Mining Q&A websites to identify collaboration patterns and user reputation

Marko Vidoni

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2017

Abstract

The Q&A websites are becoming an ever more important component of problem-solving and learning. Increasing the responsiveness of the users and ensuring the quality of the posted content is the reputation reward system found in most of the today's Q&A communities. Reputation score, however, also has the role of recognising expert users for their contributions. The research presented in this thesis explores what the different factors governing the accumulation of the user reputation are. Accordingly, user's social position and its relationship to reputation is evaluated by constructing the social graph representing user interactions and performing social network analysis. Furthermore, text content posted by the users is analysed to conclude how a style of writing influences user reputation. Lastly, users are examined how their level of engagement in the Q&A community influences their reputation score. All these aspects are jointly modelled with latent variable analysis to examine the complete system of impact factors on the user reputation. Due to the ever-increasing importance of the online reputation, the conclusions made in this study have a far reaching and prominent impact, both online and in the real world.

Acknowledgements

I would like to thank Prof. Dragan Gašević, Srečko Joksimović and Vitomir Kovanović for their guidance during this dissertation project. All the numerous discussions and meetings made this project a real joy to work on.

Furthermore, I would also like to thank my family for their constant encouragement and financial support during my MSc studies.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Marko Vidoni*)

Table of Contents

Chapter 1

Introduction

In just a few decades of its existence, the *World Wide Web* has seen an explosive growth rate and has been evolving at a tremendous pace. We have become a constantly connected society where anyone can be reached instantly and almost everywhere in the world. However, in the era of ubiquitous internet access, the old problem of not having enough information was soon replaced by the new problem of having potentially too much information (Wilson and Risk, 2002; Bekkerman and Gilpin, 2013). This new limitation, in turn, gave rise to the specialised search engines, which were developed to help the users the most relevant results for their specific queries. However, the Internet does not only represent the enormous database of readily accessible information, but it has also enabled better social interactions of the people around the world.

A particular kind of social communication which has been enabled by the development of the internet are the question-answering websites (Q&A). These websites provide the platform for the collaborative knowledge building and sharing. The Q&A websites are the source of the well organised and structured human expertise which is often needed when solving more complex problems. Q&A websites consolidate interactions between users who have open questions about some specific topic and on the other hand, expert users which can use their prior knowledge to provide answers. In the process of posing questions and providing corresponding answers, an online database of knowledge is being gradually created. The content found on Q&A portals is constantly scrutinised by the self-policing of the user community which makes sure that the substandard content is removed (Correa and Sureka, 2013). By doing so, the Q&A networks ensure a high quality level of their recorded knowledge. This, combined with the efficient

and intuitive notion of posing questions and corresponding answers provided by other users makes for the observed high popularity of the Q&A websites (Wang et al., 2013a).

Due to their wide spread, the Q&A websites can be found covering a breadth of different topics, ranging from very topic specific all the way to general. A well known example of the former group is *Quora*[1] which is used for general discussions about various topic and without any strict regulations about the content. On the other side of the spectrum are those Q&A portals which are strictly oriented toward some specific, often scientific field. Arguably the best known examples from this group are the Q&A portals operated by the *Stack Exchange*[2] organisation. Each of the Q&A portals under their umbrella are intended for the discussions about a different topic. Their largest and generally most widely known Q&A community is *StackOverflow*[3] which is primarily interested in discussions and questions related to various aspects of software development (Slag et al., 2015). Due to the rapid growth and universal use in the world of the software development, the StackOverflow Q&A community has become an indispensable tool and part of the software developers' work process. StackOverflow is often the first place where software developers look for the solutions to their specific problems, resulting in it becoming one of the important resources in the field of software development. It should be also noted that StackOverflow has also attracted a significant interest from the research community Slag et al. (2015), given its potential to help understand the social aspects of human collaboration.

On the other hand, other StackExchange Q&A communities intended for discussions in different topics have been much less explored. Specifically, in this project the main focus is given to the sister Q&A portal CrossValidated[4] which is directed towards questions about the statistics and machine learning. As both, data science and artificial intelligence are already seeing a great amount of both academic as well as industry interest, the CrossValidated is fast becoming one of the more important Q&A communities. Due to this great potential to grow and increase in its widespread use, the CrossValidated Q&A network makes for a very interesting research target.

As a certain level of expertise and effort is required to answer a complex ques-

[1] https://www.quora.com/
[2] http://stackexchange.com/
[3] https://stackoverflow.com/
[4] https://stats.stackexchange.com/

tion, a potential lack of users' motivation to contribute might hinder the success of the Q&A community. Relying only on the good will of experts to answer posted questions is often not enough. Consequently, the Q&A sites use different techniques to incentive users to contribute with their questions and even more importantly answers. Users who create noticeable questions and more importantly good informative answers are often rewarded for their posts. Q&A websites use a wide range of rewards, for example, set of collectable virtual "badges" which are awarded to the users when they achieve a certain level of participation (e.g., writing a certain number of replies, being active a certain number of consequent days). However, arguably the most important reward method is the user reputation score, which also carries a considerable weight in the real life, not just in the online communities. The reputation score is representative of the amount of user's contribution to the Q&A community and is, in turn, a sign of their expertise. Users get their reputation increased by providing useful questions which also serve the wider community. Even more notably, users get their reputations increased by providing good, well accepted answers to other users' questions. However, users' reputation can also get degraded in case they post questions or answers which are deemed by the other members of the community as already answered in other threads, useless, or non-constructive (Movshovitz-Attias et al., 2013). This way, as it has been mentioned above, the Q&A community exercises a form of self-regulation which ensures a high quality of the content.

This project focuses on gaining an understanding of how are different factors associated with user's reputation in the Q&A communities. Concretely, by using the social network analysis, we examine to what extent user's reputation is associated with network social centrality. Furthermore, the analysis also takes into account the linguistic properties of user generated posts and different factors indicating the user engagement in the Q&A community. Such analysis will provide an insight into influencing elements governing user reputation accumulation. The conclusions will also help us better understand what makes users in Q&A communities successful.

There are two specific research questions being addressed in this project. The first is to learn how user's engagement, social position, and linguistic factors influence the corresponding reputation score. After having gained the results and insights based on the science oriented Q&A web portal CrossValidated, the second research question is interested in the non-science related, everyday life

Q&A portal. For this, the comparison of the reputation accumulation impact factors is made with the Parenting Q&A network[5], which is also operated by the StackExchange organisation.

The structure of the present dissertation is as follows. In the introduction section, we briefly cover the rationale behind this project and introduced its main goals and objectives. The background section presents previously conducted research in the area of Q&A communities and their social reputation. Next follows a detailed specification of our posed research questions. After that, the methods which were used in our research are presented. With the utilised methods explained, the thesis proceeds to the results presentation and furthermore their discussion. The thesis is rounded up with the section summarising the conclusions and proposing further research directions.

[5]https://parenting.stackexchange.com/

Chapter 2

Background

Due to their popularity and often convenient availability of the data sets, the Q&A websites have already received a lot of interest in the research. One of the most popular and consequently most widely researched Q&A portals is *StackExchange*, more concretely, it's software development oriented Q&A community *StackOverflow* (Slag et al., 2015). StackOverflow network has, for example, been examined from the standpoint of its badge award system and corresponding notion of gamification (Anderson et al., 2013; Jin et al., 2015), estimating the time needed to get a question answered (Goderie et al., 2015), what questions get deleted (Correa and Sureka, 2014), the analysis of provided code snippets (Yang et al., 2016a), to name just a few *(based on my IRP*[1] *report).* This research, on the other hand, focuses on examining factors associated with how users gain their reputation on Q&A portals. In addition to some of the commonly used metrics for predicting users' reputation, such as number of produced answers or questions and different user demographics attributes (e.g. user age) (Bhanu and Chandra, 2016; Procaci et al., 2016), our research utilises methods of social network analysis in examining to what extent users' social centrality suggests their social importance and reputation. Moreover, this research also examines what are the linguistic properties users employ in asking and answering questions that might be associated with the accumulated reputation in respective communities.

The following section discusses on the previous research and is structured around three main topics. The overview first starts with the approaches using general user attribute extraction and their analysis. Next, past research involving social network analysis is presented, then follows the discussion of prior research

[1]Informatics Research Proposal

which performed text content analysis. Lastly, the background section is concluded with the examination of latent factor modelling.

2.1 User engagement attributes

The content in this section was based on my IRP report. Observing engagement factors (e.g., speed of the first response, number of comments) that are potentially associated with users' reputation, Bhanu and Chandra (2016) focused on discovering expert users within the StackOverflow community. Specifically, difficult questions that were claimed to be answered by expert users in a given domain and training random forest classifier to automatically identify such users. A question was labelled as difficult if its view count was less than the standard deviation of the dataset. To automatically distinguish between difficult and easy questions they used the described labels and trained a separate random forest classifier on features such as question score, first response time, comment number, answer length ratio, as well as several specifically calculated features. With difficult questions classified, Bhanu and Chandra (2016) used user related features to train the second random forest classifier which classified users as experts. The features used were for example: a user's total question and answer score and count, question-answer ration, post rate and several others. Authors tested several procedures of labelling training set users as experts, for example, if the number of user's accepted answers surpassed a certain threshold or if the number of difficult questions answered by the user exceeded a certain amount.

In a similar manner, Procaci et al. (2016) looked at another Stack Exchange community, which focuses on the question from the field of biology (Biology Stack Exchange). In their analysis, Procaci et al. (2016) looked at the user's reputation in connection with the user's number of questions, answers, and comments. Furthermore, Spearman and Kendall correlation coefficients were calculated to more rigorously inspect and confirm the correlation between the described features and user's reputation. Procaci et al. (2016) showed that the number of provided answers was significantly associated with the accumulated reputation. The most reputable users were found to be mainly focused on answering other users' questions rather than asking new questions.

After having presented the prior research using various general user engagement attributes, the focus is next given to more targeted methods used to analyse

Q&A networks in the past research. In the following subsections follows a brief introduction and presentation of related research in the fields of social graph and text analysis respectively.

2.2 Analysis of social interactions in Q&A forums

2.2.1 Social network analysis: a brief overview

The general network science overview presented in this subsection has been taken verbatim from my IRP report. The theoretical foundation of the social network analysis is given by the mathematical graph theory. Generally speaking, a graph G consists of a collection of vertices V (or nodes), and a set of edges E (Beyer and Pinzger, 2016), where each edge e connects two vertices (v_a, v_b). For example, a transportation network can be represented as a graph where cities and towns represent vertices, and roads and highways between them represent edges. Similarly, airline routes can be represented using a graph where vertices represent airports while edges represent routes between airports. Edges can be either un-directed, in which case there is no distinction between the vertices v_a and v_b, or directed, in which case v_a is called the *source* and v_b the *target*. For instance, the transportation graph is an example of an un-directed graph, while airline graph is an example of the directed graph (an edge (A, B) indicates the existence of a flight between airports A and B). When edges are without directions, a subsequent graph is called the un-directed graph. On the other hand, when the edges are directed, the corresponding graph is called directed graph.

The described mathematical graph formulation is used in the social sciences and social network analysis to represent people and their relationships or interactions (Scott, 2017). Each user corresponds to a node in a graph and a connection between two users are represented as an edge between the two user nodes. An important idea that arises from such social network is the notion of a person's position in the network and its benefits. By studying the mathematical properties of the networks one can gauge the importance of the network participants.

2.2.2 Constructing social graphs

Probably the most widely known application of social graphs modelling is in the case of social networks. The following discussion in this subsection is based on my IRP report. The largest social network Facebook, for example, connects its users with the notion of the friendship. Due to this mutual relationship, the users on Facebook can be represented in a graph with undirected connections. On the other hand, another widely popular social network Twitter allows users to follow each other. Contrary to the notion of mutual friendship, the action of following is one-directional. The users following each other and the community they are forming are represented as the directed graph. There, the user following another user is the source of the directed connection and the destination of the directed connection is the user that is being followed.

Contrary to the fully fledged social networks where the user network structure is inherently apparent, in the case of Q&A websites, social graph construction is not so clearly defined. This, in turn, means that there are multiple options for graph construction which have to be adapted according to the needs of the specific research problem. For example, Movshovitz-Attias et al. (2013) have presented a system for prediction of the future highly reputable users, based on just a limited amount of historical data *(from my IRP report)*. More concretely, Movshovitz-Attias et al. (2013) have used three different StackOverflow user interaction representations to build social graphs needed for their analysis. The first option was to connect the user that asked a question to all the users that answered that specific question. The second social graph construct they used was to connect the user asking the question with the user who wrote the accepted answer. Answer acceptance is a specific feature of StackOverflow portal which gives the user that posed the question the ability to mark the answer that was the most useful and in turn solved the problem. The third graph construction utilised by Movshovitz-Attias et al. (2013) connected the user which asked a question with all the users that answered and got their respective answers upvoted.

The effects of approvals and up-/down-votes in the network construction are twofold. It serves to the Q&A community as an indication that a specific post might be worth looking at. Apart from that, it also affects the reputation score of the user that produced the answer. Authors of well received high quality posts are rewarded with an increased reputation score. More concretely, the

StackExchange, which owns some of the most popular Q&A networks, operates in all of them a complex user reputation reward scheme[2] which incentivises high volume and quality of user contribution. In this discussion, only the most common and important rewards are presented. The user with an accepted answer gets 15 and for each answer upvote the user gets 10 additional reputation points. Conversely, good questions are also rewarded by five points for each question upvote. On the other hand, by having posts downvoted, the users lose two points.

A similar social graph construction approach to Movshovitz-Attias et al. (2013) was also taken by Choetkiertikul et al. (2015). In their research, they were predicting who will answer new posted questions based on different attributes of the posts and users posting the questions. One of their taken research directions focused on the social network based prediction. For this, a graph was constructed which again represented users with nodes and their social interactions with edges. Contrary to the previously discussed graph construction techniques, Choetkiertikul et al. (2015) used a weighted directed graph. They added a connection for the notions of a posted answer, the answer is accepted, or the comment is posted. However, for each of these actions, a separate new connection was not made, but the weight of the connection between the two interacting users was increased.

2.2.3 Social network analysis

The theory of social network analysis (SNA) provides a wealth of methods for understanding the ways in which social networks are constructed and utilised in user interactions. The discussion of different approaches to SNA is largely taken verbatim from my IRP report. An important concept in SNA is the idea of *centrality* which captures how a certain position in the network is beneficial for that user node (Nieminen, 1973; Freeman, 1978). The simplest form is *degree centrality*, which is a measure of the number of edges a particular node has in the graph (Zuo et al., 2012). In the context of directed graphs, there are separate *in-degree* and *out-degree* centrality measures, corresponding to the number of incoming, and outgoing edges, respectively. Besides degree centrality, there are also more complex and specialized centrality measures, such as *betweenness* centrality (Newman, 2003), *eigenvector* centrality (Ruhnau, 2000), or *closeness* centrality (Cohen et al., 2014).

[2] http://stackoverflow.com/help/whats-reputation

Apart from various centrality measures, there also exist other network node importance indicators. Movshovitz-Attias et al. (2013), for example, used the *PageRank* algorithm which is another example of node centrality measure closely related to eigenvector centrality (Jing and Baluja, 2008). Movshovitz-Attias et al. (2013) were inferring which users were the most important and reputable in the network, based on the network structure, which was constructed based on users' interactions. In addition to the PageRank algorithm, the authors further extended their analysis by using *Singular Value Decomposition (SVD)*. This measure can be also utilised as the centrality measure for the nodes in the network. Specifically, in the paper by Movshovitz-Attias et al. (2013), its primary use was for the network anomaly detection. The SVD algorithm was used to find users asking an abnormally high number of questions and those users providing an abnormally high number of useful answers. The latter could be useful for the identification of high user reputation.

Other authors have also gone for the more specialised algorithms constructed specifically for their research problem (Immorlica et al., 2015; Yang et al., 2016b; Xu et al., 2016). However, it must be noted that most of such prior research using specialised solutions to the processing of social network was not focused on the data analysis and explanation. These papers were rather interested in the development of the new network algorithms. A concrete example is the paper by Yang et al. (2016b) which researched better ways, to assign reputation to users and search for expert users. The presented a novel algorithm which jointly integrated network analysis process with the text topic modelling to predict future expert users on StackOverflow. As such, these methods are directly aimed as improved replacements for the current StackExchange reputation reward system.

