COMPGW01: COMPLEX NETWORKS AND WEB - FINAL PROJECT
The Data Science Community Network: A Stack Exchange Study Case on Experts Users and Trending Topics

Santiago Gonzalez Toral
UPI: HSGON81
Web Science and Big Data Analytics
Computer Science Department
University College London
email: hernan.toral.15@ucl.ac.uk

Abstract

During the last five years Stack Exchange Q&A site has experienced a increased popularity providing an online gamified platform for users to exchange knowledge about a wide variety of topics and where users compete to earn the best reputation on the site. The goal of this report is to analyze a use case through a network analysis approach, of a growing and vibrant community such as Data Science to provide an insight on professionals, students and field enthusiasts about the top experts and trending topics that currently evolving. Furthermore, the report also provide suggestions about which properties should be the best for the analysis of user's expertise across a network.

Content list

1 Introduction

Data Science is an interdisciplinary field which its main objective it to extract knowledge or insights from data in different structure and every day people are getting more interest. This trend produces the creation of different working groups around the world as well as different online communities. Special attention has been given to the Data Science Stack Exchange[1], which is a question and answer (Q&A) site for Data science professionals, Machine Learning specialists, and those interested in learning more about the field.

In this paper, and as a data science enthusiast, I propose the study case of this community network as a way to answer some of the questions that as a Web Science and Big Data Analytics postgraduate student have. To get involve in the industry it is important to have a good picture of who and where are the top experts around the world. Where they are the industry tend to evolve and propose novel solutions. Moreover, its also relevant to know which are the trending topics around a community. As a result, the following research questions where formulated:

Question 1: Who are the top expert/learner users in the Data Science community?

Question 2: Where are the top users contributing to the Data Science community across the world?

Question 3: What are the top Data Science topics covered by the community?

Question 4: What are the top development tools covered by the community?

To start with the study case about the Stack Exchange Data Science community and its experts and trend topics, an initial exploratory analysis was performed using the T-SQL service provided by the platform. After that, the extraction and generation of the network graph-based representation was performed by using the *Pentaho data-integration*[2] *tool*. The construction of our graph relied on two types of data: users and the tags they had been active in. Also, it could be easily integrated with a REST service to apply geocoding on the user's location property to be able to perform geo-localization analysis on the resulting network. Additionally, a combination of different statistical and graphical tools where employed for network analysis. Gephy and Matlab was useful for visualization-based analysis and the computation of network properties, and a SNAP python script were also applied for linking analysis.

The social graph between users giving the correct answer and question user owners outlined a degree distribution where a few users gives most of the correct answers to posts. The finding propose that top experts and learners can be obtained in a network by analysing the Hubs and Authority nodes, and where the obtained top 10 experts match with the ones published by the statistical community page. Furthermore, the geolocation analysis in a world map provide some good thoughts of the distribution of data scientists across the world. Moreover, by evaluating the users and tags relation network, data scientists, data science students and learners in general could acquire some knowledge about the community trending topics and a list of the most used

[1] http://datascience.stackexchange.com/

[2] http://community.pentaho.com/projects/data-integration/

frameworks in the development of solutions. It would be useful for them, when looking for additional information and support.

The study case finally showed that the reputation score a user gain through his participation activity on the Stack Exchange platform is not a good value for community analysis. This is related with the fact that participants usually compete to answer questions as quickly as possible and to earn more reputation in the network through activities not related with their expertise on the field. The paper finally concludes that linking analysis approaches generate fair better results on the discovery of experts and trending topics.

2 Background

2.1 Stack Exchange

Stack Exchange (created as Stack Overflow in 2008) is a question and answer (Q&A) social network website composed by different subject classified communities that cover a wide range of topics. It provides a high level frontend application for content collaboration wherein user interactions are initiated by a user asking a question and the flow continues when another user answers the question, and may extend further through the exchange of insightful comments.

Through the *gamification* of the platform [10], users compete to gain reputation points by providing high-ranked answers, asking questions, participating in discussions and other rules defined by the Stack Exchange community [9]. The awarded points represent their activity on the entire site (every community the user signed in) which in turn it's the key core of the site to unlock new features for them, such as being able to down-vote answers.

In 2009, the Stack Exchange network made its data freely available, including user names, locations, and artifact IDs. Additionally, the platform provides an online Data Explorer[3] where users can exploit the up-to-date data available on the network by typing T-SQL queries.

Data Science Stack Exchange[4] is a question and answer site (or subcommunity) for Data science professionals, Machine Learning specialists, and those interested in learning more about the field. Until the beginning of 2016 the beta version of the community has been online for over 600 days, and according to *Stack Exchange Area51*[5] the site has 9970 users, 76% of answers solved and users visit it 965 days per day in average.

Inside the community and in terms of the data structure, two kinds of users are identifiable. *Knowledge seekers,* or users who formulate a question in the community, are connected to the root post question, whereas *Knowledge providers*, or users who give the correct answer to a question, are commonly attached to the post accepted as the best to resolve the issue. Moreover, (question) posts contain relevant information and metrics of the interaction of this users through the network (such as scores, up-votes, answers count, interaction event dates), as well as connectivity relation with (correct) answers, comments, and tags.

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

[4] http://datascience.stackexchange.com/

[5] http://area51.stackexchange.com/proposals/55053/data-science

2.2 Foundations of Network Analysis

Social Networks and/or Q&A sites are networks in which vertices are people (or groups of people), and edges represent the interaction between them as a friendship [6]. A scientific analysis over this subfield can be done by applying a set of mathematical models and/or algorithms for graph evaluation, which help us to understand the why and how of the structure of a social community site. This kind of networks can be formalized in the form of adjacency matrices representing interactions between users, or links between tags. One can then apply link analysis methods to find the more influential nodes [3].

