Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

Social achievement and centrality in MathOverflow

Leydi Viviana Montoya, Athen Ma and Raúl J. Mondragón

Abstract This paper presents an academic web community, MathOverflow, as a network. Social network analysis is used to examine the interactions among users over a period of two and a half years. We describe relevant aspects associated with its behaviour as a result of the dynamics arisen from users participation and contribution, such as the existence of clusters, rich–club and collaborative properties within the network. We examine, in particular, the relationship between the social achievements obtained by users and node centrality derived from interactions through posting questions, answers and comments. Our study shows that the two aspects have a strong direct correlation; and active participation in the forum seems to be the most effective way to gain social recognition.

1 Introduction

Discussion communities in the format of a Question–Answer (Q&A) site, such as *Yahoo! Answers*[1] have becoming increasingly popular; and the interactions among users and the structure and dynamics of the resultant complex network present a variety of interesting research questions. For example, Rodrigues *et al.* investigated topic management by studying how users choose a topic and their 'tagging' be-

Leydi Viviana Montoya
Queen Mary University of London, Mile End Road, London E1 4NS, e-mail: lvmc3@eecs.qmul.ac.uk

Athen Ma
Queen Mary University of London, Mile End Road, London E1 4NS, e-mail: athen.ma@eecs.qmul.ac.uk

Raúl J. Mondragón
Name, Queen Mary University of London, Mile End Road, London E1 4NS, e-mail: r.j.mondragon@eecs.qmul.ac.uk

[1] http://uk.answers.yahoo.com

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

haviour [1]. Users' participation is of great interest in terms of what motivates users to be active and successful online portals often provide features that enable users to socialise, discuss, chat as well as transfer-share knowledge [2][3]. In addition, Burel *et al.* analysed how these forums operate and attempted to identify the key methods that can lead to assess of quality answers by referring to multiple or specific topics and their complexity [4].

While the benefits of a Q&A forum is evident, particularly when it comes to knowledge exchange and information flow, the reason why people are motivated to help others through these forums is still a bit of a puzzle. For example, MathOverflow[2] is an academic/research community which comprises of members who post high level questions and answers about mathematics. Tausczik *et al.* analysed the relationship between the reputation of users and the quality of their contribution, and examined how the *perceived* reputation might affect the *perceived* quality of the contribution [3]. Tausczik and Pennebaker [5] also studied the motivations for a user to contribute to a Q&A community, they suggested that building reputation is an important incentive for users to participate in the forum. In addition they also noticed that, it is important to find ways to ensure that users contribution are of good quality, and social assessment such as voting seems to be a common way to achieve this.

In this paper, to better understand how users interact with each, we represent the interactions among users in MathOverflow as a network. We find that centrality measures of an individual user give strong indications of social achievement with respect to their social status within the community and quality of their post.

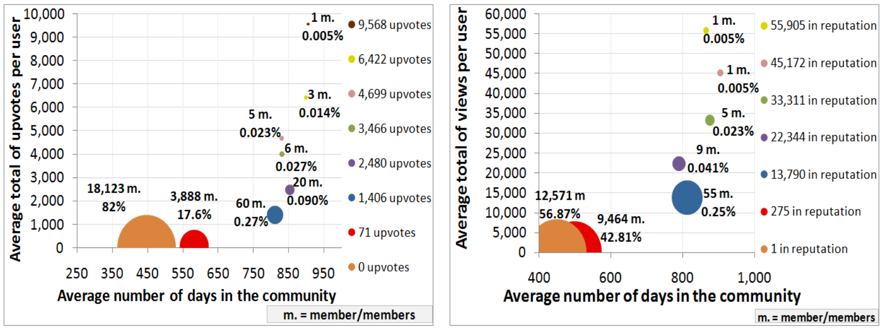
2 MathOverflow

MathOverflow was created by a group of graduates and post-doctoral students from the University of California Berkley in September 2009 [6] and has a total of 22,107 registered users (at the time in which the data was collected in April 2012 [7]). It supports a specialist mathematics Q&A forum, in which posts corresponds to Questions, Answers and Comments. This paper studies all the posts created between September 2010 and April 2012. Each member is provided with 5 social features: *views, upvotes, downvotes, reputation* and *badges*. The *views* attribute describes the number of times the profile of a member and their posts has been viewed through the forum. An *upvote* is a positive score to one‘s post; and similarly, see fig. 1(a); a *downvote* is a negative score towards one‘s contribution. *Reputation* is associated with a user‘s performance, see fig. 1(b). For instance, a user can gain reputation by accumulating *upvotes* and carrying out community and maintenance tasks for the forum, such as editing questions and deleting spam. *Badges* represent the knowledge, expertise, interest or participation of a community member. For example, the users who posted a question with more than 10,000 views are awarded a *Famous question*