2.3 Text analysis

Apart from the user's social position examined by performing the social network analysis, text content produced by the user can also be indicative of their reputation (Calefato et al., 2015). The discussion of text analysis is partly based on my IRP report. To gain insights into user produced text content of their posts various natural language processing techniques are normally employed. The text processing step is needed due to the unstructured nature of the text data type. Text in its raw form is not ready for further analysis because representative fea-

tures first need to be extracted. In the past conducted research, authors have, for example, used different text analysis software tools to automatically extract information from the input text. Bazelli et al. (2013) used *Linguistic Inquiry and Word Count (LIWC)* tool[3] to examine the linguistic properties of the text content posted on the StackOverflow Q&A network.

Procaci et al. (2016) performed a topic extraction from the StackExchange Q&A portal Biology which is, as the name suggests, intended for biology related discussions. To produce topics related to users' posts, they used the *DBpedia Spotlight tool* which automatically extracted structured information from the text content of the posts. With the DBpedia Spotlight tool, the authors identified several entities or concepts which they further utilised as topics which are discussed in the analysed texts. Apart from using DBpedia Spotlight tool, they also employed a well known natural language processing method *TF-IDF*. The method is, in essence, an improved version of simple bag-of-words text representation where just word frequencies are calculated. The *TF-IDF* method improves this approach by using word frequencies but also taking into account the spread of words across different documents in the examined corpora (Jurafsky and Martin, 2009).

Contrary to the discussed previous research that mainly used different off-the-shelf solutions for text feature extraction, Yang et al. (2016b) utilised a more involved approach. Yang et al. (2016b) incorporated topic modelling directly into their algorithm for automatic identification of the expert user. Specifically, the authors used *latent Dirichlet allocation (LDA)* to extract topics from the textual content of StackOverflow posts. In turn, the LDA modelling produced the topic distribution of users based on their answers. The topic related features were then jointly modelled with the user network structure. The described unified model was developed with the aim of improving the user reputation reward system which is currently in utilised by the Q&A portals run by StackExchange.

[3]https://liwc.wpengine.com/

2.4 Relationship between factors influencing the reputation

The relationship between the person's social position and their achievements or reputation is also observed by the notion of the social capital (Burt, 2000). Social capital is a measure of person's importance in the society based on their position in the social network. According to Burt (2004), people who have connections in several communities and thus provide a bridge between them are in an advantageous position. The mere social position, however, is not the only factor governing user's achievements. In the research by Burt (2012), the author jointly modelled the user social position together with the user engagement attributes to identify the level of user achievement in online games. The social position and thus social capital was primarily indicated by the number of user's non redundant contacts. On the other hand, the user engagement factor was represented by attributes such as the amount of time a certain game character was played by the user, the number of different characters played, and age of the user.

Similarly, Dowell et al. (2015) and Joksimović et al. (2015) model the social capital and its association with the language style used by the participants of *Massive Open Online Courses* (MOOCs). The authors examined how extracted linguistic features affect performance and social position of learners as they interact in a MOOC.

2.5 Latent variable modeling

To perform modelling of different factors which together predict the target variable a latent variable analysis can be utilised. Based on the past research presented in the previous section, the user reputation is not governed by a single type of attributes. The user reputation is jointly based on the various factors (e.g. user social position, engagement, and writing style). For such a problem, latent variable modelling enables specification of different observed variables, their representation in factors and finally relationship with the user reputation.

The latent factor modelling has already seen a substantial amount of research use. The approach has been widely utilised especially in the areas of psychology, mental health research, and other allied disciplines (Cai, 2012). A detailed and

comprehensive review of the latent factor analysis and the corresponding models can be found in the paper by Skrondal and Rabe-Hesketh (2007). According to Schreiber et al. (2006) there are two main statistical techniques that are used for the latent factor analysis. These are *confirmatory factor analysis (CFA)* and *structural equation modelling (SEM)*. The former is intended as a confirmatory technique, but the latter can be used for both confirmatory and exploratory analysis. More importantly, according to Bollen and Noble (2011), structural equation models can be utilised as multi-equation models providing multiple measures of concepts and measurement error. This, in turn, means that the models are very well suited for examining causal relationships among variables, especially in the more complex systems with interconnected variables which are often found in real life (Beran and Violato, 2010).

An actual example of the latent factor analysis application in the past research can be seen in the paper by Tzeng et al. (2015). The authors used latent variable modelling to inspect factors related to research proposals' approval after their initial review. In their study, they examined the latent factors describing investigators, vulnerability, and review process. Each of these latent factors has been constructed from the further observed variables, such as license level of the investigators, which institution they belonged to, administration time, and review frequency. Another example of latent factor analysis application in the research can be found in Beran and Violato (2010). In this paper, the authors used SEM modelling for medical and health sciences research. The two presented models were used to better understand patients' experiences of schizophrenia and examine the relationship between childhood victimisation and school achievement.

Chapter 3

Research questions

The Q&A communities are gaining popularity and the achieved reputation gives a recognition to the user in virtual, as well as in the real life. Consequently, the factors governing the user reputation accumulation have already seen interest in existing research (Movshovitz-Attias et al., 2013; Bosu et al., 2013; Procaci et al., 2016). Existing studies have mostly focused on the development of new algorithms for automatic user expertise and reputation prediction (Wang et al., 2013b; Choetkiertikul et al., 2015). Another common goal of previous studies is the development of better reputation assignment algorithms (Yang et al., 2016b). These could potentially be used by Q&A web portals in place of the current procedures for reputation accumulation.

The past papers have mainly focused on the development of the novel algorithms. However, the research has mostly left out the comprehensive analysis of the existing Q&A networks and the factors that govern the user reputation accumulation. Furthermore, previous research has often focused just on the best known Q&A network StackOverflow (Movshovitz-Attias et al., 2013; Goderie et al., 2015), but rarely on other smaller Q&A communities (Procaci et al., 2016), such as CrossValidated or Biology. Addressing these gaps in the existing literature is the main goal of this project. The aim is to perform a data analysis and assess how users accumulate their reputation in Q&A communities. A Q&A network CrossValidated, intended for statistics and machine learning topics is of primary interest for this research. However, explicit focus on science related communities is bad because the social dynamics discovered in those is not the same across all online Q&A communities. Thus, we need to explore the factors of social positioning for other types of communities. Consequently, this study

also analysed a non-science related Q&A community Parenting intended for discussions about to the raising of kids. As such, this project answers the following research questions:

Research question 1 What are factors that drive the accumulation of a user's reputation and what is the impact of such factors on the reputation accumulation?

Research question 2 The purpose is to validate the extent to which the results of research question 1 hold in a different context. The question is interested in how do results from networks concerned with topics in science compare to those from Q&A networks for non-science topics?

In order to answer these two research questions, we structured our research around three hypothesised factors that drive user reputation. First, users were inspected from several aspects to conclude how different user's attributes correlated with their reputation. Namely, we explore users' engagement within the selected community by measuring their account maturity, the ratio of provided answers compared to the questions, and the amount of sub quality posts they have produced. Second, the social network of users' interactions was constructed. Social network analysis was used to determine how a user's social position was related to their reputation. Third, the inspection of the actual text content that the users produced was performed. The text was examined for the presence of different writing styles (e.g. analytical, confident, authentic and emotional), as well as language aspects (e.g. sentence length and frequency of longer words). These factors were first separately analysed for the association with the reputation of the users. Then, a unified analysis was performed by using SEM, where all the user attributes were jointly modelled and examined as part of the system that influences their reputation.

To answer the second research question, an additional Q&A network and its data were analysed. More importantly for the second research question, the chosen Q&A network was focused on a non-technical domain, thus serving a different community of users. To answer the posed research question, the main interest was given to the results produced by the models which were jointly modelling all the user describing attributes. These results were used for the comparison, which showed how do the users from technical and non-technical Q&A communities differed in how are their reputation scores influenced.

Chapter 4

Methods

4.1 Data

This research project focuses on Q&A web sites and their communities operated by StackExchange organisation. All the data coming from their various Q&A web communities were open for public access and use, making it especially convenient for research usage.

There are two main approaches for StackExchange's data collection. The easiest way is to use the Data Explorer platform[1] provided by StackExchange organisation. The user has to specify and run the *SQL* query which in turn returns the data and can be downloaded to the local disk in a *comma separated values* (*CSV*) format. However, Data Explorer platform imposes a strict data size limit as it returns at most 50,000 records per query. This is severely limiting when comes to bulk data collection needed for data analysis such as the one presented in this report.

Due to the data size limit, for the purpose of this project, the second option of accessing StackExchange's data was utilised. The data sets associated with all their Q&A web sites are regularly published online by StackExchange[2]. In each of the respective Q&A communities' data dumps only eight main tables are published, however the full StackExchange database schema can be found online[3]. The tables present in the data dump include *Users*, *Posts*, *PostHistory*, *PostLinks*, *Comments*, *Badges*, *Votes* and *Tags*. From these, the main focus of the project

[1] https://data.stackexchange.com/

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

[3] https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

was given to the tables *Posts* and *Users*. The first one contained posts' textual content and posts' related metadata and the second table including various user information. However, the data in these data dumps was initially not in a table format but in *XML* formatted text files, with each file representing one database table. Code listings 4.1 and 4.2 represent the examples of the XML documents containing the posts and user data as they are found in the StackExchange data dump[4]. Each row element represents a row in the posts and user tables and its attributes correspond to columns in the table. As such, after downloading the data dump for a specific Q&A community, its data first needs to be imported into a database for more convenient subsequent data analysis. In this project the *SQLite* file-based database system[5] was used due to its light-weight nature and high portability. When the XML data was imported into the database the data schema was exactly the same as the one found on Data Explorer platform but with the benefit of not having the data size limit.

Listing 4.1: Excerpt from XML document containing posts data.

```
<?xml version="1.0" encoding="utf-8"?>
<posts>
  <row Id="1" PostTypeId="1" AcceptedAnswerId="15"
       CreationDate="2010-07-19T19:12:12.510" Score="36"
       ViewCount="2335"
       Body="&lt;p&gt;How should I elicit prior distributions
             from experts when fitting a Bayesian
             model?&lt;/p&gt;&#xA;"
       OwnerUserId="8" LastActivityDate="2010-09-15T21:08:26.077"
       Title="Eliciting priors from experts"
       Tags="&lt;bayesian&gt;&lt;prior&gt;&lt;elicitation&gt;"
       AnswerCount="5" CommentCount="1" FavoriteCount="23" />
  <row Id="2" PostTypeId="1" ... />
  ...
</posts>
```

Listing 4.2: Excerpt from XML document containing user data.

```
<?xml version="1.0" encoding="utf-8"?>
```

[4]https://archive.org/details/stackexchange
[5]https://www.sqlite.org/

```
<posts>
  ...
   <row Id="10" Reputation="121"
        CreationDate="2010-07-19T19:05:40.403" DisplayName="USER"
        LastAccessDate="2016-09-06T15:35:39.697"
        WebsiteUrl="http://plindenbaum.blogspot.com"
        Location="France"
        AboutMe="Bioinformatician&#xD;&#xA;Virology&#xD;&#xA;
                Genetics&#xD;&#xA;Biology&#xD;&#xA;Science&#x
                D;&#xA;Science20&#xD;&#xA;Web2.0&#xD;&#xA;Bio
                informatics&#xD;&#xA;Genotyping&#xD;&#xA;Wikipedia"
        Views="21" UpVotes="4" DownVotes="0"
        Age="47" AccountId="23234" />
  <row Id="11" Reputation="136" ... />
  ...
</posts>
```

For this project, the initial development aimed at analysing the largest of the StackExchange Q&A communities StackOverflow. This network's corresponding raw data dump size was more than 80 GB which made it extremely hard to manage with the limited computing resources and time available for this project. By using highly scalable SQL queries, data processing, and parallelised code execution some initial limited results were produced on university's computing servers. However, in the later social network analysis stage, it turned out that such a large social network was too demanding to be processed and to produce necessary analysis results in the limited amount of time available for this project. The described problem related to data size and limited computing resources has already been identified in the research proposal phase as a potential future issue. Accordingly, another smaller and more manageable StackExchange Q&A network called CrossValidated[6] was used. The main focus of the project's research questions is on the analysis of global trends in the online Q&A communities and not on StackOverflow specifically. Due to this reason, it was reasonable to use a network computationally manageable with the available resources and time. The results of global trends analysis likely did not vary considerably between different

[6] https://stats.stackexchange.com/

science oriented online Q&A networks.

CrossValidated is a Q&A network used for questions related to machine learning and statistics. The network was launched in July of 2010. The dataset used in this project consists of all the data accumulated from the website's inception until March of 2017. In the later stages of this study non-technical Parenting Q&A network[7] is also examined for results comparison with the main CrossValidated network. Parenting network has been in existence for roughly the same amount of time, being created in March of 2011.

Table 4.1 presents basic descriptive statistics of both used Q&A portals. It can be seen that the science oriented CrossValidated website received much more interest than the Parenting one, having more registered users and posts than non-technical network used primarily by parents. However, the statistics of the average number of posts that an active user which has posted at least once is very comparable for both networks. This indicates that both networks have had roughly the same amount of activity by their users.

Table 4.1: Basic network data statistics.

Q&A portal	CrossValidated	Parenting
Registered users	111,974	17,935
Active users	50,943	5,895
Active user ratio	0.455	0.329
Posts	193,024	21,487
Questions	97,832	4,889
Answers	95,192	16,598
Average #posts/user	3.789	3.647

4.2 Social graph construction

The format in which the Q&A network data was stored in the database's table form is useful in its initial state for some types of the analysis, but not for network analysis. A mathematical graph representing the network has to be constructed for the subsequent use in social network analysis.

[7] https://parenting.stackexchange.com/

In the constructed graph network representation, each user corresponds to a node in the network. Connections (edges) between network nodes represent interactions between users in a form of question asking and answering. A directed connection u_1u_2 from user u_1 to user u_2 is created if user u_1 asked a question and user u_2 answered it (Movshovitz-Attias et al., 2013; Scott, 2017). This network construction principle is presented in Figure 4.1 where, for example, user 3 answered a question from users 1, 4, and 5.

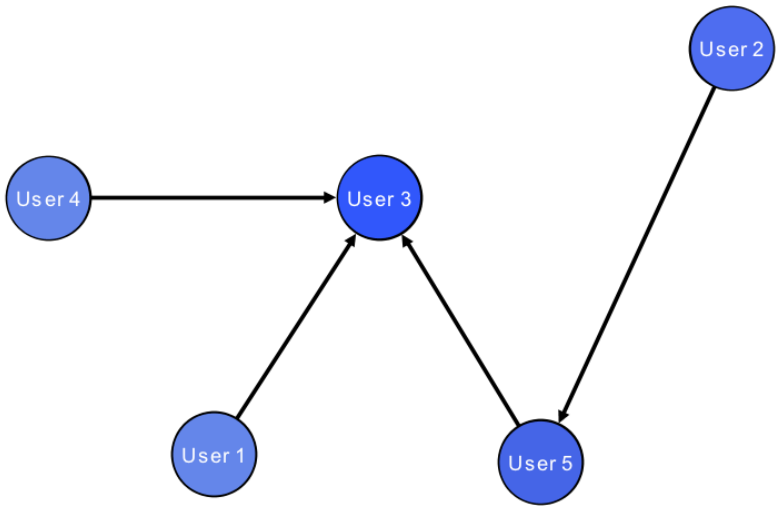


Figure 4.1: Q&A network represented as the social graph.

For network construction and later network analysis, Python-based network analysis library *NetworkX*[8] was used. To construct a network using this library, we had to provide a list of network node pairs where each pair represents a connection in the network. The NetworkX library then built a social graph and stored it as an object in the computer memory. Such constructed graph object could then be analysed using network analysis algorithms implemented in the NetworkX library.

4.3 Social network analysis

With the Q&A social network represented as a mathematical graph, social network analysis can be performed. For this, since the final selected Q&A networks are of a reasonable size, *NetworkX* library was used.

[8]https://networkx.github.io/

4.3.1 Degree centrality

Centrality measures were used to indicate how important a certain node in the network was. The basic centrality measure is degree centrality which represents how many connections a node has with neighbouring nodes (Salter-Townshend et al., 2012). An illustrative example can be found in Figure 4.1, where the node representing *user 4* has degree centrality of 1, while more active *user 3* has the degree centrality of 3. The same logic applies to *user 5*, who has one in-going and one out-going edge, in total making their node degree equal to 2.

