Chapter 1

The role of trust in knowledge based communities

Information and communications technologies (ICTs) have enabled faster and easier creation and sharing of knowledge. Furthermore, they have provided access to a large amount of data which enabled a detailed study of their emergence and evolution [1], as well as user's roles [2], patterns of their activity [3, 4, 5]. However, relatively small attention was given to sustainability of SE communities. Most of the research was focused on the activity and factors that influence the increase of the users' activity in these communities. Factors such as need for experts and the quality of their contributions have been thoroughly investigated [6]. It was shown that growth of communities and mechanisms that drive it may depend on the topic around which the community was created [7].

The **Stack Exchange** is a network of question-answer websites on diverse topics. In the beginning, the focus was on computer programming questions with StackOverflow [1] community. Its popularity led to the creation of the Stack Exchange network that these days counts more than 100 communities on different topics. The SE communities are self-moderating, and the questions and answers can be voted, allowing users to earn Stack Exchange reputation and privileges on the site.

The new site topics are proposed through site Area51 [2], and if the community finds them relevant, they are created. Every proposed StackExchange site needs interested users to commit to the community and contribute by posting questions, answers and comments. After a successful private beta phase site reaches the public beta phase, other members are allowed to join the community. The site can be in the public beta phase for a long time until it meets specific SE evaluation criteria for graduation. Otherwise, it may be closed with a decline in users' activity.

We focused analysis on four pairs of SE communities with the same topic. Astronomy, Literature and Economics are active communities [3] The first time, these communities were unsuccessful and thus closed. We also compare closed Theoretical Physics with the Physics site, considering that those two topics engage similar type of users.

[1] More information about StackOverlflow is available at: https://stackoverflow.co/ and broad introduction to StackExchange network is available at: https://stackexchange.com/tour.

[2] Visit https://area51.stackexchange.com/faq for more details about closed and beta SE communities and the review process.

[3] Astronomy, Literature and Economics graduated on December 2021 and during our research, they were still in the public beta phase.

1.1 Network properties of Stack Exchange data

On Stack Exchange sites, the interaction between users happens through posts. As we are interested in examining the characteristics of the users, we map interaction data to the networks. Using complex network theory, we can quantify the properties of obtained networks and compare different SE communities, e.g. alive and closed SE sites.

In the user interaction network, the link between two nodes, user i and j, exists if user i answers or comments on the question posted by user j, or user i comments on the answer posted by user j. The created network is undirected and unweighted, meaning that we do not consider multiply interactions between users or the direction of the interaction.

First approach is to aggregate all interactions in the first 180 days, and study the properties of static network. Many local and global network measures are dependent [8], and it was shown that degree distribution, degree-degree correlations and clustering coefficient are sufficient for description of the properties of complex networks [9].

We calculate the **degree distribution**, figure 1.1, and compare the distributions of active and closed communities of the same topic. Degree distributions between active and closed communities follow similar lines.

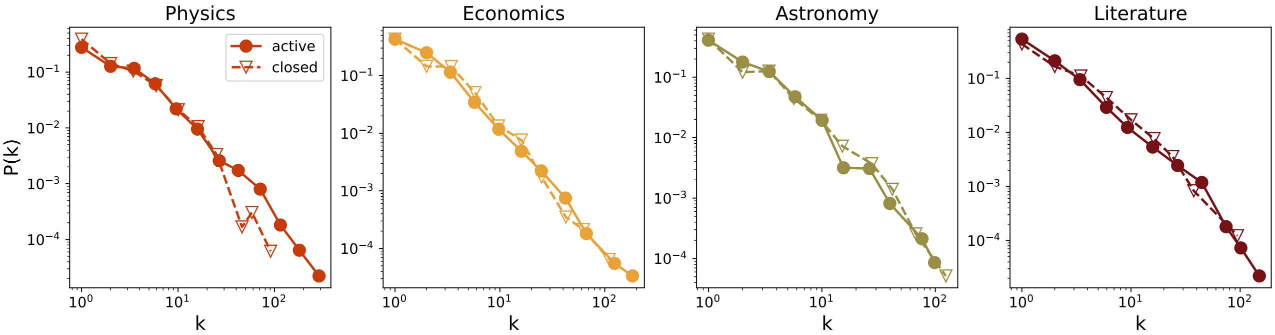

Figure 1.1: Degree distribution.