[2] http://MathOverflow.net

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

badge. In addition, badges are divided into Gold, Silver and Bronze to reflect the level of achievement.



(a) (b)

Fig. 1 (a) Upvotes per member between 30-Sept-2009 to 01-April-2012. (b) Reputation score per member between 30-Sept-2009 to 01-April-2012

2.1 *MathOverflow as a complex network*

This paper studies the interactions among users generated through posting questions, answers and comments. When a question is posted on the forum by a user, other users can answer the question; in addition, they can also comment on a question or an answer. During the period of study, there were a total of 28,360 questions and 78.7% of these questions have at least one answer. The average of answers posted per question was 2.51. The average numbers of questions, answers and comments posted per day is 30.92, 61.09 and 201.45 respectively, describing a significant difference on the volume of comments posted compared to the volume of answers and questions. There are a total of 184,727 comments, 47% were posted on a question and 53% were posted on an answer.

We represent the former interaction as (Q-A) for question-answer interactions and the latter as (QA-C) for question-comment and answer-comment interactions. The interaction among MathOverflow users are described by a graph $\mathscr{G}(\mathscr{E},\mathscr{V})$ which consist of a set of edges $\mathscr{E} = \{n_1,\ldots,n_N\}$ and a set of nodes $\mathscr{V} = \{v_1,\ldots,v_L\}$. The former consists of the users who have taken part in posting and the latter represents a (Q-A) or (QA-C) between two individual users. There are 11,743 nodes and 96,616 unique edges in the MathOverflow network where duplicate edges have been removed.

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

3 Network analysis

3.1 Centrality

We first examine the centrality of the MathOverflow network by referring to the *degree*, *betweenness*, *closeness* and *eigenvector* centralities [8].

The degree centrality is based on the number of edges k_i that node n_i has. Nodes with high degree are considered more important in the overall network's structure. The relative importance of the nodes is assessed using the degree distribution. In this network the degree distribution looks like a power law with an exponent $\alpha = -1.2$, see Fig. 2(a), which suggest that the network is scale-free. This value is similar to previous studies on co–authorship networks among physicists and computer scientists which is between 0.91 and 1.3 [9]. It has been suggested that if $\alpha < 2$ the network is dominated by few high degree nodes. In the MathOverflow network, most members have less than ten interactions but a small number of users (0.0085%) have interacted with over a thousand members.

The betweenness centrality measures the fraction of geodesic paths that pass through node n_i. In MathOvervlow, the betweenness is low among all the nodes, and the highest (normalised) betweenness is 0.05. We observed a direct correlation between a user's degree and betweenness, as illustrated in Fig. 2(b).