The main idea behind the use of degree centrality measure is that a well connected node with more social interactions is possibly a more important and reputable member of the community. Since the user gains reputation points for answering questions as well as asking good questions, the overall node degree was used instead of node in-degree representing only the number of answered questions. The reasoning behind this is that the user with no produced answers, but many good questions, will have a larger reputation than another user that answered only a small number of questions.

4.3.2 Betweenness centrality

The betweenness centrality measure shows to what extent a certain node in the network is in control of the transfer and spread of information across the network and indicates the information brokerage factor of the node (Abbasi et al., 2014). Betweenness centrality is also one of the possible measures of a node's social capital (Burt, 2000). Social capital is a measure of person's importance in the society based on his or her position in the social network. Users with bridging connections in different communities are in an advantageous position (Burt, 2004).

Betweenness centrality measures how much information could flow through a certain node. It is formally described as a measure of how often some node lies on the shortest path between any two nodes in the network. If multiple shortest paths exist between a certain pair of nodes, each path is given an equal weight (Newman, 2003). The betweenness of the node i is the fraction of shortest paths between different pairs of nodes in the network that also pass through node i (Freeman, 1978). The calculation of the betweenness centrality is formally specified in Equation 4.1 (Newman, 2003), where $p_i^{(st)}$ stands for the number of

paths from node s to node t and passing through node i. Conversely, $n^{(st)}$ stands for the number of all paths going from node s to node t. Lastly, N is the total number of nodes in the network.

$$b_i = \frac{1}{\binom{N}{2}} \sum_{s<t} \frac{p_i^{(st)}}{n^{(st)}} \tag{4.1}$$

An example of the betweenness centrality measure is presented in Figure 4.2. The nodes coloured in red have an important position in the network for the control or brokerage of the information flow. Both red nodes lie on all the shortest paths between blue and green communities. Consequently, the red nodes can potentially prevent free information flow between the blue and green nodes. Thus, these two (red) nodes garner a high betweenness centrality score.

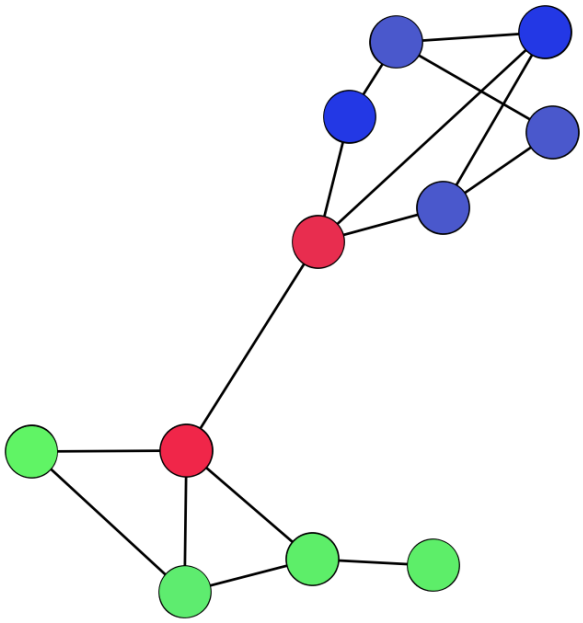


Figure 4.2: Example of a network with high betweenness centrality nodes.

4.3.3 PageRank

The algorithm by Page et al. (1998) was primarily developed for the use in web search engines, where it is used to score the importance of the webpages which are represented as nodes interconnected with hyperlinks to form the network. In turn, these scores help the search engine determine which websites are first shown in the web search result (Newman, 2010). However, the algorithm can generally

be applied to various network oriented problems where the importance of the nodes needs to be determined.

The main idea of the algorithm is that a certain node in a network is considered important if other important nodes point to this node. In the case of this study, the users that answer important questions posted by other reputable users will have a high PageRank score (Wang et al., 2013b). The common implementation of the PageRank algorithm executes iteratively, meaning that it is repeatedly updating scores of the nodes in the network. The procedure starts by initialising scores of all the nodes in the system to the uniform value of $1/N$, where N stands for the number of nodes present in the network. Then the algorithm proceeds with the iterative execution where scores are repeatedly updated for each of the nodes in the network using the following equation (Newman, 2010):

$$PR(a_i,t) = \frac{1-d}{N} + d \sum_{a_j \in IN(a_i)}^{N} \frac{PR(a_j,t-1)}{N_{out}(a_j)} \tag{4.2}$$

In Equation 4.2 for calculating the PageRank score updates, variable d stands for the damping factor which is set to a value between 0 and 1, though Page et al. (1998) suggest using the value of 0.85. Variable $PR(a_i,t)$ represents the new PageRank score at time t that is being calculated for node a_i. Conversely $PR(a_j,t-1)$ represents an old PageRank score calculated in the previous iteration $t-1$ for node a_j which is one of the nodes having the connection pointing to the current node a_i. Next, $IN(a_i)$ is a list of all nodes with a connection directed towards a_i and lastly, variable $N_{out}(a_j)$ stands for number of the outgoing connections from node a_j. The PageRank algorithm iteratively updates scores of the nodes in the network until the system's convergence or some preset number of iterations have been executed.

4.4 Text feature extraction and analysis

A user's position in a social network is not the only aspect that potentially drives how much reputation they accumulate. Especially on the Q&A portals where the primary goal is knowledge sharing, the content of the post itself can be highly indicative of a user's expertise and reputation. To account for this important contributor to the reputation score the analysis is also performed on the actual text content of the user submitted posts.

Text data in its raw form is highly unstructured, thus it is not ready to be used with different data analysis algorithms. An approach to deal with unstructured nature of textual data is to transform it into a vectorised attribute form. In other words, each text example in the data set is represented as a feature vector in a data matrix. To perform data vectorisation in this research project, the *Linguistic Inquiry and Word Count* (*LIWC*)[9] tool is used. This study utilised LIWC tool due to its ability to produce informative summary variables which can be used to describe the writing style of the user. Other possible techniques such as TF-IDF or POS tag extraction all require further processing to get informative text style representations (Kao and Wang, 2010; Maity et al., 2017). On the other hand, the LIWC tool accepts a body of text as an input and based on word frequency counts evaluates different aspects and metrics for the given text. The full list of the produced text output features is available in the tool's documentation[10].

LIWC produces a large number of various features, but most importantly for this project, the tool also provides four summary variables. These can be used to sum up all the insights provided by *LIWC* tool. The summary variables[11] are:

Analytical thinking This feature captures the degree to which people use words that suggest formal, logical, and hierarchical thinking patterns (Pennebaker et al., 2015).

Clout It measures a relative amount of confidence and social status that people display in their writing (Kacewicz et al., 2014).

Authenticity Measures to what degree the authors of a piece of text are more personal, humble, vulnerable and potentially reveal themselves in an authentic or honest way (Newman et al., 2003).

Emotional tone The feature indicates how positive or negative textual content is. Namely, the higher the feature value, the more positive the tone. Values below 50 suggest a more negative emotional tone (Cohn et al., 2004).

In addition to four main text attributes, three language metrics were also incorporated in this research. These include: the number of words per sentence, the number of long words (more than six letters), and the proportion of the

[9] http://liwc.wpengine.com/

[10] https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

[11] https://liwc.wpengine.com/interpreting-liwc-output/

used words that can be found in the LIWC dictionary (common words). The first summary attribute for the number of words per sentence indicates user's tendency to write short and simple sentences or longer, more complex sentences. The number of longer words further emphasises the notion of producing more complex text content. It is often the case that complex scientific topics require longer sentences and use long scientific terms. Lastly, highly specific terms used in certain science field are also possibly not covered by a general language dictionary. This lead to the inclusion of the third language attribute for the proportion of used dictionary words.

The users in the Q&A community and their corresponding reputation were the main point of interest in this research project. Consequently, the *LIWC* text features were produced for each of the users, based on all their textual content posted to a Q&A community. But before using *LIWC*, a text preprocessing steps were needed. StackExchange stores text content in their data dumps in the *HTML format*. Consequently, the post's text content had to be extracted to produce a clean text without any HTML tags. For this, the Python library *BeautifulSoup*[12] was used, which provided a simple programming API for specifying what parts of the HTML document should be extracted. The text extraction process included removal of the computer code listings and HTML tags. With the clean textual content extracted, all the text data for each user was concatenated into a single text body representing all the clean text produced by a certain user. LIWC was then run to produce text describing features for each of the registered users.

4.5 Statistical analysis

The results produced with the methods described in previous sections are subsequently analysed using different statistical techniques. This was used to gain a better understanding of the factors driving the user reputation in the online Q&A communities. The first step of the statistical analysis was to analyse each feature separately and evaluate its correlation with users' reputation scores. Then, the analysis was extended with latent variable modelling to model all the features together as a cohesive system which together influences the user's reputation.

[12]https://www.crummy.com/software/BeautifulSoup/

4.5.1 Variable correlation

As part of individual feature correlation analysis, the *Spearman rank* correlation and *Kendall's tau* were used. These two specific correlation measures were used because the examined data was not normally distributed. A popular Pearson correlation coefficient requires the data to be normally distributed. Spearman correlation rank (Procaci et al., 2016) and Kendall's tau (Field et al., 2012), on the other hand, do not have such limitations and can be used for non-normally distributed data.

To calculate the Spearman correlation rank, the data was first ranked and then Pearson correlation was calculated on produced ranks (Field et al., 2012). As indicated by Field et al. (2012), Kendall's tau often provides a better characterisation of the correlation than Spearman's rank. The procedure to calculate Kendall's tau is similar to that of the Spearman rank, where the data is again first ranked. Then, the number of concordant and discordant pairs are counted and are in turn both summed up respectively. Finally the following equation is used to calculate Kendall's tau: $(S_{con} - S_{dis})/(S_{con} + S_{dis})$, where S_{con} and S_{dis} stand for the sum of concordant and discordant pairs respectively (Kendall, 1938).

Both correlation ranks range from -1 to 1, where positive values mean that compared variables have an increasing trend or both variables have a decreasing trend. This means that two tested variables are strongly correlated. On the other hand, if the value of the correlation rank is negative, this indicates that when one variable's values are increasing, the other variable is decreasing in value, implying the negative correlation. The correlation rank value of 0 shows that there is no correlation between the two variables (Field et al., 2012).

4.5.2 Latent variable models

To gain a better understanding of which features are the most important for reputation accumulation we conducted a simple regression analysis. However, not all features were independent and did not directly affect users' reputation, but contribute as a part to a larger system of interconnected factors. This is the reason for a decision to utilise latent variable analysis. This analysis can be used to model interrelated influences of observed variables on the latent factors and in turn on the dependent variable, in this case user's reputation score.

To perform latent variable modelling in this research project, the R program-

ming language and its package *lavaan* (Rosseel, 2012) were used. When using latent variable models in lavaan, we first had to specify the model in question by providing the model describing formula with a similar syntax to that which was used for basic linear regression. However, there are several important additional features that can be used in lavaan models. Apart from the basic observed variable relationships, the user can also use reflective latent variables which are not directly observed but are related with several observed (or other latent) variables (Yong and Pearce, 2013). Such latent variables can also further influence other observed variables, meaning that a hierarchy of influence propagation can be built.

With the produced model in our research, its fit to the data had to be checked to indicate the reliability of the model. According to Levesque et al. (2004), several fit measures have to be inspected to measure the fit of a model. The first measure is *Root Mean Square Error of Approximation* (RMSEA) where the aim is to reduce this error measure as low as possible, however, the models with RMSEA lower than 0.05 threshold are acceptable. Apart from RMSEA, the *Comparative Fit Index* (CFI) and *Tucker-Lewis Index* (TLI) are also the measures that normally need to be evaluated to measure models performance. For both measures, the value needs to be above the threshold of 0.9 in order to be able to say that the model has a good fit (Levesque et al., 2004).

In the produced models with a satisfactory fit, the resulting variable coefficients can be interpreted as weights. These indicate how important different variables are. In this project, the interest is to examine how important and what are the effects of different observed and latent variables on user's reputation score.

Chapter 5

Results

This chapter presents the results of the conducted study. The primary interest of the analysis was the examination of the factors which contribute to users' reputation scores on Q&A platforms. User reputation scores in online Q&A communities generally follow a strong power-law distribution (Slag et al., 2015) where a large number of users have low reputation and only a handful of community members reach high reputation.

In the following sections, this chapter first starts by presenting the results of the analysis conducted on the CrossValidated Q&A dataset. The results first report how reputation was correlated with users' engagement, social position, and writing style of their posts. Then, the results of latent variable models, which integrated all three groups of factors to predict reputation, are presented. Second, the results of the same analysis on the Parenting dataset are reported.

5.1 User engagement

This section presents the results of the analysis into user engagement attributes and their correlations with the reputation score. The association of the user engagement attributes with the user reputation were first visually inspected and then by the calculation of correlation scores. The examined attributes were: user account maturity, answer/question ratio and a number of closed posts that the user produced.

The plot showing the connection between user account maturity and user reputation is presented in Figure 5.1a. It can be seen that high reputation accumulation period lasts for roughly 1000 days (2.5 years). All the users with high

reputation scores in the CrossValidated Q&A network had been the members of the community for at least 1000 days (at the time of the dataset creation). The topmost users had been members for even longer (2000 days or roughly 5 years) and some were even early adopters of the CrossValidated platform, meaning that they joined the community right at its inception back in 2010.

To gain further insights into the user engagement factors, the correlation analysis was performed (Table 5.1). User account maturity was found to have the largest positive correlation with the user reputation score in comparison to the other examined user engagement attributes. This observation was confirmed by both calculated correlation scores, which were furthermore in agreement for all the examined attributes and their corresponding correlation scores.

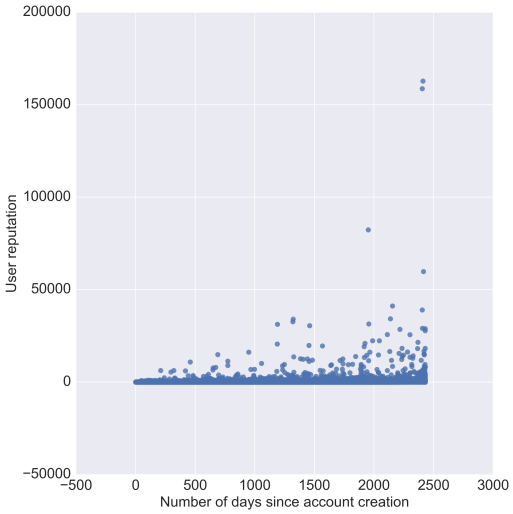
Table 5.1: User engagement attributes' correlation with user reputation.

User engagement type	Spearman rank	Kendall's tau
Account maturity	0.3618***	0.2499***
Log answer/question ratio	0.2790***	0.2207***
Number of closed posts	0.0656***	0.0542***

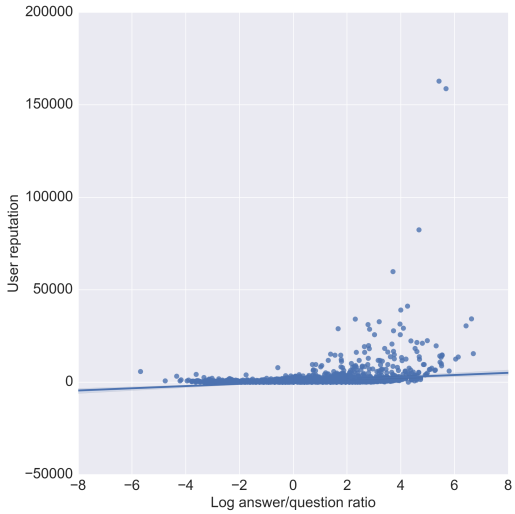
Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

The next factor that potentially influences users' reputation was the ratio between the number of posted questions and the number of questions that the user has answered (Bhanu and Chandra, 2016). In the case of users who asked more questions than produced answers, all their answer/questions ratios were concentrated just on the interval between 0 and 1. The ratios of the users with more answers than questions could range from 1 and into the infinity. To separate the former, smaller data points and make them more easily interpretable, the logarithm was applied to the ratio. Taking the logarithm helps spread the data initially in the 0 to 1 interval. The results of this analysis are presented in Figure 5.1b.