If we take look into **neighbor degree** dependece on the node degree $k_{nn}(k)$, figure 1.2 we find that there are structural differences between networks formed in the active and closed communities. On average k-degree users in active communities have neighbors with larger degree than it is case in closed communities. The results are consistent for Physics, Economics and Literature. For Astronomy we find different behavior, where the $k_{nn}(k)$ distributions of closed communities are on the top of distributions of the active one.

The **clustering coefficient** of a node quantifies the average connectivity of between its neighbours and cohesion of its neighborhood [8]. It is a probability that two neighbours of a node are also neighbours, and is calculated using the following formula:

$$c_i = \frac{e_i}{\frac{1}{2}k_i(k_1 - 1)} \ . \tag{1.1}$$

Here e_i is the number of links between neighbours of the node i in a network, while $\frac{1}{2}k_i(k_i - 1)$ is the maximal possible number of links determined by the node degree k_i. The clustering coefficient of network C is the value of clustering averaged over all nodes. Study on dynamics of social group growth shows that that links between one's friends that are members of a social group increase the probability that that individual will join the social group [10]. Furthermore, successful social diffusion

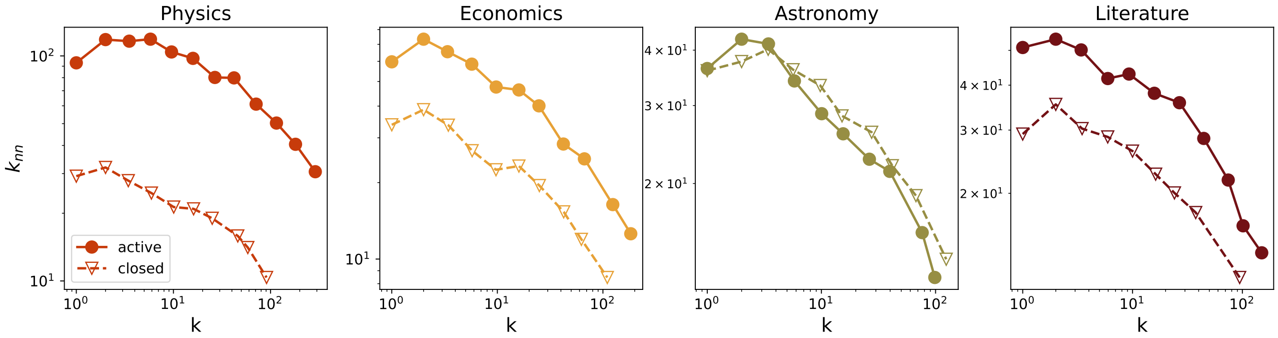

Figure 1.2: Neighbour degree.

typically occur in networks with high value of clustering coefficient [11]. These results suggest that high local cohesion should be a characteristic of sustainable communities. The dependence of the clustering coefficient on the node degree is shown on figure 1.3. As expected we find that active communities are more clustered.

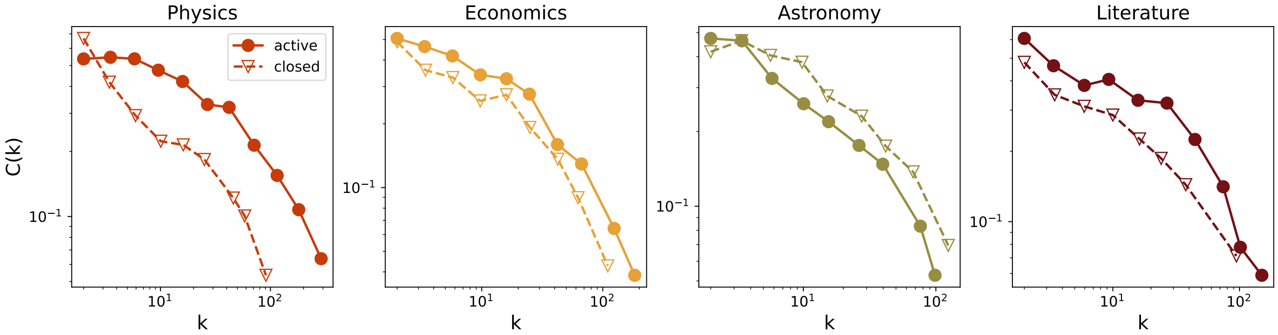

Figure 1.3: Clustering coefficient.