The *(In/Out) node's degree* and its distribution are first order properties often used to analyse the node's connectivity in a network where the latter is very useful as a initial step to evaluate and predict structure, i.e., a network with a small number of users generating large amounts of content follows a power-law distribution. These properties are useful when we want to know the nodes with the largest amount of neighbours(friends).

On the other hand, a set of second order properties are available to consider the connectivity between nodes. Centrality measures are important when the position of an actor in a network is the main part of the aims of a study. *Betweenness centrality* take into account bridges nodes. It measures how often a node appears on a shortest path between a pair of nodes in the network. But sometimes is not so important for a node to have many direct friend or be well positioned between others. *Closeness centrality* measures the effectiveness of a node to spread information to all other nodes. In brief, the use of centrality measures to analyse analyse the connectivity of nodes depends on a project objectives and/or how a network topology is built.

Network Science also offers another set of tools that are useful to analyse the possible clusters and communities inside a network. Modularity is very important for community detection. The iterative model proposed by Newman [12] evaluates the betweenness of edges to measure how well a network is partitioned into communities. Additionally, the K-Core decomposition is suitable to visualize large scale networks. Basically, its implementation prune the least connected nodes and is helpful to show the hierarchical structure of a network.

Link analysis approaches are very important to compute the importance of a node in a network. In fact, the effectiveness of algorithms like Page Rank and HITS (Hubs and Authorities) is valuable when trying to find hubs and other types of influential nodes. Both are solution to the same problem but assessed in different ways. The former considers the links pointing to a target node whereas the latter performs an evaluation that depends on the links that comes from a source node.

3 Related Work

Every day an increasing number of users are interacting with Q&A websites like Stack Exchange to either share their knowledge with others or trying to solve a problem they have. Also, due to the wide range of topics and communities Stack Exchange offers, and the fact that the its datasets become public in 2009 [6]more people are getting involved on studying its structure, getting the full attention of networks analysts who constantly makes some research over its data with an special emphasis on expert user evaluation.

[6] https://archive.org/details/stackexchange

Vasilescu et.al. [8] evaluates a case study of the data analysis software R user collaboration network between the official mailing list and the Stack Exchange community, to show the changes in behaviour of contributors to use more a gamified environment. The analysis was based on three main questions: 1)can they find evidence of mailing list popularity?, 2)what is the difference between contributors in both single communities?, and 3)can they find difference in activity between active users in both communities?. To answer this research questions, the built a longitudinal dataset combining both data sources where participants overlapped and were identifiable. Authors found that user supports activities and experts migrated to Stack Exchange R community due to the incentive of gathering public reputation. Experts presented more activity in the Q&A community and provided faster answers.

Movshovitz-Attias [5] analysed the reputation system and user contributions of the popular Stack Overflow software development community. He used Page Rank and Singular Value Decomposition (SVD) as indicators of user expertise to highlight their importance in detecting anomalous users. His findings stated that high reputation users are the primary source of answers, and especially of high quality answers. He also found that on average a high reputation user asks more questions than a user with low reputation. The final results obtained were useful to develop an application that evaluates the user contributions over a month to predict who will become an influential contributor in the long-term.

On the other hand, other network research studies go beyond the reputation and score values assigned to a user. In [4], MacLeod performed an exploratory analysis of the data to determine a possible correlation between user reputation scores and the diversity of tags a user contributed to. The conclusions are positive, stating that there is a correlation between those measures and thus high reputation experts contributes to a wide range of topics. The analysis also showed the poor community structures in the network studied. “This is consistent with the prevailing literature on Stack Exchange use and the power law”.

4 Methodology

4.1 Dataset Extraction

To start with the study case about the Stack Exchange Data Science community and its experts and trend topics, an initial exploratory analysis was performed using the T-SQL service provided by the platform. Table 1 shows some basic statistics about the community. Due to the size of the network and the fact that it is still in a beta launch stage, a sampling process was not considered necessary in benefit of having a full picture of what is going on in the emerging data science site. As a consequence, an up-to-date data dump snapshot of the model could be extracted as CSV files by using the platform Data Explorer.

1937	question
2827	answers
76%	answered
9970	users
964	visitors/day

Table 1: Statistics of the Data Science community

After analysing the documented platform data schema [7], the entity tables Posts and Users were considered most important for the aims of our network evaluation.

4.2 User-Friendly Graph Generation Approach

As a way to help non-expert users to preview, extract and generate a graph-based network from relevant information from Stack Exchange public data dumps, an extensible data integration flow was built using *Pentaho data-integration*[7] *tool*. Figure 1 shows the built workspace to extract and transform Stack Exchange data into a network (set of nodes and relationship edges). The construction of our graph relied on two types of data: users and the tags they had been active in. By taking advantage of the pluggable architecture and the ease for a user to select the relevant data he wants to analyse through the interfaces, we could easily integrate a REST service using Gisgraphy[8] to apply geocoding on the users location property to be able to perform geo-localization analysis on the resulting network.