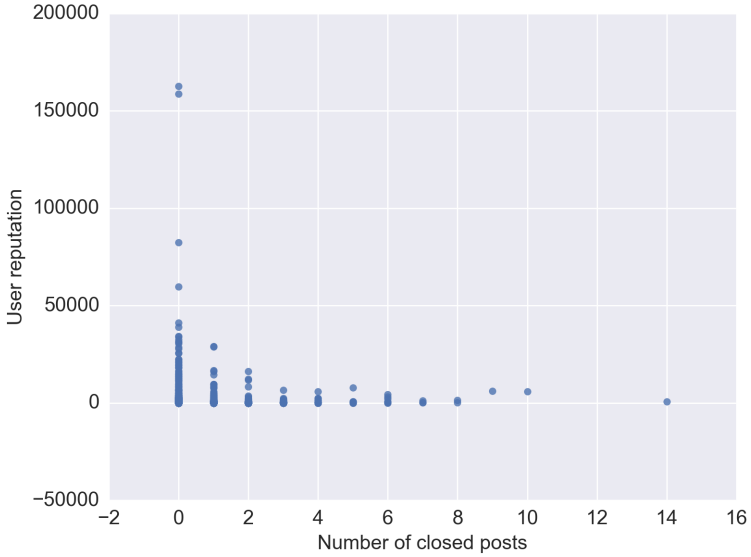
The plot shows that the users' reputation increased with the increase in answer/question ratio. Moreover, it is also indicative that users who wrote substantially more answers than questions were those who were the most reputable members of the Q&A community. On the other hand, there were also points representing users with the log answer/question ratio lower than 0, meaning that they had more questions than answers. Here, the shift in reputation was quite



(a) Number of days since user's account creation in relationship to user reputation.



(b) Log answer/question amount ratio in relationship to user reputation.



(c) Number of user's closed posts in relationship to user reputation.

Figure 5.1: User engagement attributes.

noticeable as none of the users with the log ratio lower than 0 reached higher reputation scores (more than 10,000) which can be observed for users that produce more answers than questions.

The visual inspection of answer/question ratio and its association with the user reputation was extended with the calculated correlation scores, presented in Table 5.1. The log ratio of the answer/question ratio has a slightly smaller correlation with the reputation than the previously examined user account maturity attribute. However, it can be still concluded, based on the results of the both used correlation scores, that the answer/question ratio was indeed positively associated with the user reputation.

The last examined attribute related to user's engagement is the number of closed posts that a certain user produced. The plot comparing the number of closed posts to the reputation is presented in Figure 5.1c. It indicates that the most reputable users did not have any closed posts. On the other hand, the resulting plot also suggests that users who accumulated a large number of closed posts were generally the less reputable ones. However, the results in Table 5.1 show that the number of closed posts had only a barely positive correlation which was extremely close to the turning point. This is especially noticeable when compared to the other two attributes and their correlation measures. Consequently, this small positive correlation could possibly be attributed to the potential outliers and the fact that the correlation was separately calculated for each respective attribute.

5.2 Network analysis

With the user engagement results examined, this subsection turns the attention to the interactions between users and the corresponding network analysis. The social network analysis is performed to gain an understanding how the position of the users in the network is related to their reputation. The users' social positions indicated by the network analysis attributes and their association to the user reputation was again first plotted. After that, the results were further extended by performing the correlation analysis which produced more concrete insights.

The results of the degree centrality analysis are plotted in Figure 5.2a, where they are compared to the user reputation score. From the presented plot it can be seen that the users' node degree had a strong positive relationship with the

reputation. This indicates that the volume of the posts written had an important role when it comes to the reputation accumulation. Nearly all of the most reputable users wrote at least 500 posts, while, as it shown in Table 4.1, the average number of posts per user was only 3.789. Thus the results show that the most reputable users wrote considerably more posts than the Q&A network's average.

After the initial visual inspection of the results, a further analysis of the user social position attributes was performed. The produced results are presented in Table 5.2. It can be seen that both used correlation measures agree on the rank of the tested network attributes' correlation strength. Specifically, degree centrality had the highest correlation with the user reputation, again confirming the observations from the plots.

The resulting betweenness centrality comparison to the user reputation is plotted in Figure 5.2b. A certain degree of linear association could again be observed. Additionally, there were some outliers from this general trend, where reputable users have a slightly lower betweenness score, or on the other hand, a number of less reputable users have a slightly larger betweenness than the rest of the population. However, even with the outliers, a correlation between the betweenness centrality and the user reputation was visible. This conclusion is also confirmed by linear regression line that was fitted and shown in the plot to indicate a positive association between the two examined variables.

Further correlation analysis of betweenness centrality is presented in Table 5.2. This result is complementary to the visual inspection of the association with the reputation score presented above. The correlation measures confirm that the betweenness centrality scores were slightly less well associated with the user reputation than the degree centrality.

Lastly, the resulting plot comparing the PageRank score to the user reputation score is presented in Figure 5.2c. It can be seen that PageRank also achieved a positive linear association with the reputation score. Especially in the case of the highly reputable users, their PageRank scores were associated with the reputation. Apart from that, there were also some outliers, especially those with lower reputation than expected, but with high PageRank score. However, it can be observed from both the plotted user points and the fitted regression line that the user reputation was positively associated with the PageRank score.

To conclude the network analysis, the correlations of the PageRank score with the user reputation were calculated. The results are again presented in Table

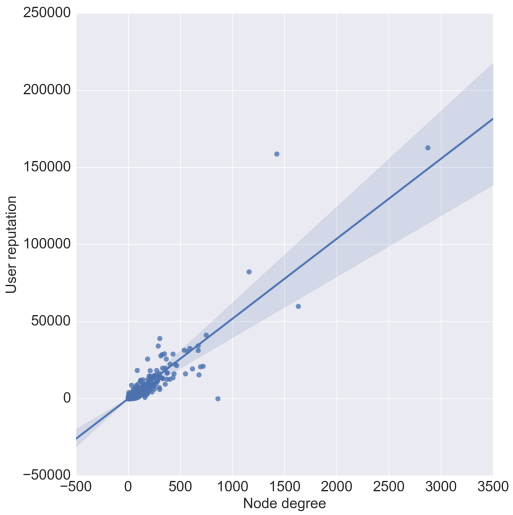
5.2. The PageRank score was found to be even slightly less correlated with the user reputation score than the previously discussed betweenness centrality. This makes the PageRank score the least correlated network analysis measure which was utilised in this study.

However, by looking at all the presented correlations in Table 5.2 it can be seen that network attributes were overall more positively correlated with the reputation than the user engagement attributes (table 5.1). There, for example, according to Kendall's tau, the strongest correlation was found for the user's account maturity with the value of 0.2499. This is just roughly higher than the smallest correlation with the reputation found for the related network metric PageRank with the value of 0.2364.

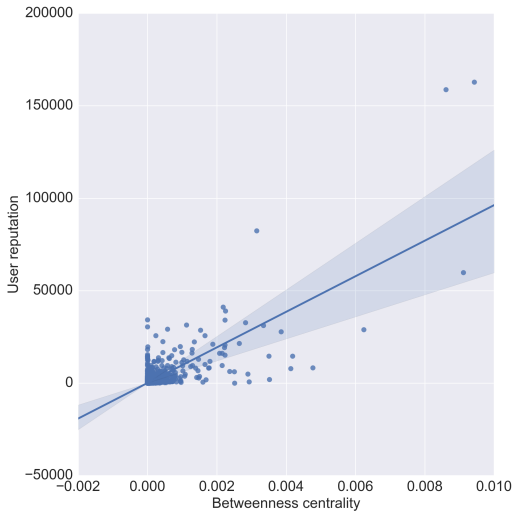
Table 5.2: Network analysis metrics' correlation with user reputation.

Network analysis metrics	Spearman rank	Kendall's tau
Degree centrality	0.5608^{***}	0.4533^{***}
Betweenness centrality	0.3685^{***}	0.3037^{***}
PageRank	0.3043^{***}	0.2364^{***}

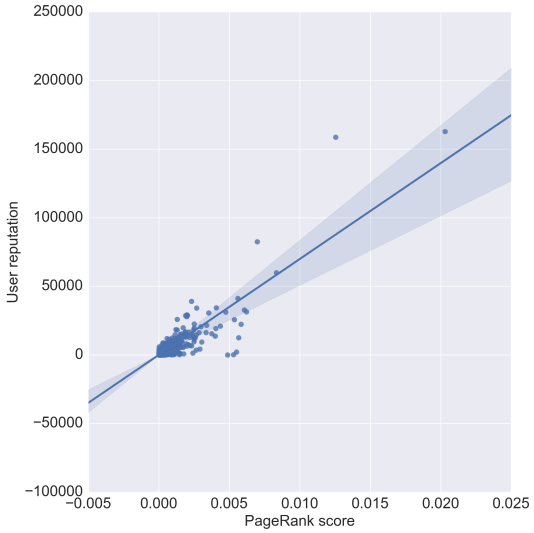
Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$



(a) User node degree centrality in relationship to user reputation.



(b) User betweenness centrality in relationship to user reputation.



(c) User PageRank score in relationship to user reputation.

Figure 5.2: Social network analysis results.

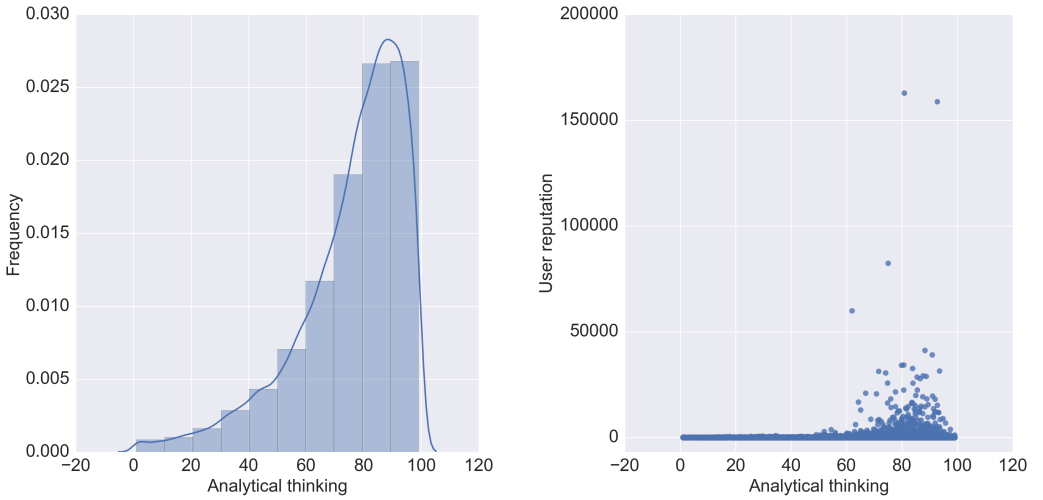
5.3 Text analysis

This section presents the results of the analysis inspecting textual content posted by the users of the Q&A community. The text related factors were explored by plotting and by calculating correlation scores to evaluate association with the user reputation. It can be immediately observed that in the case of text analysis (Table 5.3), the correlations were substantially lower than those reported in the previous sections (Tables 5.1 and 5.2). While the best performing correlations in the previous two sections had values of at least 0.30, in the case of text correlations, these are barely larger than 0.0.

The observations in Table 5.3 can be explained with the results presented in the following subsections where the text features will be examined with plots (Figures 5.3, 5.4, 5.5, and 5.6). The plots show that the textual features did not have a completely linear relationship with the reputation. They rather had a positive relationship with the reputation only up until some certain point after which with a further increasing values reputation started to decrease. This, in turn, had a diminishing effect on the overall correlation of the selected text features when compared to the reputation.

The first feature extracted from text was the notion of the analytical thinking found in the user produced posts. Since the analysed Q&A network CrossValidated mainly deals with the technology related topics, it was expected that the user produced texts would show some degree of technical, analytical writing. The plot of users' analytical writing score compared to the reputation is presented in Figure 5.3. It can be seen from the subplot (a), showing the distribution of the analytical thinking scores, that the vast majority of users produced textual content with the analytical thinking scores in the upper range of a score's value interval. In the subplot, (b) the same pattern can be also observed for the reputable users. All of the most reputable users were also concentrated in the same value range which was the average for the whole population. Thus, the plots indicated that reputable users do not particularly deviate from the analytical thinking score which is average for all the members of the investigated Q&A community.

To better outline the idea of analytical thinking score we present two actual example texts. First, the following text was assigned a low analytical thinking score:



(a) Analytical thinking score distribution.

(b) Analytical thinking score in relationship with reputation.

Figure 5.3: User text content analytical thinking score results.

> "So the title is a bit vague, but I am trying to solve some basic probability inequality. Given that CODE, how can I show that CODE? I started by showing that CODE, but I am not sure if this helps. Thoughts? Thanks."

On the other hand, this example represents the textual content, which scored high with regards to the analytical thinking:

> "For a design study, I have protocols from 5 different design groups with two persons in each team. The verbal utterance of each group (two persons) are transcribed for each sentence. The data are derived from each line of transcription for the analysis. I want to analyze two cases: I want to compare the differences between two persons in different groups. I want to compare different groups functions.
> Which types of analysis should I take for the two above?"

Lastly, the correlation scores were calculated to measure the attribute's association with the user reputation. It can be seen from Table 5.3 that analytical thinking style had a mild positive correlation with the user reputation. In fact, from the examined LIWC summary features, the analytical thinking showed the smallest positive correlation with the user reputation.

The next examined measure was for clout, which indicates how confident the

writer seems to be based on his or her text. Here, the differences between the most reputable users and the rest of the community could be observed. Subplot (a) in Figure 5.4 presents the clout score distribution for the complete user population. It shows that most of the users in the community had clout scores concentrated in the lower range of the clout score range. On the other hand, the subplot (b), presents the relationship between the clout score and user's reputation. Here, it can be seen that most of the reputable users had their clout scores skewed towards higher clout scores. Reputable users are shown to have had their clout scores in the medium range instead of the lower range as it was true for the whole user population.

For a better understanding of the text clout score, we again present two actual example texts. First, the following text was assigned a low clout score:

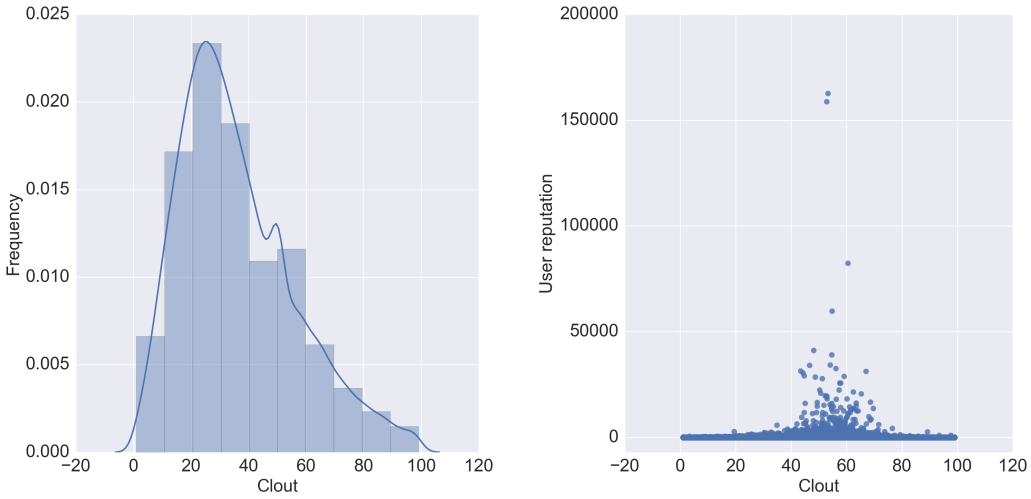
> "I have a set of data in which the response variable is continuous and the other independent variables are dichotomous (not naturally occurring). I think I have to perform Biserial correlation (I an aware of the difference between point biserial and biserial correlation). How can I do this on SPSS or any other statistical software?"

On the other hand, this example represents the text content, which received high value with regards to the clout score:

> "Do you have identical rows in you input matrix? If yes, try to remove them, e.g. by adding a small noise."