1.1.1 Properties of evolving complex networks

Instead of creating a static network from the data in the first 180 days of community life, we study how network snapshots evolve. At each time step t, we create network snapshot $G(t, t+\tau)$, for time window of the length τ. We fix the time window to $\tau = 30$ days and slide it by $t = 1$ day through time. Discussion of how the length of the sliding window influences the results is given in appendix A. Sliding the time window by one day, we can capture changes in the network structure daily, as two 30 days consecutive networks overlap significantly.

Here we investigate how clustering coefficient in a SE community is changing with time by calculating its value for all network snapshots. We compare the behavior of clustering for active and closed communities on the same topic in order to better understand how cohesion of these communities is changing over time. Figure 1.4 shows the evolution of mean clustering coefficient for all eight communities. All communities that are still alive are clustered, with the value of mean clustering coefficient higher than 0.1. Physics, the only launched community, has the value of clustering coefficient above 0.2 for the first 180 days.

During larger part of the observed period, the clustering coefficient of an active community is higher compared to the clustering coefficient of its closed pair. If we compare active communities

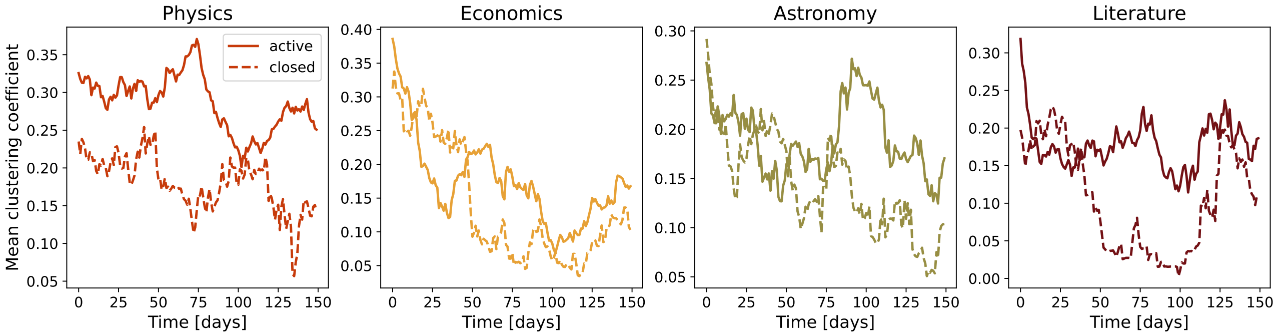

Figure 1.4: Mean clustering coefficient.

with their closed counterpart, the closed communities have higher value of the mean clustering coefficient in the early phase while later communities that are still active have higher values of clustering coefficient. These results suggest that all communities have relatively high local cohesiveness, and that lower values of clustering coefficient in the later phase of community life may be an indicator of its decline.

1.2 Core-periphery structure

Previous research on Stack Exchange communities have attempted to explain how different types of users interact. In Question-Answer communities are expected popular and casual users [3, 7]. Popular users generate the majority of interactions in the system, they are experts in community and take care on answering questions and engage the discussions through comments. As popular users they considered the 10% of the most active users, and showed that popular users are highly connected not only among themselves but also with casual users.

We tested this theory on all eight communities. We focused on 30 days sub-networks and showed how the number of links per node among popular users and between popular and casual users, evolve over time, figure 1.5. We also compare active and closed communities of the same topic, so links per nodes in active sites are larger than in closed communities.

