on the role of either asking or answering questions. Community 8 has ratios close to 1, meaning that the roles are either very balanced or the individuals in that community tend to participate very little. Some communities seem to have tendencies to unbalanced or balanced members Within all communities, members tend to be more balanced or participate very little. As we can see in the graphs almost all ratios are concentrated around the value of 1.
We can also see that some communities are specialized in some types of relationships. Community 12 has more members who comment than members who ask or answer questions; the same can be said about community 0 that has a more defined role at asking and answering rather than commenting.

Other potential approaches could be

1. Use node identified with unbalanced and balanced roles and compared them with connectors. We define nodes that are connectors, as the ones that have 1 or more connections to nodes of other communities.
2. Another alternative will be to analyze the heterogeneity coefficient based on community membership and contribution-role type.

Future research

We identify additional aspects that could be explored to increase the validity and value of the contribution-role metrics:

- The research requires more exploration regarding the contribution-roles that proved to be balanced, and the communities that have higher balanced ratios. Overall, our analysis shows that the community proves to have more balanced ratios, rather than unbalanced ones.
- We could analyze the network based on its temporal information. The analysis of the dynamic network of information exchange, could help to explore how roles and participation volume change over time. Furthermore, it would be possible to check for the stability and reliability of the contribution-role metrics across specific time episodes. Finally, by including time variables, we could explore the dissolution or emergence of communities and the relevance of contribution-roles on that matter.
- We could combine commenting practices (C2Q and C2A), as they could fit under the same theoretical construct. This transformation of the ratios could help to understand commenting interactions as a whole.
- We need to explore the effect on the ratios of other centrality measures. Based on centralities alone (eigenvector and katz centrality), no much difference was observed for the 20 nodes with highest values. However, we haven't explored yet if any changes would appear using the ratios.
- Finally, we could compare our data with other Stack Exchange sites. For instance, we find particularly interesting the comparison with Stack overflow. Our impression is that Stackoverflow works as a CQA, where experienced users tend to help less experienced ones. Therefore, we could expect less symmetric behavior than in a network of math professionals, which seems to have a symmetric or balanced form of interaction.

Limitations

Our explorative research found multiple limitations regarding the data. Some of them include:

- The lack of question identifiers or related information regarding questions.
- Lack of information regarding node history, reputation, popularity.
- No content information regarding contribution
- No interpretation of self-loops
- Quality checks of the data might be required

The availability of these information with a better crawler protocol, would increase the possibilities to explore the uses of contribution-role metrics and their power to explain CQAs dynamics.

Final comments

Our research explores the identification of generative role-taking practices, based on structural information of knowledge exchange in a CQA. We proposed the operationalization of a generative role as contribution-role, based on a2q, c2q and c2a relationships.

Our measurements help us to identify a set of nodes that have a marked style, as well as to explore the variations of ratios within and across communities. We speculate that future research could use our metrics to analyze the impact of how patterns of contribution behavior of some members could affect information flow. Furthermore, we believe that it would be possible to identify the emergence of communities based on the presence of unbalanced or balanced roles. This metrics have potential for future studies in knowledge sharing in online communities.

Our results prove to be particularly useful to analyze the role of specific group members within a complex network. Due to the big amount of nodes that are disconnected across specific types of interactions, and have an overall low level of contribution. As Online communities are usually characterized by a high number of individuals that contribute for

a short period of time or in a low volume. Even though their participation and retention is important, sustainability of online communities is usually attributed to the retention of a core groups of few members (Kane, 2009). Therefore, our analysis of the 10 nodes with most relevant contribution ratios could apply for the studies that interest on the analysis of core members or contributors within an online community.

Finally, from a learning experience point of view, working with real network data, with a huge size and sparse data, proved to be a challenging and interesting experience. Sparse networks tend to add complexities for interpretation, calculation and data wrangling, due to the possibility to disconnect members through analytical process. Furthermore, we were quite surprised that based on a relatively short data in terms of attributes, we were able to create a series of attributes based on weight, structure and combining different network ties.

References

Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science*, *22*(5), 1224–1239.