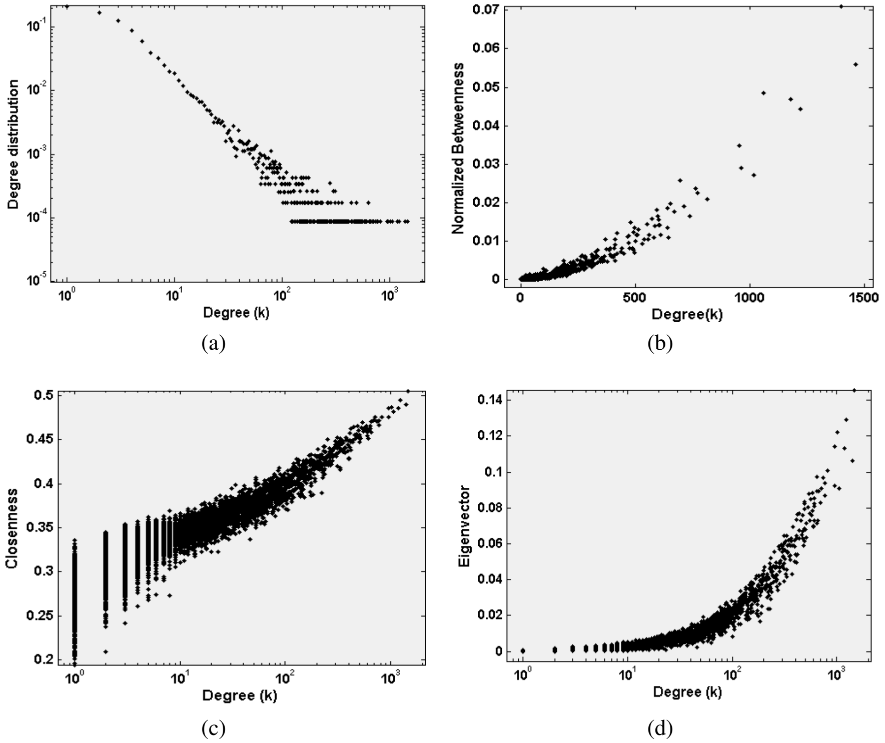
The closeness centrality of a node n_i is the mean geodesic distance from it to every other node. This centrality provides a crude measure on how quickly information spreads. In this network, nodes have a closeness in a range between 0.19 and 0.51; and similar to the betweenness centrality, it is highly related to degree centrality, see fig. 2(c).

The eigenvector centrality measures how well connected a node is and how much direct influence it may have over other well connected nodes in the network. While the degree centrality provides a simple count of the number of connections that node has, the eigenvector centrality assumes that not all connections are equal as the connections to nodes which are themselves influential (with a high degree) will result in a higher score. The eigenvector centrality is the eigenvector corresponding to the largest eigenvalue obtained from the diagonalisation of the adjacency matrix. About 96.5% of the community members have an eigenvector of 0.02 or less, which implies little influence originated from these nodes. However, a small group of 26 users (out of 11,743) have a eigenvector over 0.1, see fig. 2(d).

3.2 Global network characteristics

The radius of a network refers to the minimum eccentricity of any node and the diameter is the maximum geodesic distance between two nodes in the network [10]; the MathOverflow network has a radius and a diameter of 4 and 7 respectively. The network density refers to the number of existing edges in the network divided by

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38



(a) (b)

(c) (d)

Fig. 2 (a) Degree distribution of nodes in the MathOverflow network. (b) Betweenness of individual nodes against their degree. (c) Closeness of individual nodes against their degree. (d) Eigenvector centrality of individual nodes against their degree.

the total possible number of edges [11], our network has a density of 0.014 which means that there is still a huge potential for the members to interact with other users.

The assortativity of a network is given by

$$r = \frac{\langle kk' \rangle - \langle k \rangle \langle k' \rangle}{\langle k^2 \rangle - \langle k \rangle^2} \tag{1}$$

where the angle brackets $\langle \ldots \rangle$ mean the average over all links and k and k' are the degree of the end nodes of a link. If $r = 0$, there is no correlation between the nodes. If $r > 0$, the network is assortative and disassortative otherwise. The network is disassortative as r has a value of -0.21, and this indicates that high centrality users tend to interact with low centrality users and the presence of a mixing pattern. Network disassortativity is typically found in biological networks and computing and information networks [12].

Transitivity in a social network is associated with the presence of a heightened number of triads (three nodes fully connected) in the network [10]. The transitivity

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

of the network is relatively low with a value of 0.09 so the chance of finding closed triads in the network is quite small.

The Global efficiency of a network is defined as [13]:

$$E_{\text{glob}}(\mathscr{G}) = \frac{1}{N(N-1)} \sum_{i \neq j \in \mathscr{G}} \frac{1}{d_{ij}} \tag{2}$$

where d_{ij} is the geodesic distance from node i to node j and $0 \leqslant E_{glob}(\mathscr{G}) \leqslant 1$. E_{glob} can be seen as a measure of how efficiently information is exchanged over a network, given that all nodes are communicating with all other nodes concurrently [13]. The global efficiency of the network is 0.33.