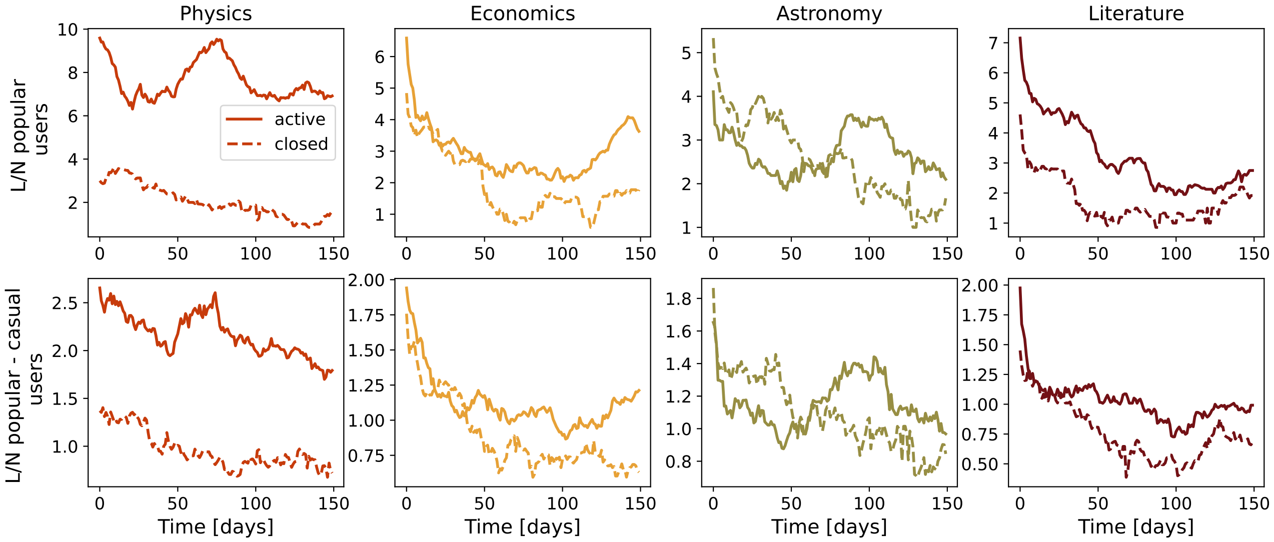

Figure 1.5: Links per node among popular users (top 10% of users) and between popular and casual users (everyone but popular users).

Although, we find the difference between active and closed communities, the split according to 10% most active users does not guaranty that all popular users will be considered. Furthermore, the smaller group of frequently active users is similar to the core users in core-periphery structure. This is why we are going to detect the core of the each 30day network. By this, separation is based on the network structure, and more consistent, as using algorithmic approach are optimized the connectivity inside the core, periphery and among them. Core-periphery structure has core that is densely connected group of nodes, while the periphery has low density [12, 13].

We use Stochastic Block Model (SBM) to infer the core-periphery structure of each 30 days network snapshot and analyses how core structure evolve over time. The SBM algorithm is adapted for inferring the core-periphery structure, [13]. For each 30 days network we run the sample of 50 iterations and choose the modes parameters according to minimum description length. As stochastic models start from the random configuration, they can converge to different states. This is why we analyzed the stability of the inferred structures. More details are given in the appendix. We found that obtained structures differ, but minimum description length does not fluctuate too much. Also, different similarity measures between infered core configurations take values higher than 0.9, indicating that core structure is stable.

1.2.1 Jaccard index

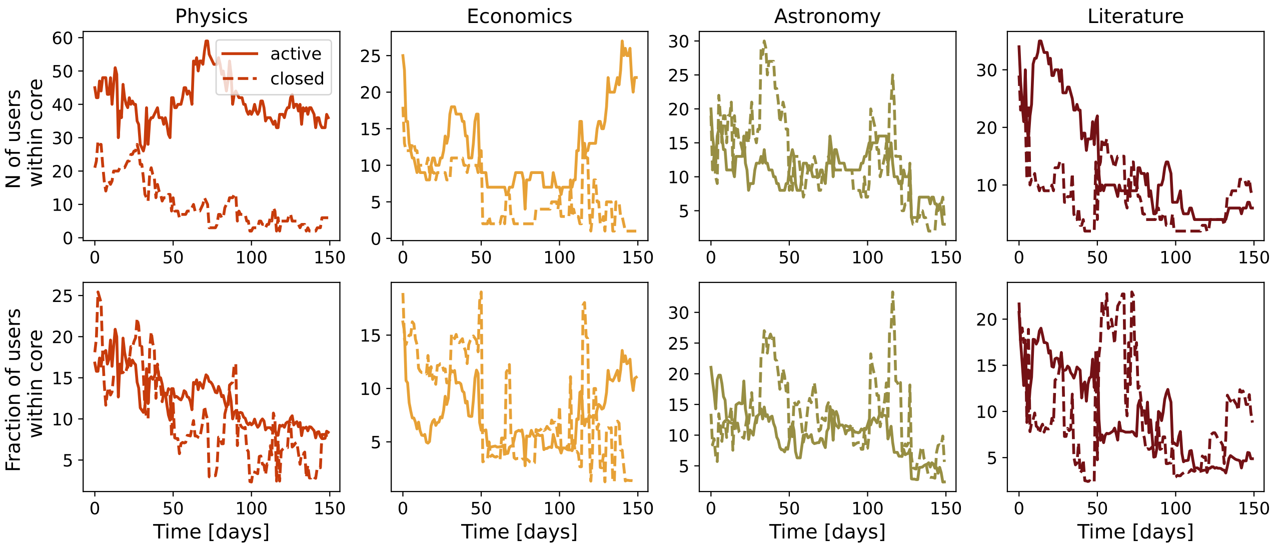

Figure 1.6: Just for reference size of the core (top) and fraction of users in core (bottom). Solid lines - active sites; dashed lines - closed sites.