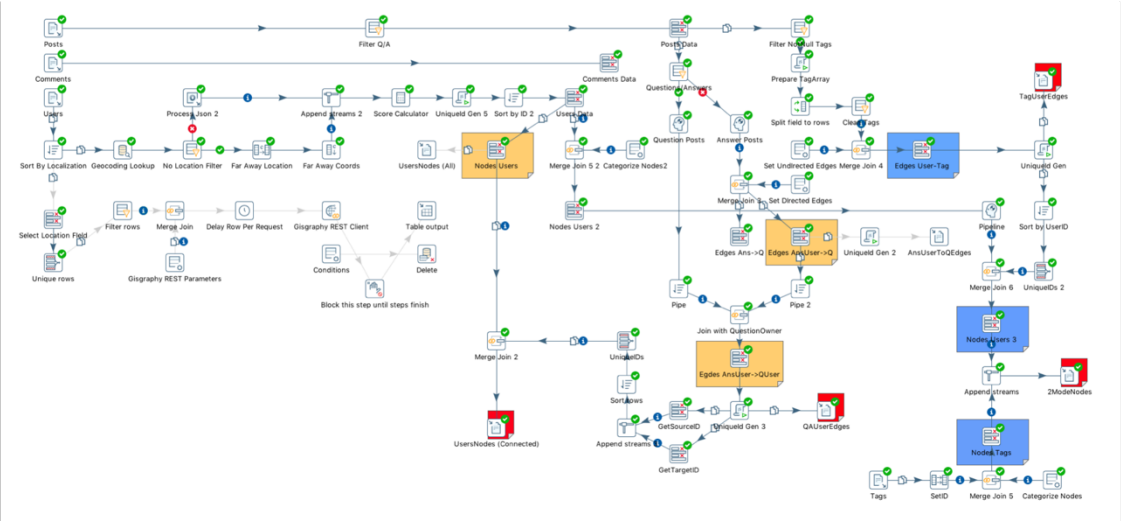

Figure 1. Customizable data-integration flow for Stack Exchange Network generation

In brief, with the aim to cover all the research question proposed for our analysis, we created two different types of graph representation: 1)a weighted directed graph of users with the relations [*Owner Question*] - [*Owner Accepted Response*], now called *experiment 1* network, and 2) an undirected weighted bipartite graph, or in other words, a 2-mode graph, of relations between users and used tags (experiment 2). Finally, both representations were serialized as CSV files (more information about the output network representation on appendix).

4.3 Methods and Tools for Network Analysis

For the network analysis stage, a combination of different statistical and graphical tools available for network analysis where employed. First of all, Gephy was the main platform used across the experiments. It was useful for visualization-based analysis and the computation of at least Third order network properties (degree distribution, centrality measures, modularity). Furthermore, a SNAP python script was developed to apply linking analysis algorithms (Page Rank, HITS), as well a Matlab library (developed by Clauset et. al. [2]) and its plotting tools to analyze the power law degree on a network.

[7] http://community.pentaho.com/projects/data-integration/

[8] http://www.gisgraphy.com/

5 Results

5.1 Experiment 1

The generated directed graph of [Answer User] – [Question User] for experiment 1 has 1589 user nodes with a 37% of them where geo-localized correctly, and 2618 edges weighted by the given score to the correct answer. As can be seen in Figure 2 and Figure 3, the social graph between users giving the correct answer (assessed and accepted by the question owner) and question user owners follows a power-law degree distribution. In other words, the data science community has few users giving most of the correct answers to posts. It gave an initial insight of how many expert users the community could have.

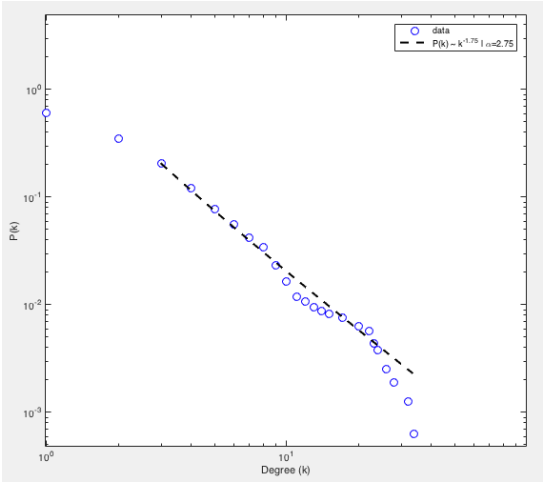

Figure 2. Power-law degree distribution

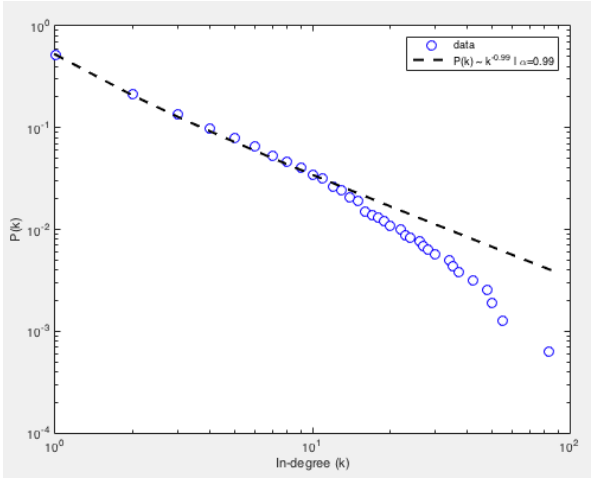

Figure 3. Power-law out degree distribution

The reputation of the community users was analysed by dividing them into two categories: expert askers (or *learners*) and answer specialists (or *experts*). Results through analysing the differences between the given user reputation score gained by the Stack Exchange *gamified* schema and the analysis of nodes and network properties can be compared on Figures 4 to 9. Link analysis was also performed, getting the top 10 experts and learners by using the HITS algorithm. Table 2 and Table 3 shows the main properties of the best community users.