The clout score's association with the user reputation is further examined by calculating correlation scores. According to the results presented in Table 5.3, the clout score exhibits a considerable positive correlation with the user reputation. Furthermore, the clout score or in other words confidence of the user's writing has the strongest positive correlation when compared to other tested text summary attributes.

Next, the analysis of the authenticity score was performed. This score indicates how authentic and personal the style of the text author is. Keeping in mind that the analysed Q&A network deals with the science related topics, it can be seen that this score was quite low. Namely, the scientific writing style is often very objective and does not show signs of the writer's personal authenticity. The plots in Figure 5.5 confirm this thinking. Subplot (a) shows that the overall dis-



(a) Clout score distribution.

(b) Clout score in relationship with reputation.

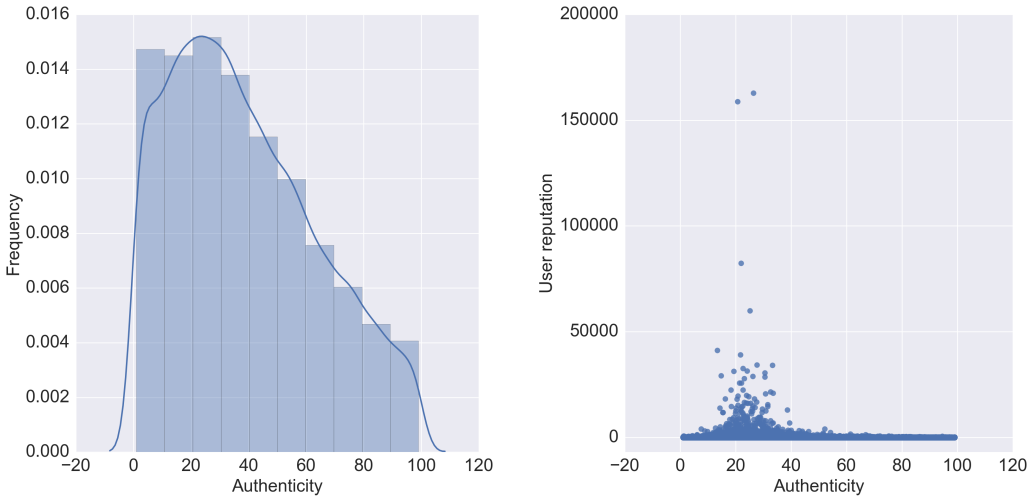
Figure 5.4: User text content clout score results.

tribution of the authenticity score in the CrossValidated Q&A network tends to be quite low. The same is true for the more reputable users (subplot (b)) of the network, which did not deviate from the network average with the authenticity scores in the lower range.

For a better understanding of the text authenticity score, we again present two actual example texts. First, the following text was assigned a low authenticity score:

> "Why would the results of the ANOVA be non-significant, while a pairwise comparison using Tukey's Wholly Significant Difference (WSD) is significant? Is their a general pattern in the means of the data that would typically produce this result?"

On the other hand, this example represents the text content, which received high value with regards to the authenticity score:



(a) Authenticity score distribution.

(b) Authenticity score in relationship with reputation.

Figure 5.5: User text content authenticity score results.

> "I am an absolute beginner in field of machine learning, I started doing titanic assignment in Kaggle and found(read some where) Random Forest is the best fit. I started reading about random forest and found the Explanation by Edwin Chen in this question intuitive. This made me "understand" how I can solve the Titanic assignment which predicts if one survives or not(classification). But I cannot understand How random Forest will work for regression which is continuous. Please don't mind to point out any mistakes in my assumptions or the way I started things. Any advice would be helpful, This looks very vast and Don't even know where to begin."

The authentic style of writing was further examined by performing the correlation analysis. The results in Table 5.3 indicate a negative correlation between user reputation and the authenticity. Furthermore, it can be observed that authentic style of writing was the only found text summary attribute negatively associated with the user reputation.

The last of the analysed LIWC summary feature was the emotional tone score. It indicated how positive or negative the text content is. The results of this analysis are presented in Figure 5.6. Here the emotional scores presented in subplot (b) indicate that the most reputable users generally aligned with the overall average of the whole user population presented in the subplot (a). The

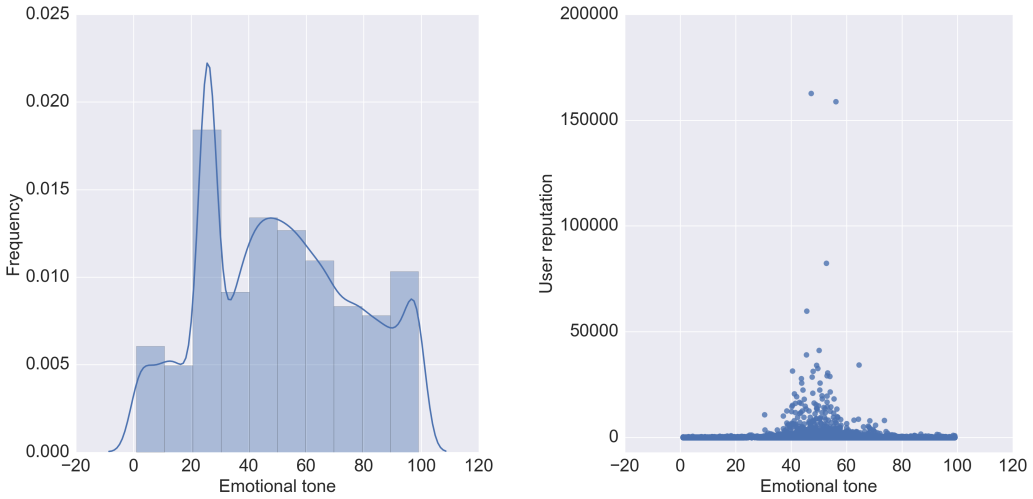
vast majority of users produced text content with emotional scores in the medium range of the score interval. The outliers visible in the subplot (a) at the emotional tone score of 25.77 could be explained by looking at the actual text content these users produced. It turned out that these were the users which have posted only a small number of posts. Moreover, their posts were very short as they often included only a block of code and a short side note. As LIWC can not work on the code block content and they were in turn removed prior the analysis, the remaining text content for these users was very short. This was probably problematic for LIWC tool as it tried to calculate the scores on a body of text which was too short.

Overall, it can be seen that the vast majority of the users wrote in a neutral tone with the most of the user population's scores around the 50 mark. More specifically the median emotional score was 49.53 and the mean was equal to 50.84. Lastly, turning the attention to the most reputable users, it has been demonstrated that they did not deviate from the population's average and they wrote in an emotion neutral style.

For a better understanding of the text emotional tone score, we again present two actual example texts. First, the following text was assigned a low emotional tone score:

> "My spouse frequently works with (expensive, hard to obtain) data samples; for example route information for commuting bicyclists collected using a smartphone app. More often than not, these samples suffer from some kind of known demographic over-representation that they'd like correct for various applications. Mindful of Karl Roves' and friends "corrections" to "obvious" democrat oversampling for the Nov. 2012 election polls which led to rather embarrassingly incorrect predictions, is there any theoretically appropriate way of doing this? I'm not even sure what to call what I'm looking for – is this what in some places is called reject inference"

On the other hand, this example represents the text content, which received high value with regards to the emotional tone score:



(a) Emotional tone score distribution.

(b) Emotional tone score in relationship with reputation.

Figure 5.6: User text emotional tone results.

> "Lets assume we have a dependent varible which can take on three values: 1, 2 and 3. Is there any differences in running multiple binary logit regressions(ie. 1 vs 2 and 2 vs 3) or the multinomial logit regressions with 2 as the base group? Will the results be the same as the base group is the same in both cases? I've tried to look it up but can't seem to find an answer. Thanks in advantage, have a nice weekend. Regards, Martin"

Lastly, the results in Table 5.3 indicate that the emotional tone was overall second most positively correlated examined summary text feature. There could be seen a decrease of about 0.1 in the correlation when compared to the previously presented summary feature for clout. However, emotional tone, none-the-less, had a slightly positive and second largest correlation with the user reputation.

Apart from the examined main four text summary attributes, three additional language metrics were extracted. The final correlation analysis results for these support text features are presented in Table 5.3. These indicated a slight positive correlation in the case of words per sentence number and ratio of the six letter words. Especially, the variable for the number of words per sentence showed relatively high correlation when compared to other text related features. Lastly, the ratio of words found in LIWC dictionary showed an extremely small negative correlation and furthermore strongly insignificant result, thus it could not be

accepted as necessarily correct. All the other presented results in Table 5.3 showed a really small *p-value* (below 0.001) and consequently strong significance of the results.

Table 5.3: Network analysis metrics' correlation with user reputation.

Network analysis metrics	Spearman rank	Kendall's tau
Analytical thinking	0.0188^{***}	0.0131^{***}
Clout	0.1365^{***}	0.0955^{***}
Authenticity	-0.0453^{***}	-0.0330^{***}
Emotional tone	0.0237^{***}	0.0161^{***}
Words/sentence	0.1100^{***}	0.0749^{***}
Six letter words	0.0502^{***}	0.0344^{***}
Dictionary words	-0.0002	-0.0005

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

5.4 Latent variable model analysis

As part of latent variable regression analysis the following models were built and evaluated:

- Model with two latent factors: social position factor and text contribution quality factor.
- Model with three latent factors: social position factor, text contribution quality factor and user engagement factor.
- Hierarchical model with two main, first order latent factors and one second order latent factor: first order social position and user engagement latent factors and text contribution quality latent factor as part of the main social position factor.

According to the presented background section, the online user reputation is not solely influenced by a single user describing factor. Reputation accumulation in this study has been based on three main factors: user's social position, engagement, and text quality. This motivated the construction of the three latent variable models presented here.

The first, two-latent factor model serves the role of the simple baseline, Namely, similar to Kao and Wang (2010), the user reputation is based on social position and user text content quality which are possibly primal indicators of reputation. In the subsequent two models, the user engagement factor was also added to explore its effect on the reputation. Furthermore, first two models had the latent variables directly influencing the reputation. However, in the last model the indirect contribution of the post quality on the reputation captured by constructing a hierarchical model.

5.4.1 Two-factor model for social position and text contribution quality

The first and simplest analysed model used only 2 latent factors, one for user's social position indicated by social network analysis metric and another for text contribution quality factor based on LIWC features. The diagram of used model is presented in Figure 5.7.

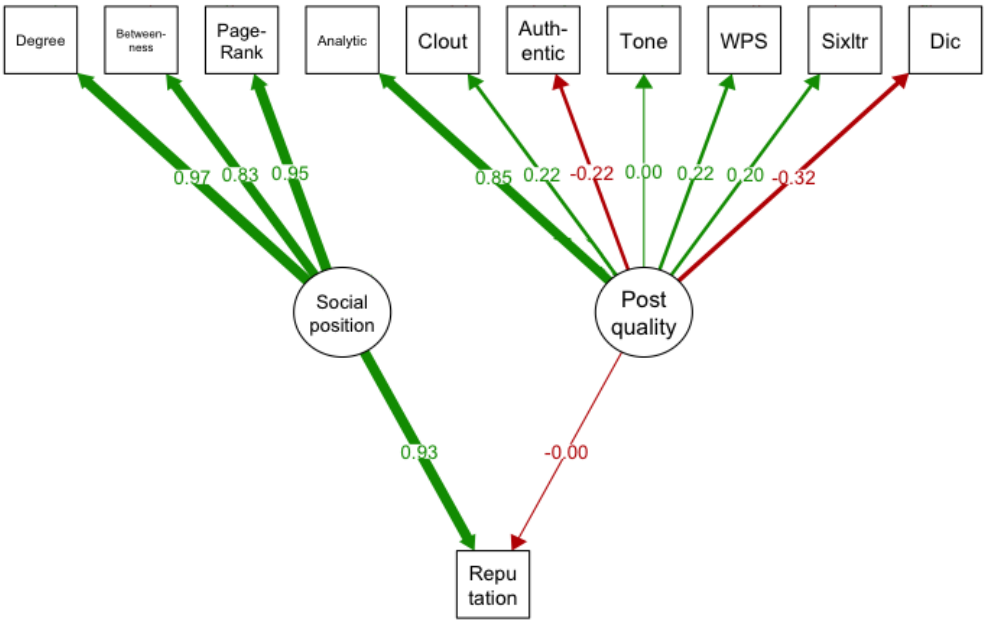


Figure 5.7: Path diagram for the model with two latent factors.

The specified two-factor model was then fitted to the data. By doing so the attribute coefficients were evaluated. These indicated how important a certain variable is for the prediction of the user reputation. The results for the utilised

two-factor model are presented in Table 5.4. The note at the bottom of the table presents the evaluated fit measures which indicate how well the model fits the data and conversely what is the quality of the model. All three model performance measures met the threshold values presented in the methods chapter. It can thus be concluded that the model has a good fit.

The result analysis starts with the importance of the observed variables produced in the social network analysis. It can be seen that node degree centrality had the strongest influence on the user's social position. Next, somewhat different to the network measure's direct correlation with the reputation, the PageRank was second most influential feature for the user's social position. The betweenness centrality had the smallest coefficient when it came to predicting social position.

Next, the LIWC text features and their influence on the post text content quality were examined. It can be seen that almost all text features had quite small coefficients when predicting the latent post quality factor. The only exception was analytical writing feature, which was comparable in coefficient value to those found for social position and was substantially larger than other text features. Conversely, the analytical writing style was the most influential for the post quality in the analysed Q&A community. Next followed clout, words/sentence and the amount of six letter words which all had a comparable influence on the positive text quality. A really small influence on the text quality was found for emotional tone. However, as it is indicated by the high *p-value*, this result is not significant and can thus not be trusted. The last remaining text features for authentic style of writing and use of dictionary words had both negative influence on the post quality.

Having described the results for the observed variables, the attention is now given to the coefficients corresponding to the two latent variables. It can be seen that user's social position factor had a large and dominant coefficient value for the prediction of the user reputation score. Post quality latent variable, on the other hand, had a really small negative coefficient. However, upon closer examination, it could be seen that the *p-value* for this result was quite large indicating that the result might not be significant and thus not to be trusted.

Table 5.4: Results for the latent model with two latent factors.

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	reputation	~	social position	0.934	0.00	235.60	0.00	0.93	0.94
2	reputation	~	post quality	-0.004	0.00	-1.58	0.11	-0.01	0.00
3	social position	=~	degree	0.966	0.00	249.98	0.00	0.96	0.97
4	social position	=~	betweenness	0.831	0.00	195.85	0.00	0.82	0.84
5	social position	=~	pageRank	0.947	0.00	236.52	0.00	0.94	0.95
6	post quality	=~	Analytic	0.847	0.02	38.85	0.00	0.80	0.89
7	post quality	=~	Clout	0.218	0.01	27.72	0.00	0.20	0.23
8	post quality	=~	Authentic	-0.215	0.01	-27.51	0.00	-0.23	-0.20
9	post quality	=~	Tone	0.002	0.01	0.36	0.72	-0.01	0.01
10	post quality	=~	WPS	0.216	0.01	27.28	0.00	0.20	0.23
11	post quality	=~	Sixltr	0.202	0.01	27.00	0.00	0.19	0.22
12	post quality	=~	Dic	-0.322	0.01	-32.95	0.00	-0.34	-0.30

Model fit: $RMSEA = 0.017$; $CFI = 0.998$; $TLI = 0.997$

5.4.2 Three-factor model for social position, text contribution quality and user engagement

The next SEM model added an additional latent factor for user engagement and related observed variables. The diagram of this model is presented in Figure 5.8.