Kane, G. C. (2009). *It's a Network, Not an Encyclopedia: A Social Network Perspective on Wikipedia Collaboration.* https://doi.org/10.5465/ambpp.2009.44243222

Lee, S.-Y., Rui, H., & Whinston, A. B. (2019). Is Best Answer Really the Best Answer? The Politeness Bias. *MIS Quarterly*, *43*(2).

Majchrzak, A., Faraj, S., Kane, G. C., & Azad, B. (2013). The contradictory influence of social media affordances on online communal knowledge sharing. *Journal of Computer-Mediated Communication*, *19*(1), 38–55.

MathOverflow. (2019). Retrieved November 15, 2019, from MathOverflow website: https://mathoverflow.net/

Paranjape, A., Benson, A. R., & Leskovec, J. (2017). Motifs in temporal networks. *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 601–610.

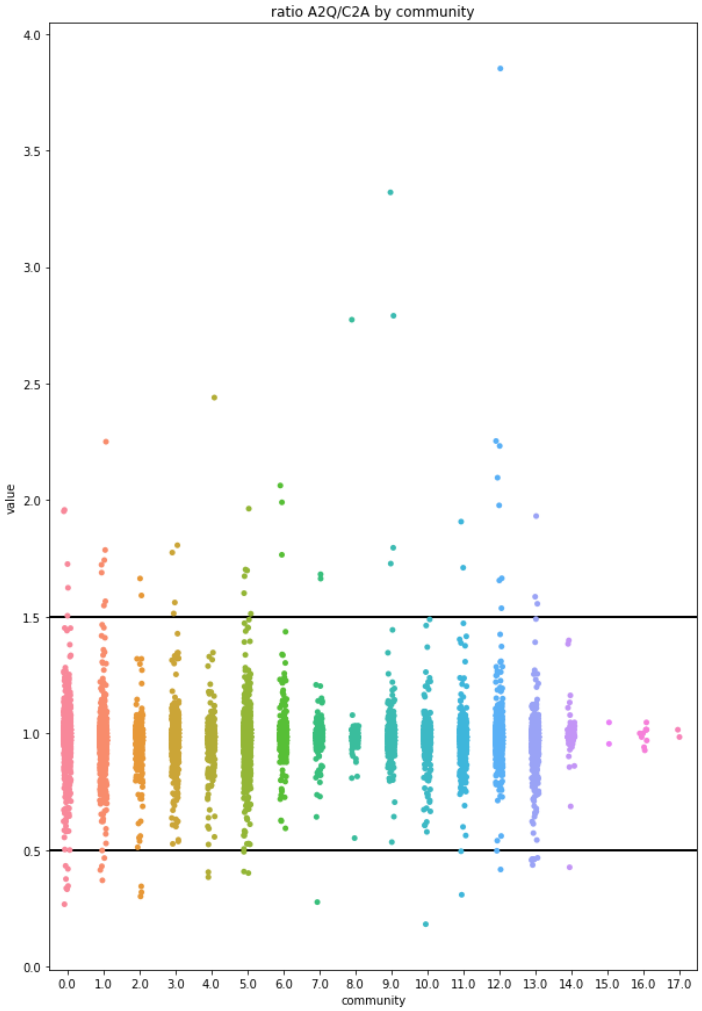
ACM.