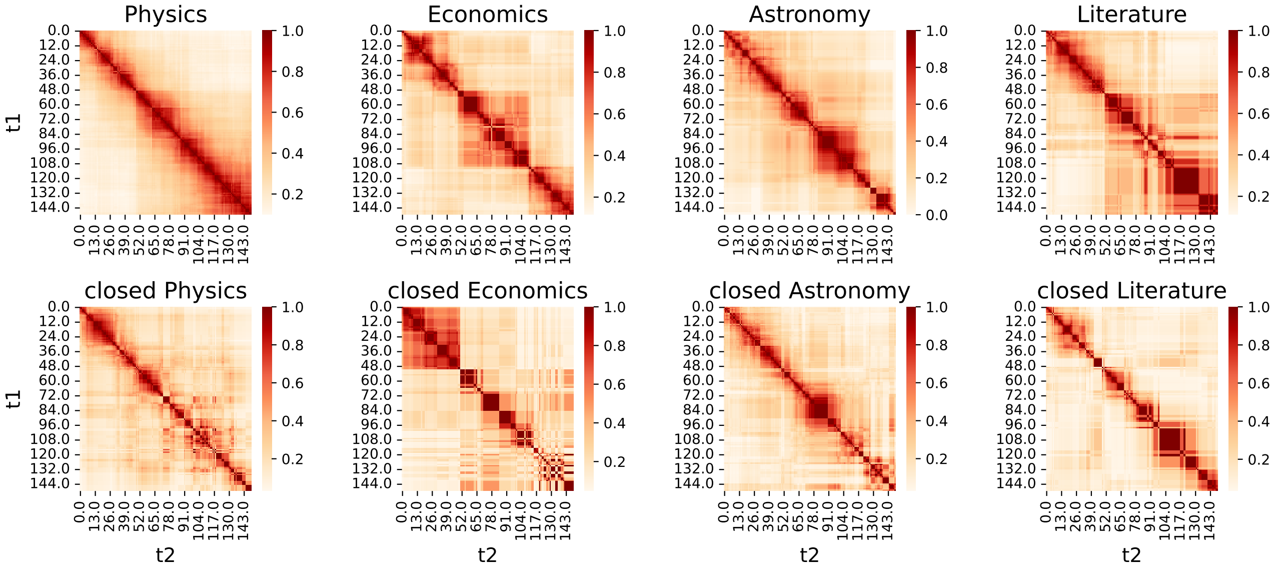

Figure 1.7: Jaccard index between core users in sub-networks at time points $t1$ and $t2$

1.2.2 core-periphery density

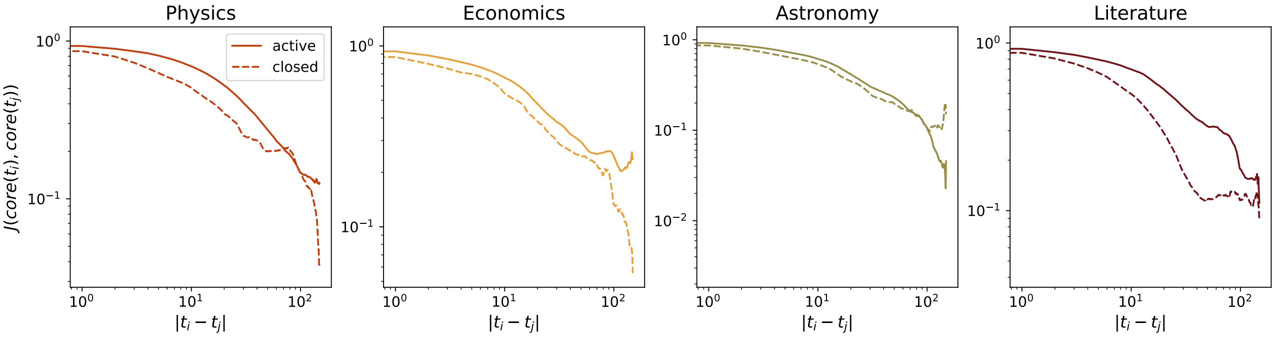

Figure 1.8: Jaccard index between core users in 30days sub-networks for all possible pairs of 30 days sub-networks separated by time interval $|t_i - t_j|$