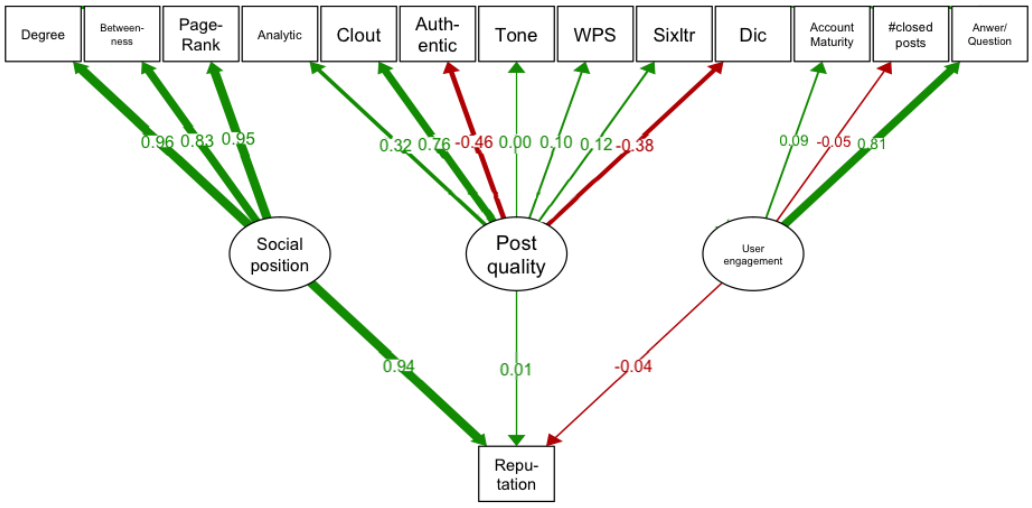


Figure 5.8: Path diagram for the model with three first order latent factors.

The explained model was fitted to the data in the same way as the previous model. The results are presented in Table 5.5. As for the previous latent model, the note at the bottom of the table presents model performance measures which again indicate a well-fitted model.

Observed variables indicating the social position showed similar coefficient values and thus importance to those analysed before in the two-factor model (Table 5.4). Namely, the node degree centrality again had the highest influence on the user's social position. Next, closely followed PageRank and the least influential was the betweenness centrality. Although again, as it was also the case in the two-factor model, all of the social position observed variables had coefficients higher than any other observed variable that was included in this model.

In the case of the post quality latent variable, it can be seen that its corresponding observed variables retained the coefficients similar as before. In addition to similar coefficient size ranking of features, a high *p-value* and thus insignificant result can again be seen for the emotional tone variable. However, there are a few differences that need to be pointed out. First and most obvious is that in this model analytical writing feature was not the most important as it was in the two-factor model. Here it was in the second place, after the clout feature which had the largest coefficient. Another prominent change was in the case of features with negative coefficients, thus negative influence. These features were again authentic writing style and dictionary words. However, the values of their coefficients were even more negative. This was especially true for the authentic style of writing which had the negative coefficient value of almost double of that in the two-factor model. According to this, more extensive and detailed model, the authentic writing had an even more negative impact on the text post quality.

The new addition to this three latent factor model was user engagement as a latent factor and its corresponding observed variables. It can be seen that log answer/question number ratio had by far the largest coefficient value. Second, with a substantially smaller coefficient value was the user account maturity. Last, with the reciprocal influence indicated by the negative coefficient was the number of closed posts. It is also worth noting, that all the fitted coefficients for the observed variables corresponding to the user engagement factor were significant results indicated by the *p-values* below the threshold of 0.001.

Lastly, the top level (latent) factor variables and their importance for the

user reputation are presented. The same as in the previous factor model, the most important latent variable with the largest evaluated coefficient was for the user's social position in the network. Next followed user's text post quality factor with a small positive coefficient. This result also had a slightly increased *p-value* meaning a slightly diminished significance of this specific result. Finally, the user engagement latent variable got a slightly negative coefficient indicating its slightly negative influence effect on the user reputation score. However, similarly to the result for the two-factor model, the latent factor for user's social position had really a prominent influence on the reputation. Social position factor had a variable coefficient substantially larger than the other two evaluated latent coefficients.

Table 5.5: Results for the latent model with three first order latent factors.

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	reputation	~	social position	0.941	0.00	229.98	0.00	0.93	0.95
2	reputation	~	post quality	0.015	0.01	2.25	0.02	0.00	0.03
3	reputation	~	user engagement	-0.039	0.01	-5.59	0.00	-0.05	-0.03
4	social position	=~	degree	0.960	0.00	250.16	0.00	0.95	0.97
5	social position	=~	betweenness	0.829	0.00	195.10	0.00	0.82	0.84
6	social position	=~	pageRank	0.948	0.00	236.64	0.00	0.94	0.96
7	post quality	=~	Analytic	0.318	0.01	50.31	0.00	0.31	0.33
8	post quality	=~	Clout	0.755	0.01	76.17	0.00	0.74	0.77
9	post quality	=~	Authentic	-0.460	0.01	-51.38	0.00	-0.48	-0.44
10	post quality	=~	Tone	0.001	0.01	0.13	0.90	-0.01	0.01
11	post quality	=~	WPS	0.102	0.01	16.22	0.00	0.09	0.11
12	post quality	=~	Sixltr	0.116	0.01	19.06	0.00	0.10	0.13
13	post quality	=~	Dic	-0.379	0.01	-43.29	0.00	-0.40	-0.36
14	user engagement	=~	day_since_cre	0.088	0.01	13.77	0.00	0.08	0.10
15	user engagement	=~	num_closed_posts	-0.048	0.01	-7.87	0.00	-0.06	-0.04
16	user engagement	=~	log_answ_quest_ratio	0.814	0.03	25.08	0.00	0.75	0.88

Model fit: $RMSEA = 0.04742$; $CFI = 0.97620$; $TLI = 0.96390$

5.4.3 Hierarchical three-factor model

The last type of the latent factor model was the hierarchical model with two first-order factors for social position and user engagement and one second-order for the post quality. The second order text post quality factor did not influence user reputation directly, but indirectly as part of user's social position. The described

model's diagram is presented in Figure 5.9.

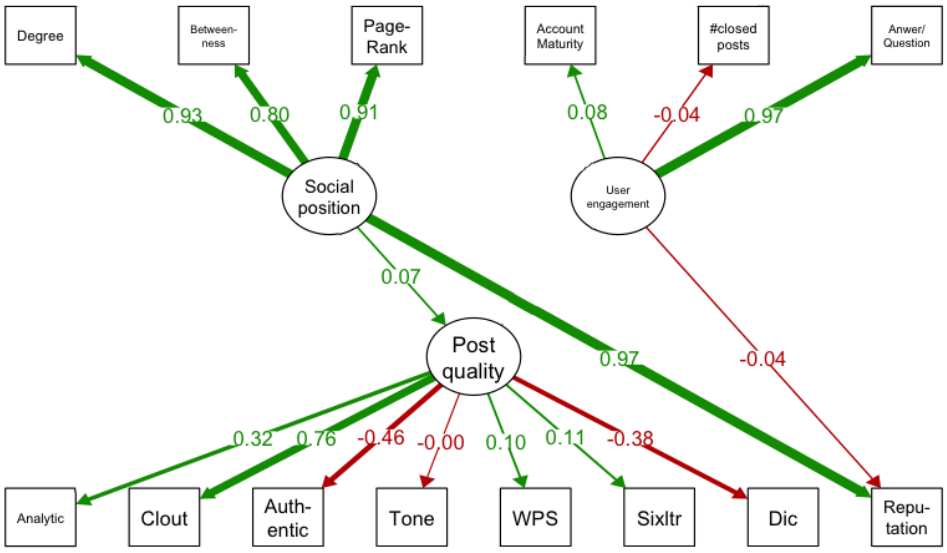


Figure 5.9: Path diagram for the hierarchical model with two first order latent factors and one second order latent factor.

The above-specified model was fitted to the data and the resulting coefficients are presented in Table 5.6. Again, the note at the bottom of the table presents model performance measures which indicate a well-fitted model.

First, the variables forming the social position latent variable are examined. It can be seen from the table that initial three observed network analysis variables stayed more or less the same as in the previously introduced models. The added second order post quality latent variable had a really small coefficient when compared to other three network analysis variables.

When looking at post quality measures, it can be seen that clout feature had again the largest positive coefficient. The table shows that the second largest positive coefficient was attributed to the analytical writing style feature. This was followed by features of emotional tone, words/sentence and the amount of six letter words which have substantially lower coefficients. Here, it must be again noted that a coefficient of 0.0 for the emotional tone feature is not a significant result which can be completely trusted, due to its high *p-value*. Apart from the presented text features with the positive coefficients, the evaluated model again produced negative coefficients for authentic writing style and high use of LIWC

dictionary words. Moreover, when compared merely by the absolute value of the coefficient, it can be seen that these two features had a really strong negative influence on the post quality. The authentic writing style, for example, had the second largest absolute coefficient value when compared to other features extracted from the text. The observed results are in agreement with what has also been seen for the previous model in Table 5.5.

The remaining group of observed variables formed the user engagement latent factor. The log answer/question ratio was again the feature with the largest coefficient value which turned out to be substantially larger than the other user engagement features. The number of days since account creation showed only a small positive influence on the user engagement factor and the number of closed posts variable's negative coefficient indicated negative influence.

With the second order variables' results presented, the attention is now given to the top level factor variables. The latent variable representing the user's social position again had by far the largest positive coefficient value. On the other hand, the other latent variable influencing the user reputation is user engagement, which got a small negative coefficient.

Table 5.6: Results for hierarchical model with two first order latent factors and one second order latent factor.

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	reputation	~	social position	0.974	0.00	246.21	0.00	0.97	0.98
2	reputation	~	user engagement	-0.038	0.00	-14.42	0.00	-0.04	-0.03
3	social position	=~	degree	0.927	0.00	233.32	0.00	0.92	0.94
4	social position	=~	betweenness	0.800	0.00	184.30	0.00	0.79	0.81
5	social position	=~	pageRank	0.915	0.00	225.26	0.00	0.91	0.92
6	social position	=~	post quality	0.075	0.01	11.40	0.00	0.06	0.09
7	post quality	=~	Analytic	0.317	0.01	50.30	0.00	0.31	0.33
8	post quality	=~	Clout	0.757	0.01	75.57	0.00	0.74	0.78
9	post quality	=~	Authentic	-0.460	0.01	-51.10	0.00	-0.48	-0.44
10	post quality	=~	Tone	0.000	0.01	0.00	1.00	-0.01	0.01
11	post quality	=~	WPS	0.101	0.01	16.12	0.00	0.09	0.11
12	post quality	=~	Sixltr	0.115	0.01	18.89	0.00	0.10	0.13
13	post quality	=~	Dic	-0.384	0.01	-43.60	0.00	-0.40	-0.37
14	user engagement	=~	day_since_cre	0.076	0.01	12.93	0.00	0.06	0.09
15	user engagement	=~	num_closed_posts	-0.041	0.01	-7.61	0.00	-0.05	-0.03
16	user engagement	=~	log_answ_quest_ratio	0.972	0.04	27.46	0.00	0.90	1.04

Model fit: $RMSEA = 0.04723$; $CFI = 0.97639$; $TLI = 0.96420$

5.5 Comparison with the non-technical network

After the results for CrossValidated Q&A network were produced, the analysis was extended to a non-technical Q&A network. The network selected for this further analysis was the Parenting Q&A community, also operated by the Stack-Exchange organisation. This ensured the comparability of the results since both networks have exactly the same format. The same analysis as in the case of CrossValidated was performed on this new network and corresponding dataset.

Here, only the results of the last best fitting latent model (based on *RMSEA*, *CFI* and *TLI*) are presented and compared to the results previously examined for CrossValidated network (Table 5.6). The model had a hierarchical structure with two first order latent factors, for social position and user engagement, and one second order latent factor for post quality. The described structure was the same as had been used for the last presented CrossValidated model. The model for Parenting Q&A community is depicted in Figure 5.10.

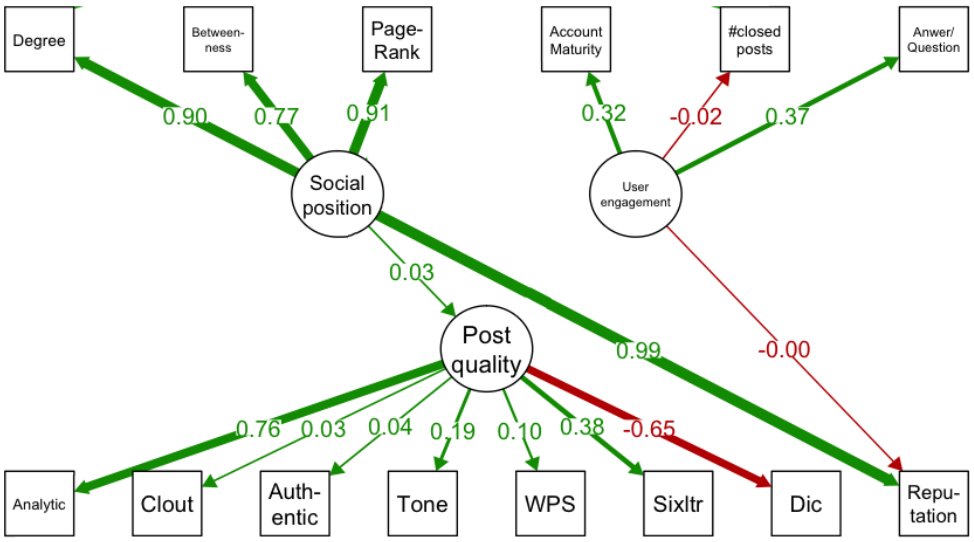


Figure 5.10: Path diagram for latent factor analysis of the Parenting Q&A network.

The resulting coefficients of the fitted model for the Parenting network are shown in Table 5.7. First, the fitted coefficients for observed variables forming user's social position latent variable are presented. It can be seen that the ranking of the variables according to their coefficient size had slightly changed. The node degree centrality did not have the largest positive coefficient as in the models for

CrossValidated network. For Parenting network, the PageRank had a slightly larger coefficient than the node degree centrality. The betweenness centrality was however again in the third place when looking at the size of the fitted coefficient. The latent variable for post quality which also contributed to the user's social position had a positive coefficient, but was, when compared to the other mentioned variables, very small in value.

Next, textual variables forming the post quality latent variable were examined. The largest coefficient was assigned to the analytical writing style. On the other hand, a sharp decrease in the coefficient value could be seen for the clout text feature, which normally had one of the highest coefficient values in the CrossValidated network. It must be noted that the evaluated result had a slightly increased *p-vale*, indicative of statistical insignificance of the link between Clout and the latent factor. However, the most interesting change in fitted coefficients could be seen for the authentic writing style. This feature consistently has had a negative coefficient value in the CrossValidated network but was positive when evaluated on Parenting network dataset.

The attention is now given to the resulting coefficients for observed variables forming the user engagement latent factor. As in the case of CrossValidated network, the log answer/question number ratio had again the largest positive coefficient. Next followed the account maturity which got only a slightly lower coefficient, where in the previously analysed network, the difference was quite significant. As before, the number of closed posts feature produced a small negative coefficient. However this time, for the Parenting network, the result had a high *p-value* and is thus insignificant.

Lastly, the top level latent variables influencing the user reputation itself are examined. The results for the Parenting network were quite similar to those seen before in the CrossValidated Q&A community. The user's social position latent variable was dominant with a comparably high positive coefficient. On the other hand, the user engagement latent factor produced a slightly negative coefficient value, which was comparable with the results seen in the previously analysed CrossValidated Q&A network. However, it must be noted that due to the high *p-value* this result is not completely significant.

Table 5.7: Results for the Parenting Q&A portal using hierarchical model with two first order latent factors and one second order latent factor.

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	reputation	~	social position	0.990	0.01	95.31	0.00	0.97	1.01
2	reputation	~	user engagement	-0.004	0.01	-0.45	0.65	-0.02	0.01
3	social position	=~	degree	0.891	0.01	86.74	0.00	0.87	0.91
4	social position	=~	betweenness	0.766	0.01	67.59	0.00	0.74	0.79
5	social position	=~	pageRank	0.907	0.01	86.68	0.00	0.89	0.93
6	social position	=~	post quality	0.035	0.02	2.11	0.04	0.00	0.07
7	post quality	=~	Analytic	0.757	0.02	41.35	0.00	0.72	0.79
8	post quality	=~	Clout	0.030	0.02	1.82	0.07	-0.00	0.06
9	post quality	=~	Authentic	0.041	0.02	2.44	0.01	0.01	0.07
10	post quality	=~	Tone	0.188	0.02	11.19	0.00	0.16	0.22
11	post quality	=~	WPS	0.104	0.02	6.06	0.00	0.07	0.14
12	post quality	=~	Sixltr	0.380	0.02	24.97	0.00	0.35	0.41
13	post quality	=~	Dic	-0.653	0.02	-37.29	0.00	-0.69	-0.62
14	user engagement	=~	day_since_cre	0.317	0.02	13.88	0.00	0.27	0.36
15	user engagement	=~	num_closed_posts	-0.017	0.02	-0.92	0.36	-0.05	0.02
16	user engagement	=~	log_answ_quest_ratio	0.368	0.03	14.53	0.00	0.32	0.42

Model fit: $RMSEA = 0.059$; $CFI = 0.963$; $TLI = 0.944$

Chapter 6

Discussion

The study presented in this thesis was conducted to answer two research questions. The first research question focused on analysing different factors governing the user reputation accumulation in the online Q&A communities, more concretely, CrossValidated. Accordingly, the user's social position, engagement, and writing quality were examined for their relationship with the user reputation. The second research question, on the other hand, was interested in validating the extent to which the results from the first research question also apply in a different context. To produce the necessary comparison, another Q&A community was analysed for reputation governing factors. The focus of this study was a Q&A community which was focused on the non-science related topics.