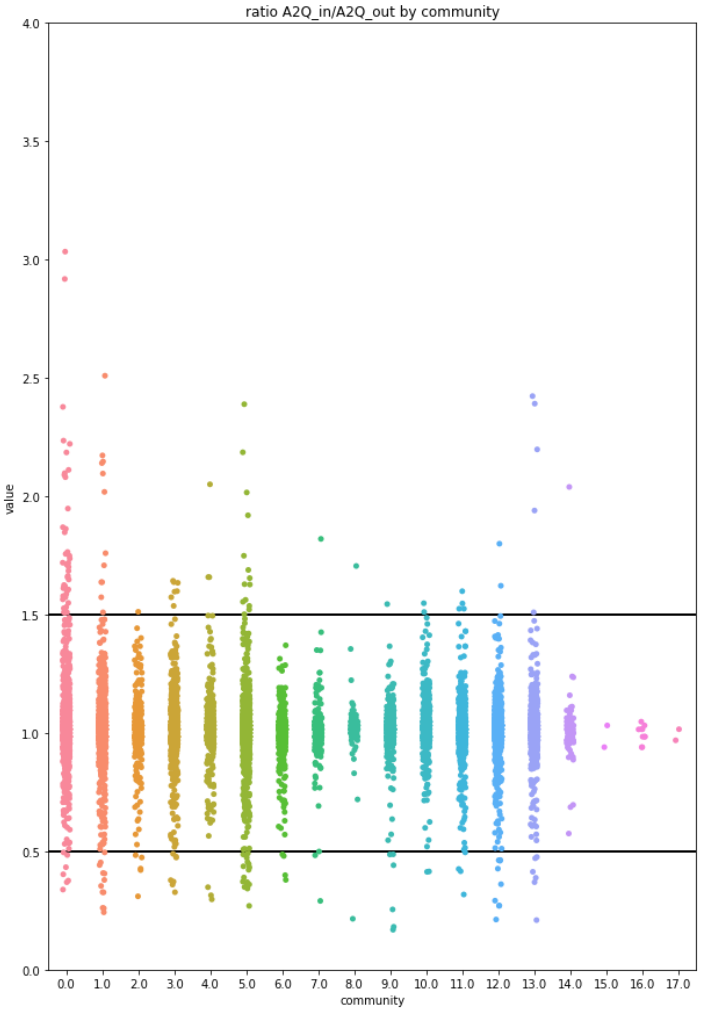
Sharratt, M., & Usoro, A. (2003). Understanding knowledge-sharing in online communities of practice. *Electronic Journal on Knowledge Management*, *1*(2), 187–196.

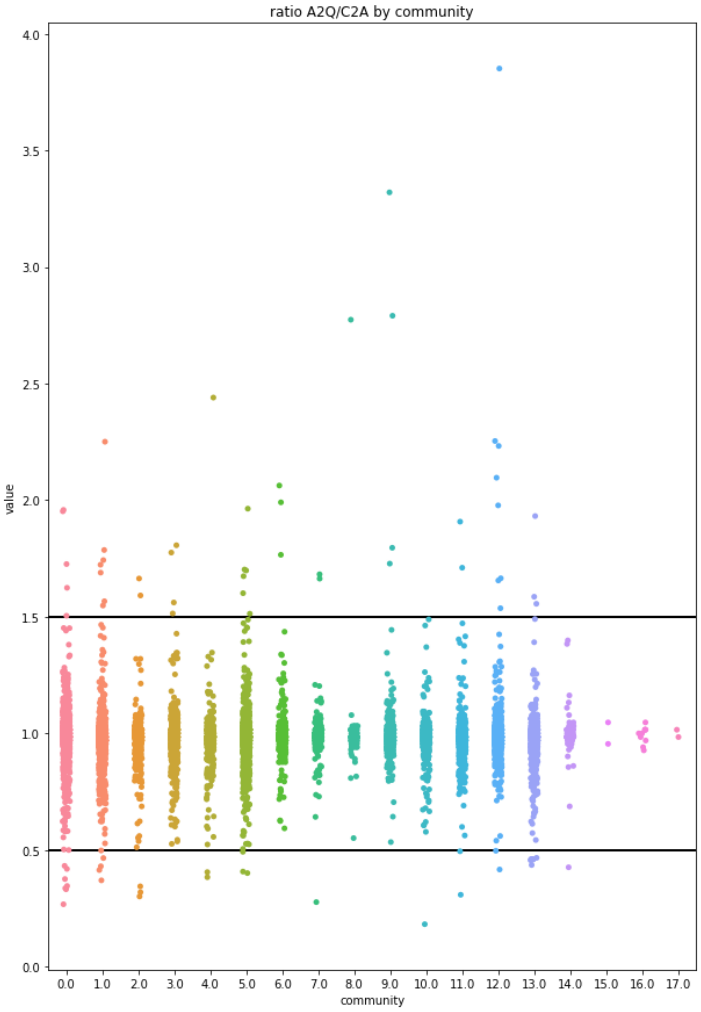
SNAP: Network datasets: Math Overflow temporal network. (n.d.). Retrieved November 15, 2019, from http://snap.stanford.edu/data/sx-mathoverflow.html

Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly*, 35–57.

Appendix




ratio A2Q_in/A2Q_out by community
value
community
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0


ratio A2Q/C2A by community
value
community
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0