	Location	Rep	Score	In-degree	Out-degree	Closeness	Betweenness	Cluster Coeff.
Aleksandr Blekh	Atlanta, US	4158	83	2	335	2873	437	0.002
Emre	Far away	1836	223	0	146	3528	0	0.002
Anony-Mousse	Series of Tubes(Far away)	1220	-37	0	101	2857	0	0.007
Steve Kallestand	Southern California	1971	158	126	146	2879	305.60	0.009
MrMeritology	Far away	1230	23	0	92	3187	0	0.0017
Seen Owen	United Kingdom	2069	363	0	176	3553	0	0.0016
Dacuny 33	Gurguen, India	1699	470	86	89	3362	23901.91	0.007
jamesmf	Far away	1255	12	3	84	3281	3867.37	0.009
damienfrancois	Far away	896	17	0	74	3.62	0	0.045
Indico	Boston	2189	25	0	216	2988	0	0.008

Table 2. Top 10 user experts by HITS

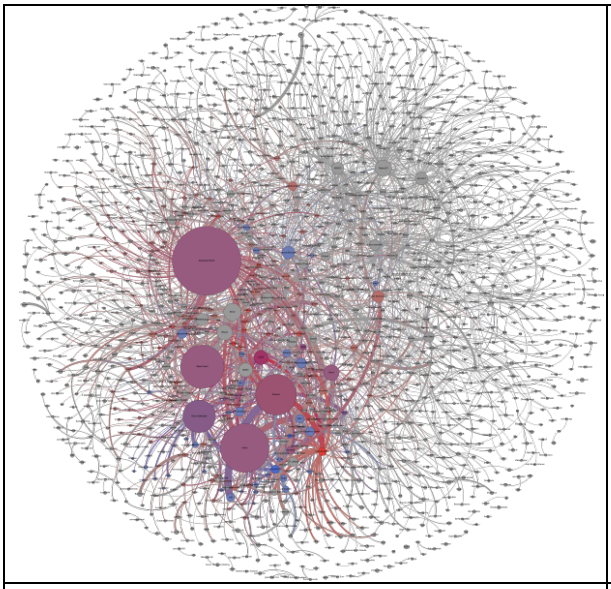

Figure 4. Weighted nodes by reputation

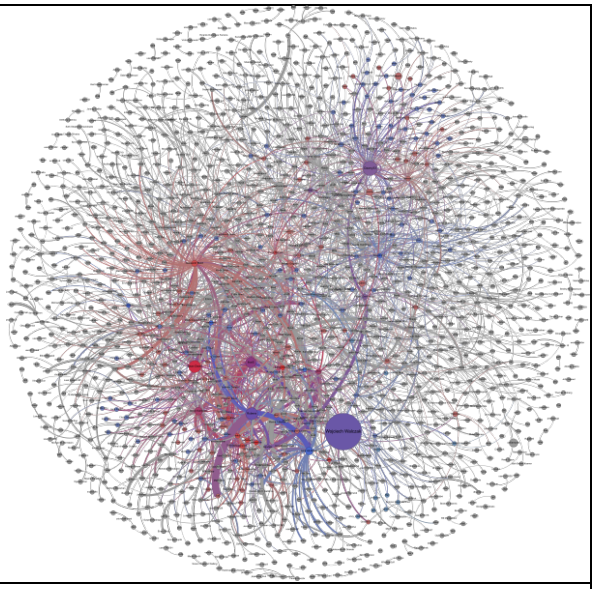

Figure 5. Weighted network by answer score

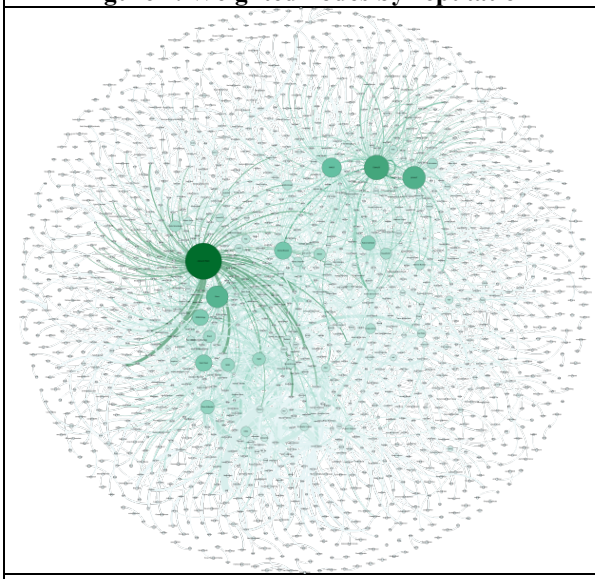

Figure 6. Weighted nodes by out-degree

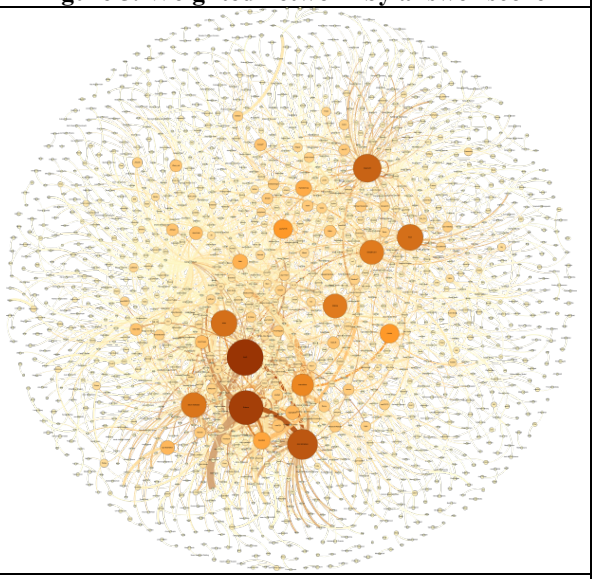

Figure 7. Weighted nodes by in-degree

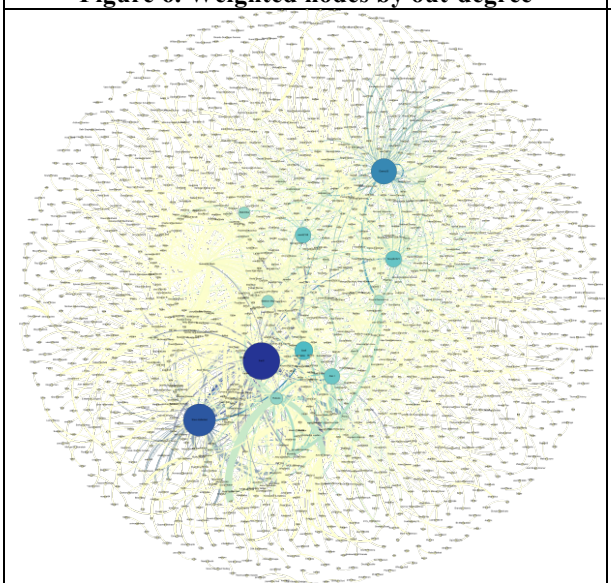

Figure 8. Weighted nodes by betweenness centrality

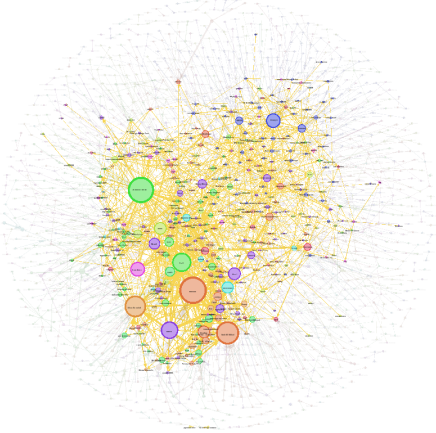