3.3 Rich–club

Figure 2 strongly suggest that the nodes of high degree are the most important in the network. To study the interactions between this high degree nodes we use the notion of *rich–club* to examine how the interactions are distributed in this community. High degree nodes that are also highly interconnected to each other are referred as rich nodes [14]. The rich–club coefficient is used to characterise the density of connections between the rich nodes and is given by:

$$\Phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)} \tag{3}$$

where $E_{>k}$ corresponds to the number of edges among the $N_{>k}$ nodes having a degree higher than a given value k [15]. $\Phi(k)$ represents the ratio of the real number to the maximally possible number of edges linking the $N_{>k}$ nodes [16].

We considered the top 0.5% of the nodes (58 nodes) with highest degree to define a subgraph of rich nodes, where $k_s = 409$ is the lowest degree in the subgraph. Network measures such as assortativity, clustering coefficient, rich–club coefficient and average shortest path were calculated for the new subgraph.

The subgraph has a high clustering coefficient of 0.8 which means that users who have interacted with a same user tended to make ties (interact) between them. In addition, the subgraph has a negative assortativity value which means that high degree users tend to interact with lower degree users. This could be due to the fact that the rich nodes have a wide range of degrees (between 409 to 1,462). Zhou *et al.* state that the rich–club coefficient $\Phi(k)$ is the ratio of the total actual number of links to the maximum possible number of links between members of the rich-club, and a coefficient of 1 or close to 1 means that the members within the club form a fully connected network [16]. This implies that the top 0.5% of the users with highest degree in the network demonstrated to be highly connected. Nevertheless, Zhang et.al [17] suggest that k_{max}/k_s is a convenient index in complex networks with any degree distribution to show the proportion of links (or degrees) the rich nodes possess in comparison with the rest of nodes in a network. In this paper, k_{max}/k_s

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

= 3.5745 which means rich nodes were far closer connected among them than the majority of the nodes, forming a rich–club.

Zhou *et al.* recommend to compare the clustering coefficient, assortativity and average path length for the rich–club to the same parameters in the network as way to confirm the existence of the rich–club. If these values were quite different then that confirms the rich–club has a different behaviour to that of the whole network [17]. The existence of a rich–club in the network was confirmed as the rich–club has different behaviour than the rest of the network, because its clustering coefficient is much closer to 1 and its geodesic path is small. Both, rich nodes and the whole network have a negative assortativity which is associated to the wide range of existing degree levels inside the network (from 1 to 1,462).

4 Users attributes and network measures

As mentioned previously, MathOverflow has incorporated a range of social features that encourage participation among users. We examine a number of social achievements, namely, reputation, total number of views and number of upvotes; and see how one's social achievements tie in with the centrality measures. Firstly, we define a set of hypotheses with reference to social achievements and centrality. Secondly, the Spearman's rank correlation is used to test the validity of the proposed hypotheses.

Hypothesis 1 – A user's reputation score is closely related to his/her degree centrality. This hypothesis assumes that a user's reputation score is closely related to the number of interactions of the user. On one hand, for a MathOverflow member, reputation is defined as "how much the community trust them?" [18] and a user would need to participate and be active in the network to earn reputation. On the other hand, degree centrality can be seen as a measure of the activity of an actor [19]. As a result, a positive correlation is expected as a user who participates more would have a higher degree and reputation value.

Hypothesis 2 – The total number of views obtained by a user is related to his/her eigenvector centrality. The hypothesis takes into consideration that the measure of eigenvector centrality reflects one's influence in the network and assumes that the higher one's influence, the more likely to attract other users to view your post.

Hypothesis 3 – The number of upvotes obtained by a user is related to his/her closeness centrality. Finally, this refers to one's closeness to other users as an indication of their level of expertise as knowledgeable users are likely to attract followers. Similarly, the number of upvotes obtained by a user shows that such users have been providing reliable or useful posts. A positive correlation is expected as a nodes with a significant volume of upvotes should be more centralised in the network as well as closer to the different kind of users, as the community is comprised of users

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

who would like to learn and share knowledge, it is reasonable to assume that the best voted users are likely to linked to the others.