6.1 Research question 1: the factors governing user reputation

To answer the first research question, the focus is first given to the user engagement and its influence on the accumulated user reputation. The results showed that the users need to be active members of the Q&A community for a longer period of time to be able to accumulate a substantial amount of reputation (Figure 5.1a and Tables 5.1, 5.4, 5.5, and 5.6). All the most reputable users have been members longer than 1000 days, most of them even more than 2000 days. A completely new user, no matter how brilliant in their contributions, is not going to be able to immediately become very reputable. Similar conclusions were also presented by Movshovitz-Attias et al. (2013), where the authors found that the

number of upvotes and accepted answers grows with the account maturity. Both, upvotes and accepted answers are responsible for the accumulation of the user reputation. Thus, these results complement what has been found in our study. It can be concluded that to build a high reputation it takes time, perseverance, and knowledge. Even with a vast knowledge base, it is basically impossible to become very reputable in a really short amount of time.

Apart from the user account maturity, the important factor governing the gained reputation is also the answer/question number ratio. The users in the community can vote to award reputation points for good answers as well as good questions. However, more reputation is gained from providing relevant answers than from asking useful questions, thus making question answering more beneficial to reputation increase. Based on the presented results the most reputable users all produce substantially more answers than questions. In other words, the most reputable users are those whose primary role is to help others with relevant answers. They do not ask for much help in return from the other members of the Q&A community. On the other hand, there are the users which primarily use the Q&A portal to ask questions and do not contribute a lot of answers. Based on the produced results, this group of users does not achieve such a high reputation which could be comparable to that of the users whose primary activity is question answering (Figure 5.1b and Tables 5.1, 5.4, 5.5, and 5.6). However, the analysis has also shown, that just a large frequency of posts is not enough for a high reputation accumulation. A large number of produced closed posts can also be a sign of poor quality and inexperience. Based on the produced results, posting sub-quality posts which get closed by the community has a negative effect on the reputation (Tables 5.4, 5.5, and 5.6). This, in turn, means that quantity is not more important than quality when trying to attain high reputation.

The above made conclusions are further confirmed in the research by Procaci et al. (2016) and Adamic et al. (2008). The former article also found the most reputable users to be mostly concentrated on providing answers and not on asking questions. Similar is also true for the later article. Here, the authors found that participating with answers in the topics with a large answer to question ratio is not the optimal strategy. In those topics, one has a strong competition from other users who also posted a large number of answers. Our conclusions regarding the negative effects of closed posts are also further confirmed in Correa and Sureka (2013). The authors have observed a decrease in the number of closed

posts produced by the more experienced and reputable users. All these discussed results confirm that to become a reputable member of the Q&A community one needs to work primarily on answering questions. Some reputation can be gained by asking a lot of relevant questions but to achieve a really high reputation, high quality question answering is the right activity. Next, the user social position and its relationship with the reputation is discussed.

User interactions play an important role in the online communities and perceived importance of its members. For the successful reputation accumulation, the most influential social position attribute has been found to be the user node degree centrality (Figure 5.2a and Tables 5.2, 5.4, 5.5, and 5.6). Consequently, one of the most important features governing the user reputation accumulation is the sheer volume of produced posts. In the social network terms, this means that the users that have a lot of interactions with other users will, in turn, have a high reputation. This builds on the previous discussion, where it has been concluded that producing predominantly answers to other users' questions, will have a very positive effect on one's reputation. The combined interpretation is that in fact producing a large number of answers will have the most positive effect on the reputation. Our findings can be supported by the Nikolaev et al. (2016). The authors have found a strong association between the degree centrality and the ability to engage with the community and be highly influential member of such community.

Apart from the node degree centrality, there are also other network measures indicating node's importance in a certain social network. One of them which was also used in this project is PageRank. There exists a strong association between PageRank score and the user reputation (Tables 5.2, 5.4, 5.5, and 5.6). This consequently means that answering questions posted by other active and possible reputable users will drive the reputation up. Conversely, answering questions from less active users, which post only rarely, is not too beneficial from the reputation standpoint. These less active users are potentially also less experienced and do not produce good quality questions. Such low quality questions probably do not garner a lot of attention from the wider community. This consequently means that corresponding answers are also not seen by a large number of interested users which would potentially vote for a reputation increase. Taking this into account, it can be concluded that it is most beneficial for the reputation increase to answer questions from other active users which themselves produce a large

number of posts (Nikolaev et al., 2016). However, this conclusion also needs a disclaimer, as it is well known that PageRank score is also affected by a sheer number of connections the user has (node degree), in this case, the number of posts. The results produced in our study are further confirmed in Movshovitz-Attias et al. (2013). The authors have also observed similar results in their research of the PageRank score and its association with the user reputation on StackOverflow Q&A portal. The same conclusion was made by Macleod (2014). It has been observed that in different StackExchange Q&A communities, users with high PageRank score tend to be more reputable.

The last examined social position metric governing user reputation was betweenness centrality. It has been shown that the betweenness centrality is slightly less well associated with the user reputation and social position (Tables 5.2, 5.4, 5.5, and 5.6). Furthermore, the plot in Figure 5.2b shows a certain degree of linear association between the betweenness centrality and user's reputation. However, as it has already been observed there is a sizeable number of outlier users. Similar decreased impact of the betweenness centrality on the user reputation was presented by Low and Svetinovic (2015). In their study, they observed not only a reduced importance of the betweenness centrality for the user reputation but actually a very small negative impact.

This diversion from the expected results can be attributed to the fact that betweenness centrality does not directly measure the connectedness of the nodes but rather their importance as information transmitters (Newman, 2010). The Q&A networks online are not completely sealed communities and users can bring in the external knowledge. Consequently, the knowledge propagation through the network is not the only possible way of learning new information. This is also the reason that some of the users could be gaining a lot of reputation even though they do not lie on a large number of shortest paths, thus being important for the information propagation through the network. In other words, some user in need of learning the new information is not relying only on the users with the high betweenness to enable the free flow of the information. Such user can basically circumvent other users with high betweenness and just read the desired information directly from the source user. In the Q&A network setting, the user does not need to directly interact (ask a question) with any of the users to learn new information. The user can just read an old post that has already been published and answered by some other members of the community. As

such the network structure stays the same (no new connections created), but the reputation of the user that provided an answer would still potentially increase.

However, the problematic outliers aside, the betweenness centrality results still give an important insight into the impacts of the reputation accumulation. It is mostly beneficial for the users to have a high betweenness as this means that they are forming bridges between different communities in the network. This can be interpreted with the reasoning that such users are experts in more than one distinct topic around which the network communities are formed (Ye et al., 2016). As such the knowledge of these users is more diverse and they are consequently able to provide their answers to multiple network communities. This, in turn, means that they can reach a larger number of people and thus have a higher potential to gain reputation based on their posts.

The last examined aspect of the user reputation accumulation was their posted text content. Highly reputable users of the CrossValidated Q&A community write in an analytical, objective and emotion neutral style (Figures 5.3 and 5.6) which is common for the science related topics. This is further confirmed by the fact that CrossValidated network is indeed mostly concerned with the topics related to statistics and machine learning, both strongly scientific fields. The described observation is further complemented by the fact that on the other hand, the results have shown that writing in an authentic and personal style is not beneficial for the overall user reputation (Figure 5.5). Namely, users who write in a personal subjective and emotional style will see a decrease in the ability to successfully gain reputation. Most reputable users also write longer and more evolved sentences than just basic, very short sentences. This result is potentially also aligned with the fact that analytical writing style is desired. To express a complex scientific idea, more words are normally needed and thus the longer sentences. Science related topics also use a lot of specialised technical terms which are often longer words. These conclusions are also supported in the research by Hwong et al. (2017). The authors have also observed similar writing style in the science-related social media posts. Namely, posts were found to be written in a highly analytical, objective and emotion neutral way.

Another prominent conclusion that can be made based on the results of the text content analysis is related to the clout score measuring expressed confidence. In the case of the more reputable users, they are more confident in their writing than average users (Figure 5.4). The confident writing can be explained with

the reasoning that these users are knowledgeable experts in the field they are contributing to. The experts in their respective fields commonly show a certain degree of authority and confidence in their scientific writing. This conclusion can also be supported by Hwong et al. (2017). They have observed that science-oriented social media content is also written in confident style. However, over confident writing can, on the other hand, have a negative effect on the user reputation. Too much confidence can be perceived by other people as arrogant or even unfriendly. It is obvious that in addition to multiple other factors, a correct amount of confidence needs to be displayed in answers for the maximum reputation gains.

To round the discussion of the results produced in this study we focus on the interpretation of the built latent variable models. User engagement factor has been found to have quite a small impact on the user reputation which is predominantly influenced by the user's social position (Tables 5.5 and 5.6). Furthermore, the latent factor for user engagement has a slightly negative coefficient indicating a small negative influence on the reputation. According to Brown et al. (1998); Smith and Crandall (1969), this result is due to the overwhelming influence of the competing latent factor for social position. Turning the attention to the influence of the observed variables, the most important user engagement attribute was the log answer/question ratio. The significantly less important attributes for user account maturity and even negatively associated number of closed posts have both shown only a minute impact on the user engagement latent factor. Accordingly, it can be said that being a frequent provider of good answers is by far the most beneficial activity for the reputation accumulation. In other words, highly reputable users are those which are the source of knowledge for other members of the community. This conclusion is aligned with those made in the previous paragraphs.

The social position attributes have been modelled as part of all three examined latent models (Tables 5.4, 5.5, and 5.6) which were in agreement that user's social position plays the most important role in determining the received reputation. The social position and its corresponding attributes are by far the most important for the user's reputation prediction. This is especially true in the case of the third model that was used (Table 5.6). It fit the social position factor with the overall largest observed factor loading (coefficient). The interpretation of this model is that users' interactions and thus their positioning in the social network are partly

influenced by their way of writing. Namely, better the user's style of writing, the better their position in the social network. This social position is in turn highly reflective on their reputation. When looking at the observed variables which form the social position factor, all the models are again in agreement with node degree being the most influential. For a good social positioning, it is the most important to be a very active user which produces a large number of posts. Further, slightly less important for one's social position is the act of answering questions from other active users. Lastly, it is also beneficial to be answering questions from several different communities, thus bridging the gap between them and in turn reaching the wider audience.

Finally, the attention is given to the text quality factors in latent models and the corresponding influence on the user reputation. Each of the models had a slightly different setting of text quality influence on the user reputation. For the two-factor model (Table 5.4), the same small negative factor loading effect can be observed as has been discussed in the case of user engagement in three-factor models. The explanation for this again lies in the overwhelming influence of the social position factor (Brown et al., 1998; Smith and Crandall, 1969). The post quality was further modelled in the three-factor model (Table 5.5) and a hierarchical three-factor model (Table 5.6). In the latter, the post quality was represented as a second order factor influencing user's social position instead of directly the user's reputation. Both three-factor models have shown some degree of post quality influence on reputation. Although it must be noted that compared to the social position impact, the text quality has quite a small one. This can be attributed to the fact that text is very unstructured data type and its analysis is also often affected by the ambiguity (Bhonde et al., 2010). Furthermore, the preference of what is a good text content is often a very personal and subjective decision. Thus the opinions about a single text body can often vary between different readers.

When looking at observed variables and their predictive importance for user reputation, it can be concluded that analytic writing style and being confident in one's writings are the most important. On the other hand, the reputation is most negatively influenced by writing in a subjective authentic style and the use of common dictionary words. Negative influence found for the proportion of the used LIWC dictionary words can be explained by the technical nature of the scientific writing. The LIWC dictionary includes most common use words, however, the

scientific writing often includes specialist words which are used only in certain technical fields. It, in turn, means the reduced chance of being captured in a general dictionary. This explanation is further supported by the apparent lower proportion of dictionary words in all the analysed texts. The average proportion for the analysed science oriented Q&A network is 76.72, while according to LIWC webpage[1], the average proportion of dictionary words in the general texts is 85.18. Consequently, it can be concluded that the most reputable users which, as it has been indicated before, are experts in their respective fields. Where necessary, they use highly specialised words which are not present in the dictionary. This, in turn, causes the observed negative influence of the attribute for the proportion of the dictionary words.

6.2 Research question 2: comparison with non-technical Q&A network

The second research question that has been answered was concerned with the comparison of main science oriented Q&A network to the Q&A community for everyday, non-technical topics. The comparison has been made to the Parenting Q&A network primarily used by parents to discuss the raising of kids.

Most of the presented results from Parenting network (Table 5.7) indicate the same conclusions as before in the case of the CrossValidated network. First order factor loadings are in agreement with those observed for the CrossValidated network (Table 5.6). The same is also true in the case of the user engagement observed variables and corresponding fitted coefficients. However, there is a slight reordering in the importance ranking of the observed network analysis attributes. Namely, the sheer number of posts is no longer the most important attribute for user's social position and thus reputation gathering. In the case of Parenting network, the PageRank score takes the first place (by a small margin). The observed change in the ranking of importance can be attributed, to the fact that PageRank score is known to be often influenced by the node degree. In turn, the observed high PageRank score can be also partly attributed to the high node degree. The second and a substantially more indicative of the change in nature of the compared Q&A network is the observed change in authentic writing

[1] http://liwc.wpengine.com/compare-dictionaries/

style. In the science oriented CrossValidated network, as it has already been discussed, the authentic and subjective writing has been discouraged. However, in the case of the non-technical Parenting Q&A network, the results indicated that authentic writing style has actually a positive effect on user's perceived post quality and conversely the reputation. This result indicates that not all text content should be strictly scientific and objective along the lines of academic research papers. When contributing to the to the Q&A network interesting in everyday topics such as parenting it is completely acceptable to show some degree of authenticity and subjective thinking. Another conclusion that can be made from the conducted analysis is the positive influence of the emotional writing tone. In the CrossValidated analysis, it has been concluded that emotional tone does not have any influence on the post quality. It was further observed that the best suited emotional tone for high reputation is a neutral tone. However, in the case of the Parenting network, this does not completely hold true as the emotional tone has a positive influence on the post quality. It can thus be concluded that in the non-science related Q&A communities it is required to show some degree of positive emotions in one's writings to successfully accumulate large reputation. When there is no strict need for academic writing style, people generally like to read content where authors express their emotions, especially positive. Apart from these more influential observations all the other text related observed and latent variables have shown a good degree of similarity to the previously discussed models fitted on CrossValidated data.

In conclusion, the second research question has been answered by comparing the two Q&A networks which are distinctly different in their main topics of interest. Science oriented network's users need to adhere to certain writing rules and styles. These, are inherently connected to the style found in the academic writing. On the other hand, the further examination of the Parenting network revealed that in the case of the network not oriented around science related topics, the previously inferred rules are more relaxed. Especially prominent was the shift in direction of the authentic writing style influence. Where it had a negative influence in the case of the CrossValidated network, in the Parenting network it has a positive impact. A similar argument has also been made in the case of showing emotions in one's writings. This leads to the conclusion that in the non-technical Q&A network expressing one's subjective side of thinking is not punished but has rather a positive influence.

Chapter 7

Conclusion and further work

7.1 Summary

This study has examined the Q&A communities which are growing in importance, both online and in the real life. One of the main social aspects of such communities is user reputation score. Its role is the recognition of the Q&A communities' expert members. As such, the high reputation score is perceived as very important and indicative of the person with a great wealth of knowledge in their respective field. Due to these important factors, user reputation accumulation process on Q&A websites offers various interesting research directions.