Figure 9. 3-core decomposition

	Location	Reputation	Score	In-degree	Out-degree	Closeness	Betweenness
IharS	Minsk Belarus	1413	155	83	3171	3544	0.0081
Rubens	Minas Gerais	2051	236	117	4758	10246.73	0.0202
Alvas	Singapore	219	105	0	1	44.5	0.015
Jack Twain	Far away	229	52	0	0	0	0.072
Matt	Far away	163	60	1	1	39	0.005
user62198	Far away	189	25	7	4154	13923.4	0.007
Rishika	India	231	38	0	0	0	0.027
Marcodena	Far away	524	140	0	0	0	0.002
JeanVuda	Edmonton, Canada	191	35	0	0	0	0.009
Piotr Migdal	Warsad, Poland	333	36	16	1.5	1504	0.007

Table 3. Top 10 learners by HITS

5.2 Experiment 2

The generated undirected 2-mode network graph of User – Tag for experiment 2 has 1184 user nodes, 229 unique tags and 4550 connections between them. Here, the Page Rank algorithm was used to get the trending topics in the data science community. Figure 10 presents a cloud tag of the top tags used by the users in theirs question posts. On the other hand, by executing the community detection algorithm, 13 strongly connected components where detected. Figure 11 highlights the found top tags by this approach.