The Spearman's rank correlation is a statistical measure used to test the direction and strength of the relationship between two variables [20]. The null hypothesis for this kind of test is defined as "there is no relationship between the two sets of data" [21]. The Spearman's rank correlation provides a value r_s which falls between -1 and +1 and a $p-$value to test its statistical significance level [22]. First of all, the data of both variables (x_i and y_i) are ranked separately. Secondly, using the ranked values x_i and y_i, the r_s coefficient is calculated as:

$$r_s = \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_i (x_i - \langle x \rangle)^2 \sum_i (y_i - \langle y \rangle)^2}} \tag{4}$$

when tied ranks existed on the data [23]. Once r_s is calculated, the existing correlation between the variables can be established depending on which range the absolute value of r_s lies on [20]: *very weak* (0.00 - 0.19); *weak* (0.20 - 0.39); *moderate* (0.40 - 0.59); *strong* (0.60 - 0.79) and *very strong* (0.80 - 1.0). Finally, a significance test is done by evaluating the $p-$value from the student's t distribution, which test the significance of r_s [24]. Table 1 shows the results for the three hypotheses.

Hypothesis 1: Reputation values vs. Degree centrality. The Spearman's rank correlation for the two variables is 0.709. The $p-$value is less than the significance level of the null hypothesis (i.e. $p < 0.05$) which therefore can be rejected. As a result, there is a strong correlation between the reputation score and the degree centrality.

Hypothesis 2: Eigenvector values vs. Total number of views. These variables also showed a strong correlation, and again, the null hypothesis was rejected. This strong correlation means the total number of views is a good indicator of how much the user has been active and participating with popular people in the network. Also, it is important to point out that the eigenvector centrality correlates better with the total number of views than the total number of upvotes, as the latter gives a moderate correlation with $r_s = 0.53$. We can conclude that users with higher eigenvector centrality have more visits to their profile than other users, but they might not have more upvotes than users with lower eigenvector values.

Hypothesis 3: Upvotes vs. Closeness centrality. In contrast, the upvotes and closeness variables are weakly correlated. The null hypothesis that there is no correlation between the variables, was tested at a significance level of 5%. As the p-value is less than the significance level, the null hypothesis was rejected, confirming the weak correlation between Upvotes and Closeness centrality. This result means the users with more popularity in the network for their knowledgable contributions are not necessarily more centralised or close to the general users.

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

Table 1 Spearman's rank correlation results for the three hypothesis (Significance level of 5%)

Hypothesis	r_s coefficient	Result
1	0.70	Strong correlation
2	0.62	Strong correlation
3	0.28	Weak correlation

5 Behaviours on subgroups

Specific subgroups in the network were identified based on the tags (topics) the users marked their questions with. When analysing the 3 subgroups with highest number of users (AG–Algebraic Geometry, SQ–Soft Question and NT–Number Theory) they appear to have similar network characteristics as the whole network. These subgroups contain 20%, 19% and 18% of the total users in the network respectively. The subgroups AG and NT have a diameter larger than the original network, which means the farthest two nodes in AG and NT are more distant; whereas the SQ subgroup has the same diameter as the original network (see Tab. 2).

Table 2 Network parameters for the whole network and the subgroups AG, SQ and NT.

Parameter	All network	Subgroup AG	Subgroup SQ	Subgroup NT
Total nodes	11,743	2,310	2,215	2,132
Total unique edges	96,616	12,903	12,392	20,628
Network diameter	7	9	7	8
Network radius	4	5	4	4
Network density	0.0014	0.005	0.005	0.005
Assortativity	-0.2156	-0.1604	-0.1328	-0.1662
Transitivity	0.0955	0.228	0.161	0.235

Identifying clusters in a network was performed in order to identify subgroups with particular behaviour or with nodes which share similar characteristics. The purpose of cluster analysis is to divide the data into groups that are meaningful, useful or both, based only on the information found in the data that describes the objects and their relationships [25]. Given that the network has less than 500,000 nodes the cluster algorithm used is the Clauset-Newman-Moore (CNM) which is based on greedy maximising the modularity function Q [26]. The quality function Q of a network division, defined as

$$Q = \sum_i (e_{ij} - a_i^2) \tag{5}$$

where e_{ij} is the fraction of edges in the network that connect vertices in group i to those in group j, and a_i is the fraction of edges that fall within communities, minus the expected value if edges fall at random [27]. A total of 105 groups were