In the presented study, two specific research questions were being answered. First, we were interested in what factors are driving the user reputation score accumulation in the science oriented CrossValidated community. The second research question was exploring how the observed results and conclusions hold when compared to the non-science oriented Parenting network.

Our conducted analysis has first been performed on each of the extracted attributes separately. Then all different user describing attributes were jointly modelled with the unified latent variable models. The user describing attributes were of three major types. The user engagement attributes were extracted directly from the database. On the other hand, to produce social position attributes a social graph representing user interactions was built and analysed. Lastly, user reputation modelling was based on the user produced text content. Here, LIWC tool has been used to produce required summary text features.

In the presented study, we have first explored different user engagement factors and their impact on the user's reputation. It has been concluded that the

most reputable users are those which are the members of the particular Q&A community for a substantial amount of time. Their main activity has been found to be question answering, as it has been shown that the most reputable users hardly ever post a question. On the other hand, they provide a lot of high quality answers to the other users' questions.

Second important aspect of the presented study looked into building the social network and performing the corresponding analysis on the constructed social graph. Based on the produced results, we have concluded that the most important aspect indicating user's social position is the degree centrality. This conclusion, connected with that from the previous paragraph, confirms that the most reputable users are those producing a large number of posts which are, more specifically, answers. Furthermore, answering questions by other reputable users will in turn also help increase our own reputation. Lastly, it has also been found that reputation is affected by the ability of the users to traverse different communities with their connections. Users which have the necessary knowledge to be able to participate in multiple social graph communities dealing with different topics will gain the most reputation.

The last examined aspect contributing to the user's reputation accumulation was the text content they publish in their posts. The most prominent conclusion was the fact that confident and analytical writing style are the most important two aspects for high reputation accumulation. Combined, these two reputation indicators were explained by the conventions governing the scientific writing style which can be also found in most of the academic papers. Next, as part of the second research question, an interesting discrepancy has been revealed between the two types of the analysed Q&A communities. In the case of CrossValidated community, authentic and emotional writing styles are discouraged. However, in the case of the non-science related Parenting network, more authentic, emotional style of the posted content had a slightly positive association with the reputation score. This writing style preference had to do with the intended target group of both examined Q&A communities. All other examined reputation impacting aspects were found to be in an agreement between the two analysed Q&A communities.

The research presented in this thesis provides insights into the different factors governing the user reputation accumulation. The conclusions have far reaching implications as the online reputation scores also play a prominent role in the

real life. More concretely, the reputation score can serve as an important factor indicating person's level of competence, especially in science related fields. Accordingly, in today's highly competitive world the process of successful reputation accumulation is becoming more and more important.

7.2 Further work

The presented research had thoroughly examined different aspects of governing the gains in reputation score. However, there are still possible further research directions. The first option is to perform the described analysis on the substantially larger Q&A community such as StackOverflow. This research direction would be primarily enabled by the availability of the considerable compute resources which are required to operate on such a large social network. Building on this, a slightly less substantial, further research direction would be to also examine the use of different social network representations. The social graph could, for example, be built by only connecting the users if one of them accepted the answer from another. Furthermore, a certain threshold could be utilised to determine how many interactions in a form of question answering two certain users need to have in order to constitute a connection between their respective nodes.

The second major research direction would be the application of the more advanced NLP procedures to further analyse text content produced by the users. One possible algorithm that could be used in this research setting is *latent dirichlet allocation* (LDA) (Blei et al., 2003), which has seen use in the past research for text topic modelling and discovery. The other possible more involved NLP procedure which could be used is *Doc2Vec* (Le and Mikolov, 2014). This is an extension of the well-known algorithm *Word2Vec* (Mikolov et al., 2013), which is used to produce a dense distributed representation of the words. Conversely, in the case of *Doc2Vec*, a representation vector would be produced for each of the users based on their respective posted text content. An important feature of these distributed representations is the fact that vectors representing words or documents with similar meanings are close together in the vector space. Such representations could then clustered to gain even further insights into what type of content is posted by the reputable users.

Bibliography

A. Abbasi, R. T. Wigand, and L. Hossain. Measuring social capital through network analysis and its influence on individual performance. *Library & Information Science Research*, 36(1):66–73, 2014. doi: 10.1016/j.lisr.2013.08.001.

L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 665–674, New York, NY, USA, 2008. ACM. doi: 10.1145/1367497.1367587.

A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Steering user behavior with badges. *WWW '13 Proceedings of the 22nd international conference on World Wide Web*, pages 95–106, 2013. doi: 10.1145/2488388.2488398.

B. Bazelli, A. Hindle, and E. Stroulia. On the personality traits of StackOverflow users. *IEEE International Conference on Software Maintenance, ICSM*, pages 460–463, 2013. doi: 10.1109/ICSM.2013.72.

A. Bekkerman and G. Gilpin. High-speed Internet growth and the demand for locally accessible information content. *Journal of Urban Economics*, 77:1–10, 2013. doi: 10.1016/j.jue.2013.03.009.

T. N. Beran and C. Violato. Structural equation modeling in medical research: a primer. *BMC Res Notes*, 3:267, 2010. doi: 10.1186/1756-0500-3-267.

S. Beyer and M. Pinzger. Grouping Android Tag Synonyms on Stack Overflow. *Proceedings of the 13th International Conference on Mining Software Repositories*, pages 430–440, 2016. doi: 10.1145/2901739.2901750.

M. Bhanu and J. Chandra. Exploiting response patterns for identifying topical experts in stackoverflow. In *2016 Eleventh International Conference on Digital*

Information Management (ICDIM), pages 139–144, Sept 2016. doi: 10.1109/ICDIM.2016.7829790.

S. B. Bhonde, R. L. Paikrao, and K. U. Rahane. Text association analysis and ambiguity in text mining. *AIP Conference Proceedings*, 1324(1):204–206, 2010. doi: 10.1063/1.3526195.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

K. A. Bollen and M. D. Noble. Colloquium Paper: Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences*, 108(Supplement_3):15639–15646, 2011. doi: 10.1073/pnas.1010661108.

A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft. Building reputation in StackOverflow: An empirical investigation. *IEEE International Working Conference on Mining Software Repositories*, pages 89–92, 2013. doi: 10.1109/MSR.2013.6624013.

T. a. Brown, B. F. Chorpita, and D. H. Barlow. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of abnormal psychology*, 107(2):179–192, 1998. doi: 10.1037/0021-843X.107.2.179.

R. S. Burt. The network structure of social capital. *Research in organizational behavior*, 22:345–423, 2000.

R. S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110 (2):349–399, 2004. doi: 10.1086/421787.

R. S. Burt. Network-Related Personality and the Agency Question: Multirole Evidence from a Virtual World1. *American Journal of Sociology*, 118(3):543–591, 2012. doi: 10.1086/667856.

L. Cai. Latent variable modeling. *Shanghai archives of psychiatry*, 24(2):118–120, 2012. doi: 10.3969/j.issn.1002-0829.2012.02.010.

F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli. Mining successful answers in stack overflow. *IEEE International Working Conference on Mining Software Repositories*, 2015-Augus:430–433, 2015. doi: 10.1109/MSR.2015.56.

M. Choetkiertikul, D. Avery, H. K. Dam, T. Tran, and A. Ghose. Who Will Answer My Question on Stack Overflow? *2015 24th Australasian Software Engineering Conference*, pages 155–164, 2015. doi: 10.1109/ASWEC.2015.28.

E. Cohen, D. Delling, T. Pajor, and R. F. Werneck. Computing classic closeness centrality, at scale. In *Proceedings of the second ACM conference on Online social networks*, pages 37–50. ACM, 2014.

M. A. Cohn, M. R. Mehl, and J. W. Pennebaker. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science*, 15(10): 687–693, 2004. doi: 10.1111/j.0956-7976.2004.00741.x. PMID: 15447640.

D. Correa and A. Sureka. Fit or unfit: Analysis and prediction of 'closed questions' on Stack Overflow. *Proceedings of the first ACM Conference on Online Social Networks*, pages 201–212, 2013. doi: 10.1145/2512938.2512954.

D. Correa and A. Sureka. Chaff from the wheat. *Proceedings of the 23rd international conference on World wide web - WWW '14*, pages 631–642, 2014. doi: 10.1145/2566486.2568036.

N. M. M. Dowell, S. Skrypnyk, S. Joksimović, A. Graesser, S. Dawson, D. Gašević, T. a. Hennis, P. D. Vries, and V. Kovanović. Modeling Learners ' Social Centrality and Performance through Language and Discourse. *Educational Data Mining - EDM'15*, pages 250–257, 2015.

A. Field, J. Miles, and Z. Field. *Discovering Statistics Using R*. SAGE Publications, 2012.

L. C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239, 1978.

J. Goderie, B. M. Georgsson, B. Van Graafeiland, and A. Bacchelli. ETA: Estimated time of answer predicting response time in stack overflow. *IEEE International Working Conference on Mining Software Repositories*, 2015-Augus: 414–417, 2015. doi: 10.1109/MSR.2015.52.

Y.-l. Hwong, C. Oliver, M. V. Kranendonk, and C. Sammut. Computers in Human Behavior What makes you tick ? The psychology of social media engagement in space science communication. *Computers in Human Behavior*, 68:480–492, 2017. doi: 10.1016/j.chb.2016.11.068.

N. Immorlica, G. Stoddard, and V. Syrgkanis. Social Status and Badge Design. *Www*, pages 473–483, 2015. doi: 10.1145/2736277.2741664.

Y. Jin, X. Yang, R. G. Kula, E. Choi, K. Inoue, and H. Iida. Quick trigger on stack overflow: A study of gamification-influenced member tendencies. *IEEE International Working Conference on Mining Software Repositories*, 2015-Augus: 434–437, 2015. doi: 10.1109/MSR.2015.57.

Y. Jing and S. Baluja. Pagerank for product image search. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 307–316, New York, NY, USA, 2008. ACM. doi: 10.1145/1367497.1367540.

S. Joksimović, V. Kovanović, N. Dowell, D. Gašević, A. C. Graesser, O. Skrypnyk, and S. Dawson. How do you connect? Analysis of social capital accumulation in connectivist MOOCs. *ACM International Conference Proceeding Series*, 16-20-Marc:64–68, 2015. doi: 10.1145/2723576.2723604.

D. Jurafsky and J. H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon, and A. C. Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014. doi: 10.1177/0261927X13502654.

W.-c. Kao and S.-w. Wang. Expert Finding in Question-Answering Websites : A Novel Hybrid Approach. *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 867–871, 2010. doi: 10.1145/1774088.1774266.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014. URL `http://arxiv.org/abs/1405.4053`.

C. Levesque, A. N. Zuehlke, L. R. Stanek, and R. M. Ryan. Autonomy and Competence in German and American University Students : A Comparative Study Based on Self-Determination Theory. 96(1):68–84, 2004. doi: 10.1037/0022-0663.96.1.68.

J. F. Low and D. Svetinovic. Data analysis of social community reputation: Good questions vs. good answers. In *2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1193–1197, Dec 2015. doi: 10.1109/IEEM.2015.7385836.

L. Macleod. Reputation on stack exchange: Tag, You're It! In *Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014*, pages 670–674, 2014. doi: 10.1109/WAINA.2014.108.

S. K. Maity, A. Kharb, and A. Mukherjee. Language Use Matters: Analysis of the Linguistic Structure of Question Texts Can Characterize Answerability in Quora. (June 2014), 2017. URL `http://arxiv.org/abs/1703.04001`.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL `http://arxiv.org/abs/1301.3781`.

D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos. Analysis of the reputation system and user contributions on a question answering website: StackOverflow. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13*, pages 886–893, 2013. doi: 10.1145/2492517.2500242.

M. Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

M. E. E. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):15, 2003. doi: 10.1016/j.socnet.2004.11.009.

M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 2003. doi: 10.1177/0146167203029005010.

U. Nieminen. On the centrality in a directed graph. *Social Science Research*, 2 (4):371 – 378, 1973.

A. Nikolaev, S. Gore, and V. Govindaraju. Engagement capacity and engaging team formation for reach maximization of online social media platforms. *Pro-*

ceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu:225–234, 2016. doi: 10.1145/2939672.2939681.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*, 54(1999-66):1–17, 1998. doi: 10.1.1.31.1768.

J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, and D. I. Beaver. When small words foretell academic success: The case of college admissions essays. *PLOS ONE*, 9(12):1–10, 12 2015. doi: 10.1371/journal.pone.0115844.

T. B. Procaci, B. P. Nunes, T. Nurmikko-fuller, and S. W. M. Siqueira. Finding Topical Experts in Question & Answer Communities. 2016. doi: 10.1109/ICALT.2016.68.

Y. Rosseel. lavaan: An R package for structural equation mdeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

B. Ruhnau. Eigenvector-centrality—a node-centrality? *Social networks*, 22(4):357–365, 2000.

M. Salter-Townshend, A. White, I. Gollini, and T. B. Murphy. Review of statistical network analysis: Models, algorithms, and software. *Statistical Analysis and Data Mining*, 5(4):243–264, 2012. doi: 10.1002/sam.11146.

J. B. Schreiber, A. Nora, F. K. Stage, E. A. Barlow, J. King, A. Nora, and E. A. Barlow. Reportig Structural Equation Modeling and Confirmatory Factor Analysis Results : A Review. *The Journal of Educational Research*, 99(6):232–338, 2006. doi: 10.3200/JOER.99.6.323-338.

J. Scott. *Social Network Analysis*. SAGE Publications, 2017. ISBN 9781526412232.

A. Skrondal and S. Rabe-Hesketh. Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, 34(4):712–745, 2007. doi: 10.1111/j.1467-9469.2007.00573.x.

R. Slag, M. De Waard, and A. Bacchelli. One-day flies on StackOverflow - Why the vast majority of StackOverflow users only posts once. *IEEE International*

Working Conference on Mining Software Repositories, 2015-Augus:458–461, 2015. doi: 10.1109/MSR.2015.63.

C. Smith and V. Crandall. *Achievement-Related Motives in Children*. Russell Sage Foundation, 1969. ISBN 9781610446938.

D.-S. Tzeng, Y.-C. Wu, and J.-Y. Hsu. Latent variable modeling and its implications for institutional review board review: variables that delay the reviewing process. *BMC Medical Ethics*, 16(1):57, 2015. doi: 10.1186/s12910-015-0050-8.

G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao. Wisdom in the social crowd: An analysis of quora. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1341–1352, New York, NY, USA, 2013a. ACM. doi: 10.1145/2488388.2488506.

G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang. ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, 54(3):1442–1451, 2013b. doi: 10.1016/j.dss.2012.12.020.

P. Wilson and A. Risk. How to find the good and avoid the bad or ugly: a short guide to tools for rating quality of health information on the internet-commentary: On the way to quality. *BMJ*, 324(7337):598–602, 2002. doi: 10.1136/bmj.324.7337.598.

B. Xu, Z. Xing, X. Xia, D. Lo, Q. Wang, and S. Li. Domain-specific cross-language relevant question retrieval. *Proceedings of the 13th International Workshop on Mining Software Repositories - MSR '16*, pages 413–424, 2016. doi: 10.1145/2901739.2901746.

D. Yang, A. Hussain, and C. V. Lopes. From Query to Usable Code: An Analysis of Stack Overflow Code Snippets. *Proceedings of the 13th International Workshop on Mining Software Repositories - MSR '16*, pages 391–402, 2016a. doi: 10.1145/2901739.2901767.

J. Yang, S. Peng, L. Wang, and B. Wu. Finding Experts in Community Question Answering Based on Topic-Sensitive Link Analysis. *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pages 54–60, 2016b. doi: 10.1109/DSC.2016.35.

D. Ye, Z. Xing, and N. Kapre. The structure and dynamics of knowledge network in domain-specific Q&A sites: a case study of stack overflow. *Empirical Software Engineering*, pages 1–32, 2016. doi: 10.1007/s10664-016-9430-z.

A. G. Yong and S. Pearce. A beginner's guide to factor analysis: Focusing on exploratory factor analysis, 2013.

X. N. Zuo, R. Ehmke, M. Mennes, D. Imperati, F. X. Castellanos, O. Sporns, and M. P. Milham. Network centrality in the human functional connectome. *Cerebral Cortex*, 22(8):1862–1875, 2012. doi: 10.1093/cercor/bhr269.