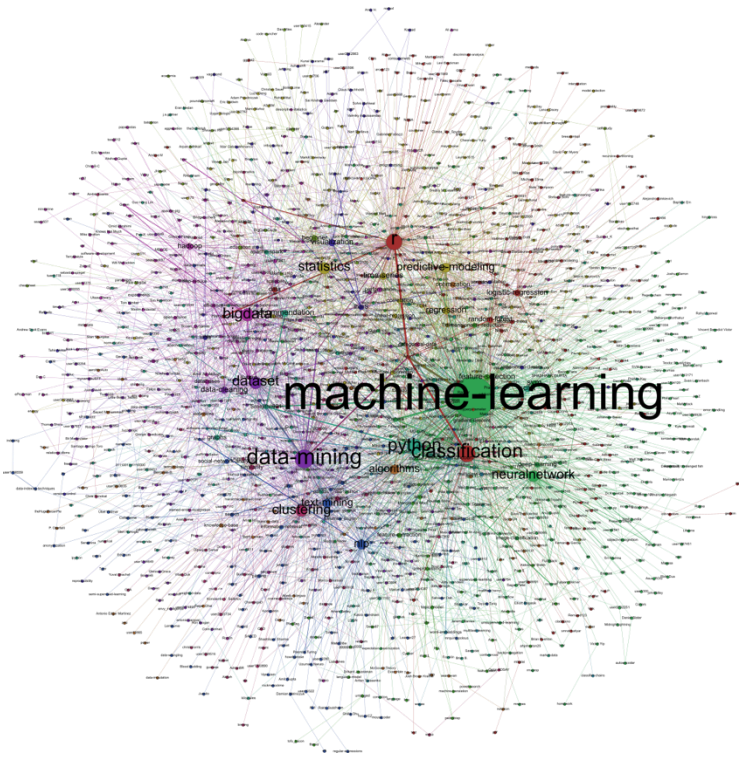

Figure 11. Modularity of the 2-mode network

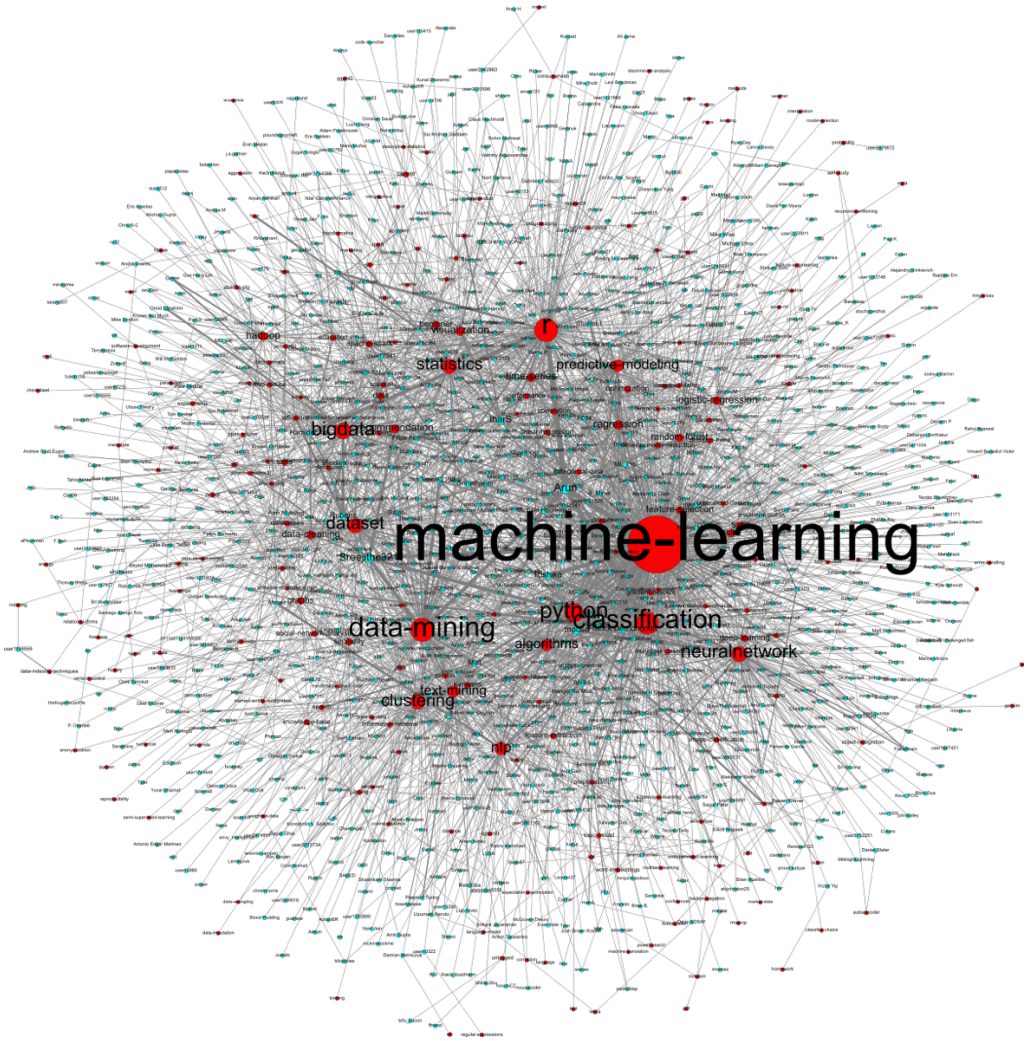

Figure 10. Top tags by Page Rank

6 Discussion

A comparison between the different gathered results leveraged the conclusions of the research questions proposed.

6.1 Question 1: Who are the top expert/learner users in the Data Science community?

First of all, the top experts and learners can be obtained in a network primarily by analysing the Hubs and Authority nodes it has. Figure 12 outlines the top experts and learners inside the network topology, where the top 10 experts coincide with the ones published by the *Area51* community page, more importantly on the winner (richer node) and data scientist Aleksandr Blekh, PhD[9], who has contributed to the expansion of the network through all the correct answers the provided. However, being an expert user is not only a matter of responding as many questions a user can, but also if he contributes with good and up-voted questions that are useful for the growing community.