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

identified when running the algorithm. Most of the clusters are quite small given that 61% of the clusters has just 2 users, see fig. 3(a). The map representing the 105 clusters in the network displayed its 6 biggest groups (denominated as G1, G2, . . . , G6) and the small box in the bottom right corner contains 99 clusters with a total of 297 nodes. Also, the 6 clusters have 97.5% of the total users in the MathOverflow network. The subgraphs (clusters) diameter, average geodesic distance and density were calculated for the 6 clusters, in order to characterise them. Cluster G2 has the same diameter than the whole network (7 edges). Clusters G6 and G3 have a diameter of 11. The six clusters show a density lower than 0.006 which is bigger than the density of the whole network (0.0014), although they are significantly low given that the density provides the ratio of direct ties in the network to the total of possible direct ties [28]. The 6 biggest clusters have a small proportion of users with the highest degree values and the highest upvotes counts, with a relevant proportion of users with low degree value and upvotes as followers. The users with a degree higher than 500 or upvotes higher than 750 or with the highest reputations are located only on the 6 biggest clusters (see fig. 3(b)). In contrast, users who have achieved a reputation of 1 (which is given when users input their personal details on the community web page) are located in the biggest clusters as followers of the main users and in the small clusters, Fig. 3(c).

The largest clusters seem to have a similar properties as the whole network. However, when looking at the users attributes in each cluster, there are no specific similitudes between the users in a cluster; the clustering algorithm does not provide clusters which represent a clear criteria of the community division, for example the topic of the questions.

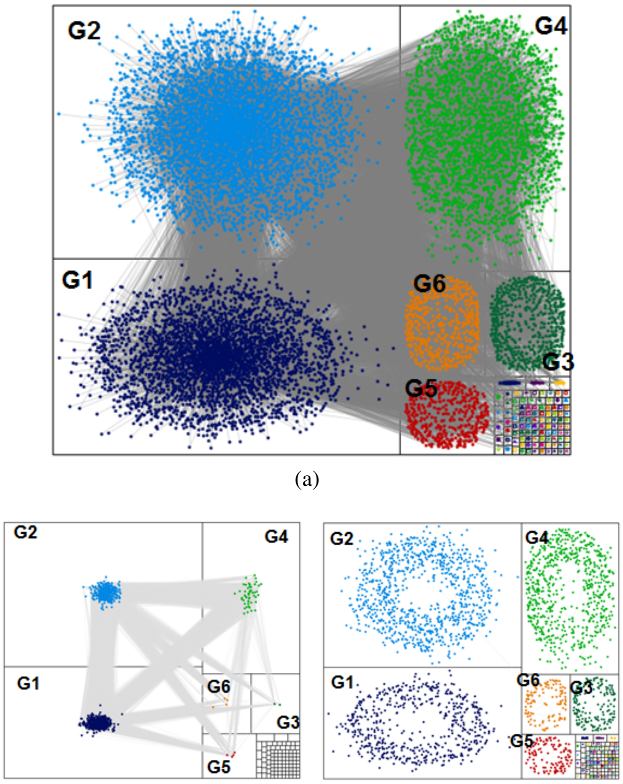
6 Conclusions

Social Network Analysis was used to establish the structure of the academic web community MathOverflow, and to identify main characteristics among its users. The network has a high number of nodes and links and its degree distribution is approximated with a power law, which suggest that the network is scale–free and interactions are dominated by a few high centrality users.

The network is found to be disassortative reflecting that general users are often advised by a *reputable* user who has achieved a certain social status in the network. The network exhibits self-similarity in the sense that when subgroups are defined using individual topics or by clustering using modularity, these subgroups show similar characteristics as the original network.

We study three different hypotheses on the relationship between centrality and social achievements. Spearman's rank correlation on these hypotheses show that a user's reputation, the number of views and the number of upvotes are related to their centrality. It suggests that allowing users to gain social recognition through these features is an effective way to encourage users' participation.