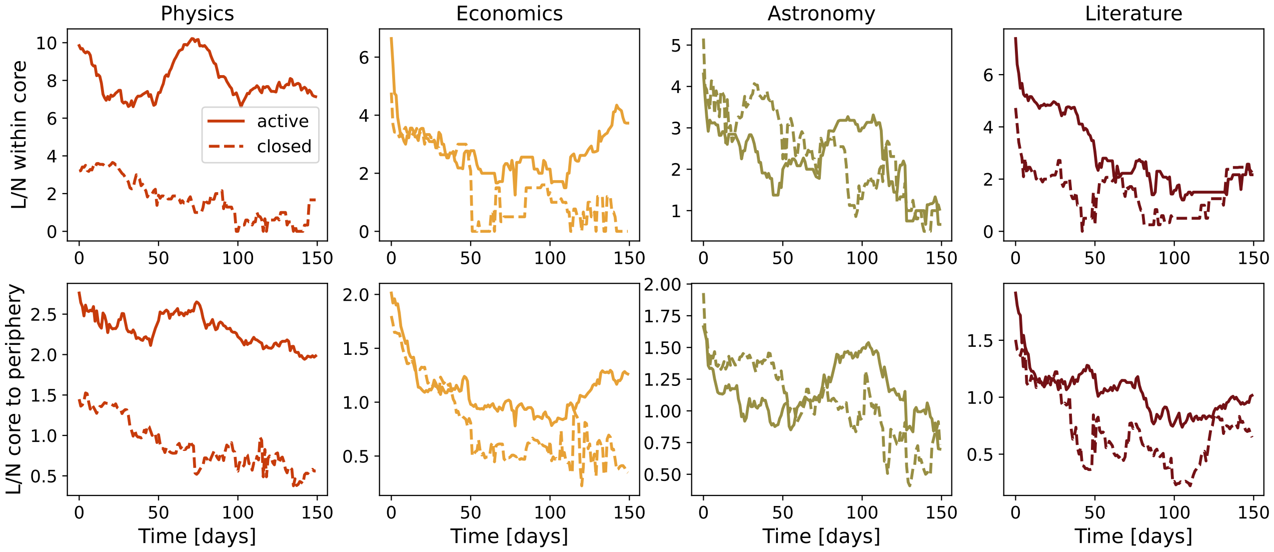

Figure 1.9: Links per node in core and links per node between core and periphery.

Bibliography

[1] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Scientific reports*, 5(1):1–10, 2015.

[2] Akrati Saxena and Harita Reddy. Users roles identification on online crowdsourced q&a platforms and encyclopedias: a survey. *Journal of Computational Social Science*, pages 1–33, 2021.

[3] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Activity archetypes in question-and-answer (q8a) websites—a study of 50 stack exchange instances. *ACM Transactions on Social Computing*, 2(1):1–23, 2019.

[4] Rogier Slag, Mike de Waard, and Alberto Bacchelli. One-day flies on stackoverflow-why the vast majority of stackoverflow users only posts once. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pages 458–461. IEEE, 2015.

[5] Anamika Chhabra and S RS Iyengar. Activity-selection behavior of users in stackexchange websites. In *Companion Proceedings of the Web Conference 2020*, pages 105–106, 2020.

[6] Himel Dev, Chase Geigle, Qingtao Hu, Jiahui Zheng, and Hari Sundaram. The size conundrum: Why online knowledge markets can fail at scale. In *Proceedings of the 2018 World Wide Web Conference*, pages 65–75, 2018.

[7] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Self-and cross-excitation in stack exchange question & answer communities. In *The World Wide Web Conference*, pages 1634–1645, 2019.

[8] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[9] Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltán Toroczkai, Marián Boguná, Guido Caldarelli, et al. Quantifying randomness in real networks. *Nature communications*, 6(1):1–10, 2015.

[10] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.

[11] Damon Centola, Víctor M Eguíluz, and Michael W Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, 2007.

[12] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[13] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *arXiv preprint arXiv:2005.10191*, 2020.