[9] https://www.linkedin.com/in/ablekh

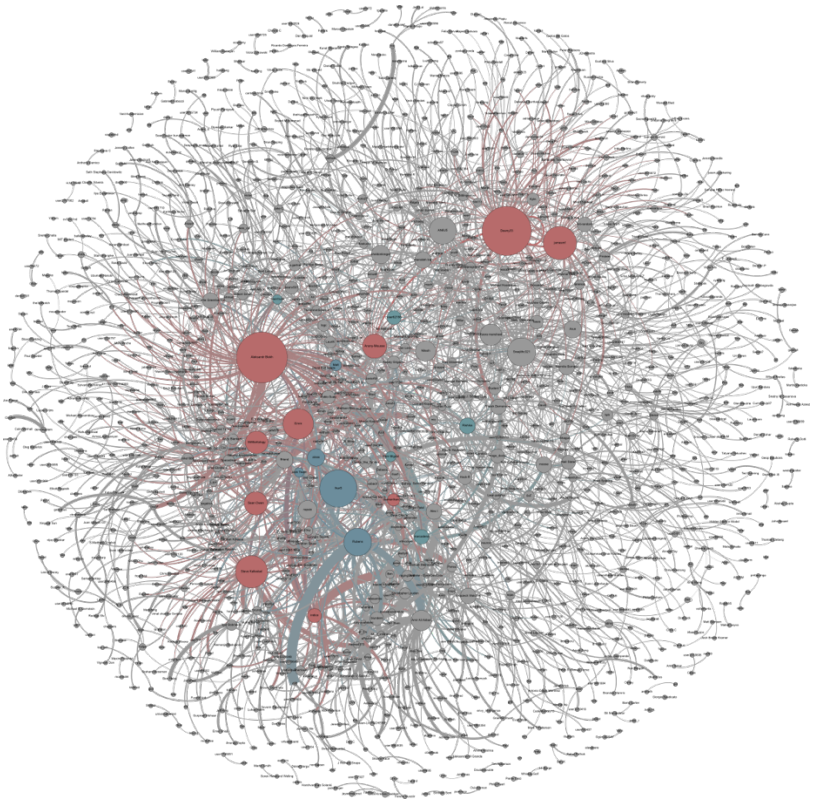

Figure 12. Top 10 experts (red) and learners (green) on the Data Science network

Node's degree measures and edges score weights are not good properties to get the best users in a community, while Betweenness centrality is possibly a good approach but it possible to get user ambiguity when trying to categorize him as learner or expert. Finally, the Stack Exchange reputation scores should not be taken in mind when analysing best users inside sub communities due to people earn points across all the groups they belong to, exploiting the reputation scheme just by interacting with the gamified network and/or giving fastest responses without any in-depth information with the goal to improve their reputation as it was a running competition.

6.2 Question 2: Where are the top users contributing to the Data Science community across the world?

A big drawback of the Stack Exchange dataset is the missing information about location of users. The majority prefer to stay anonymous on the web, but there are other people who want to get recognition of their knowledge either for finding good job positions or fame. Figure 13 draws the distribution of the community users across the world. For visualization purposes, users that could not be geo-localized, were situated in direction of the North Pole, specifically on the farthest point of Earth in that direction, and were categorized as a *Far Away* location.

As can be seen, a big amount of responses come from USA and India. The United Kingdom and China also play a good part in asking question or giving answers. On the other hand, few countries in South America and Africa are contributing or getting involved in the Data Science community. All in all, due to the big amount of users with misclassified location, this question was not completely answered, but the map obtained gave me some good thoughts of the distribution of data scientists across the world.