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38



(a)

(b)

(c)

Fig. 3 (a) Communities derived from the MathOverflow network using modularity and they are mapped using NodeXL [29]. (b) Users with a reputation score of 800 or over and their distribution across the different communities. (c) Users with a reputation score of 1 and their location.

[redacted]

References

1. Rodrigues, E., Milic-Frayling, N., Fortuna, B.: Social tagging behaviour in community-driven question answering. In: Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. Volume 1., IEEE (2008) 112–119
2. Mendes Rodrigues, E., Milic-Frayling, N.: Socializing or knowledge sharing?: characterizing social intent in community question answering. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM (2009) 1127–1136
3. Tausczik, Y., Pennebaker, J.: Predicting the perceived quality of online mathematics contributions from users' reputations. In: Proceedings of the 2011 annual conference on Human factors in computing systems, ACM (2011) 1885–1888

Complex Networks IV
Studies in Computational Intelligence Volume 476, 2013, pp 27-38

4. Burel, G., He, Y., Alani, H.: Automatic identification of best answers in online enquiry communities. The Semantic Web: Research and Applications (2012) 514–529
5. Tausczik, Y., Pennebaker, J.: Participation in an online mathematics community: differentiating motivations to add. In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM (2012) 207–216
6. Joel, S.: Cultural anthropology of stack exchange. Hacker News London Meetup - Events (June 2012) Available on http://vimeo.com/37309773.
7. MathOverflow: Dumps files: 2012-04-01 (April 2012) Available on http://dumps.mathoverflow.net.
8. Newman, M.: The mathematics of networks. Electronic Article (2005) Available on http://www-personal.umich.edu/ mejn/papers/palgrave.pdf.
9. Newman, M.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences **98**(2) (2001) 404–409
10. Newman, M.: The structure and function of complex networks. SIAM review **45**(2) (2003) 167–256
11. Faust, K.: Comparing social networks: size, density, and local structure. Metodološki zvezki **3**(2) (2006) 185–216
12. Pastor-Satorras, R., Vázquez, A., Vespignani, A.: Dynamical and correlation properties of the internet. Physical review letters **87**(25) (2001) 258701
13. Latora, V., Marchiori, M.: Efficient behavior of small-world networks. Physical Review Letters **87**(19) (2001) 198701
14. Zhou, S., Mondragón, R.J.: The rich-club phenomenon in the internet topology. IEEE Communications Letters **8**(3) (March 2004) 180–182
15. Colizza, V., Flammini, A., Serrano, M., Vespignani, A.: Detecting rich-club ordering in complex networks. Nature physics **2**(2) (2006) 110–115
16. Jiang, Z., Zhou, W.: Statistical significance of the rich-club phenomenon in complex networks. New Journal of Physics **10**(4) (2008) 043002
17. Xu, X., Zhang, J., Small, M.: Rich-club connectivity dominates assortativity and transitivity of complex networks. Physical Review E **82**(4) (2010) 046117
18. MathOverflow: Mathoverflow - frequently asked questions. Forum web page (July 2012) Available on http://mathoverflow.net/faq.
19. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Volume 8. Cambridge university press (1994)
20. Owen, A., Petrie, M., Palipana, A., Green, D., Croft, T., Jones, A., Joiner, S.: Spearman's correlation. Loughborough University (2012) Available on http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf.
21. s.n.: Ib geography notes. Web page (2012) Available on http://www.angelfire.com/ga2/ibgeography/.
22. Hossain, L., Wu, A.: Communications network centrality correlates to organisational coordination. International Journal of Project Management **27**(8) (2009) 795–811
23. Lund, A., Lund, M.: Spearman's rank-order correlation. Web page (2012) Available on https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.
24. Zar, J.: Significance testing of the spearman rank correlation coefficient. Journal of the American Statistical Association **67**(339) (1972) 578–580
25. Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson Addison Wesley (2006)
26. Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6) (2004) 066111
27. Newman, M.: Fast algorithm for detecting community structure in networks. Physical Review E **69**(6) (2004) 066133
28. Xu, G., Zhang, Y., Li, L.: Web Mining and Social Networking: Techniques and Applications. Volume 6. Springer (2010)
29. Foundation, S.M.R.: Nodexl excel template. Computer Program (2012) Available on http://www.smrfoundation.org/nodexl/.