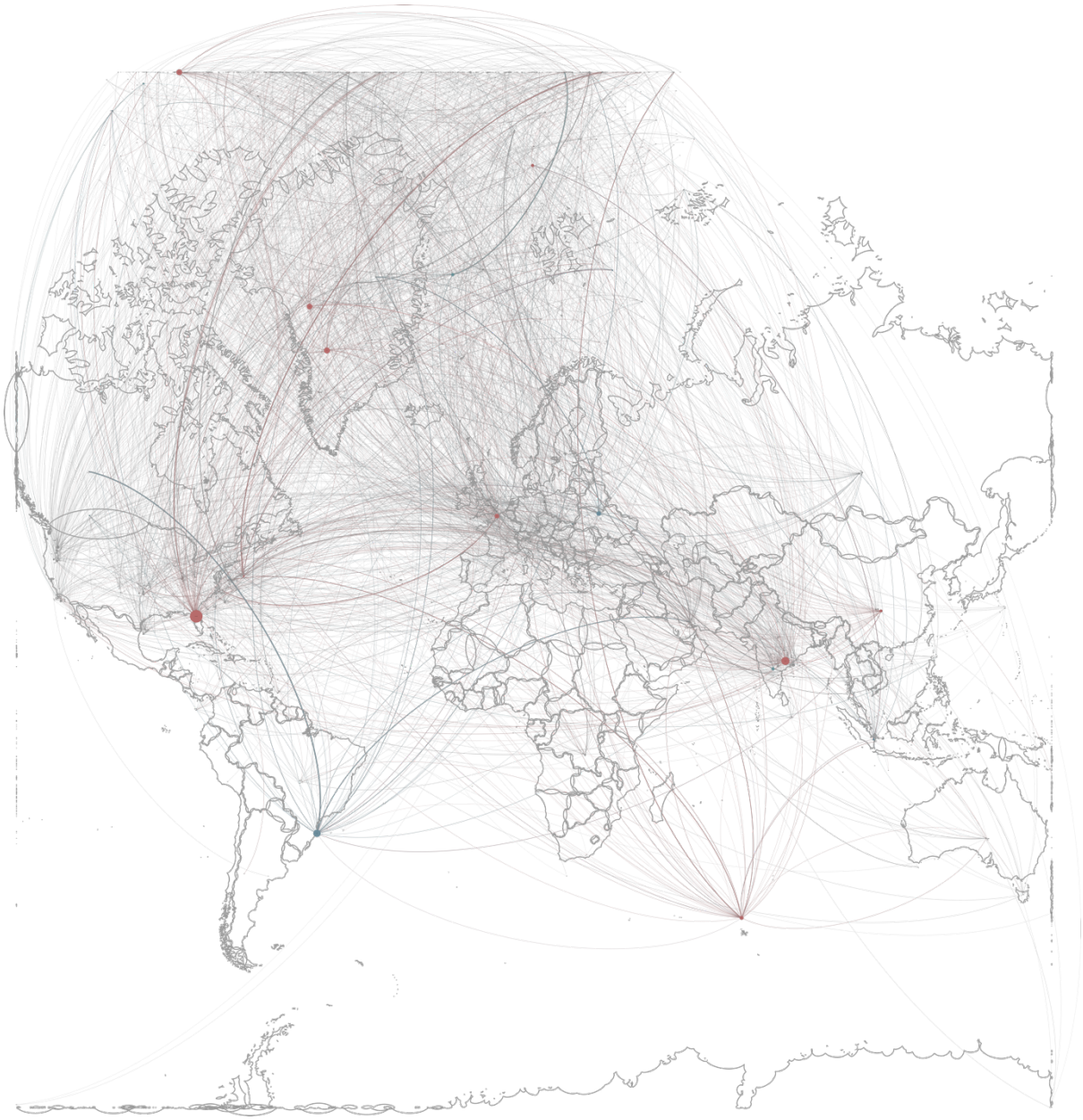

Figure 13. Data Science top experts/learners across the world

6.3 Question 3: What are the top Data Science topics covered by the community?

Thanks to the power of Page Rank, data scientists, data science students and learners in general could acquire some knowledge about the community and if it would be useful for them. Table 4 lists the trending topics spoken inside the community and that could be of big interest to UCL students. It is important to highlight the popularity of the emerging field of machine learning these days, and the still low but every day increasing interest about deep learning, which is getting more enthusiasts than social network analysis in the Data Science network, for example.

	Tag	Degree	Betweenness	Page Rank
1	Machine-learning	460	411569.33	0.06
2	Data-mining	187	103456.65	0.022
3	r	159	105783.07	0.022
4	Classification	163	81066.63	0.02
5	Python	146	83733.67	0.017
6	Bigdata	118	68998.96	0.014
7	Statistics	104	56347.32	0.012
8	Clustering	105	53306.12	0.012
9	Dataset	110	58906.62	0.012
10	neuralnetwork	99	46433.3	0.011
11	nlp	82	35408.61	0.01
15	Regression	54	17972.85	0.006
16	Visualisation	46	21080.87	0.006
19	Hadoop	40	17893.77	0.005
20	Sum	47	12963.67	0.005
29	Deep-learning	31	7362.58	0.004
35	Social-network-analysis	23	4089.27	0.003

Table 4. Trending topics on the Data Science community

6.4 Question 4: What are the top development tools covered by the community?

By using the betweenness centrality a list of the most used frameworks was obtained. Table 5 evidence that R and Python-based frameworks, platforms or programming languages are preferred above Java and Matlab. Additionally, the recently leveraged and Open Source framework for machine Intelligence Tensorflow[10], is getting more popularity than a well known Scala frameworks used for similar purposes.

Framework	Betweenness
R	105783.07
Phyton	83733.66
Java	2106.05
Apache-manout	1952.91
Matlab	1902.7
Tensorflaw	41.12
Scala	901.67

Table 5. Trending frameworks used in the Data Science community

In addition to the previous analysis made, a list of users whom are involved in most of the existing tags was extracted by its cluster coefficient. Table 6 shows the findings

Users	Cluster	Degree
Arun	0.006	15
Sreejithc321	0.006	24
IharS	0.004	30
Rishika	0.004	19
Moose	0.004	16
Frank Dermoncourt	0.003	21
Rubens	0.003	16
Alvas	0.003	18
Srinat	0.003	18
Hamidek Iraj	0.003	10

Table 6. Users involved in most of the Data Science tags

[10] http://tensorflow.org/

6.5 Lessons Learned and Future Work

The study case showed that the reputation score a user gain through his participation activity on the Stack Exchange platform is not a good value for community analysis. This is related with the *Fastest Gun in the West* problem: where with the aim to maximise their chances of collecting up-votes from their peers, participants would race to answer questions as quickly as possible, rather than as correctly or as exhaustively as possible [8]. On the other hand, Page Rank and HITS is a better approach for the discovery of experts and trending topics, and also help to find some inconsistencies between users earned score and the real reputation and expertise.

Future work and research on the Data Science community could be extended by performing a joint analysis between posts and comments. Usually comments does not contribute to the response but its score could have some correlation with the reputation of users. In addition, further network properties like the analysis of the assortative coefficient and rich clubs should be applied to have a fuller and better picture of the network. The possible finding would help to train a predictioner or a classifier to obtain a user ranking for decision making on companies when looking for employees or in the research industry to find possible experts to to build strong working groups.

7 References

[1] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, 2009.

[2] A. Clauset, "Power-law Distributions in Empirical Data," 2007. [Online]. Available: http://tuvalu.santafe.edu/~aaronc/powerlaws/. [Accessed: 28-Dec-2015].

[3] T. G. Lewis, *Network Science: Theory and Applications*, 2009th ed. New Jersey: John Wiley & Sons Inc., 2009.

[4] L. MacLeod, "Reputation on Stack Exchange: Tag, You're It!," *28th International Conference on Advanced Information Networking and Applications Workshops*, 2014.

[5] D. Movshovitz-Attias, "Analysis of the reputation system and user contributions on a question answering website: Stackoverflow," *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013*.

[6] M. E. J. Newman, *Networks: An Introduction*, 2011th ed. Oxford: Oxford University Press Inc., 2011.

[7] StackExchange Community Wiki, "Database schema documentation for the public data dump and SEDE," 2009. [Online]. Available: http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede. [Accessed: 20-Dec-2015].

[8] B. Vasilescu and A. Serebrenik, "How social Q&A sites are changing knowledge sharing in open source software communities," *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014.*

[9] StackExchange Community Wiki, "How does 'Reputation' work?," 2015. [Online]. Available: http://meta.stackexchange.com/questions/7237/how-does-reputation-work. [Accessed: 28-Dec-2015].

[10] S. Deterding, "Gamification: designing for motivation," *interactions*, 2012.

[11] A. Anderson and D. Huttenlocher, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012*

[12] M. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences, 103(23), 8577-8582*, 2006.