Summer 6-15-2016

TALKING TO ME? CREATING NETWORKS FROM ONLINE COMMUNITY LOGS

Remko W. Helms
Utrecht University, remko.helms@ou.nl

W. Ai
University of Utrecht, w.ai@uu.nl

Jocelyn Cranefield
Victoria University of Wellington, jocelyn.cranefield@vuw.ac.nz

TALKING TO ME? CREATING NETWORKS FROM ONLINE COMMUNITY LOGS

Research Paper

Helms, R.W., University of Utrecht, Department of Information and Computing Science, Utrecht, the Netherlands, r.w.helms@uu.nl; Open University, Heerlen, the Netherlands, Faculty of Management, Science and Technology, remko.helms@ou.nl

Ai, W., University of Utrecht, Department of Information and Computing Science, Utrecht, the Netherlands, w.ai@uu.nl

Jocelyn Cranefield, Victoria University of Wellington, School of Information Management, Wellington, New Zealand, jocelyn.cranefield@vuw.ac.nz

Abstract

Online communities offer many potential sources of value to individuals and organisations. However, the effectiveness of online communities in delivering benefits such as knowledge sharing depends on the network of social relations within a community. Research in this area aims to understand and optimize such networks. Researchers in this area employ diverse network creation methods, with little focus on the selection process, the fit of the selected method, or its relative accuracy. In this study we evaluate and compare the performance of four network creation methods. First we review the literature to identify four network creation methods (algorithms) and their underlying assumptions. Using several data sets from an online community we test and compare the accuracy of each method against a baseline ('actual') network determined by content analysis. We use visual inspection, network correlation analysis and sensitivity analysis to highlight similarities and differences between the methods, and find some differences significant enough to impact study results. Based on our observations we argue for more careful selection of network creation methods. We propose two key guidelines for research into social networks that uses unstructured data from online communities. The study contributes to the rigour of methodological decisions underpinning research in this area.

Keywords: online community, social network analysis, network creation, enterprise social network.

1 Introduction

In the current digitized and networked world, online communities are all around us. Online communities are groups of people that communicate online in a shared virtual space, bound by a common purpose or interest, and guided by shared norms (De Souza and Preece, 2004; Phang, Kankanhalli and Sabherwal, 2009; Preece and Schneiderman, 2009). The fundamental activity is social interaction through which the members of online communities share information (Ridings and Gefen, 2004). The nature of interaction is related to the common purpose or interest of members, which may range from a shared passion for a place or object to a shared hobby (e.g. Füller, Jawecki and Mühlbacher, 2007), medical interest, or profession. Online communities thereby provide diverse and complementary benefits to their members: access to information, social interaction (Iriberri and Leroy, 2009), a sense of membership and belonging (Blanchard and Markus, 2004) and friendship and social support (Ridings and Gefen, 2004). Participation leads to the creation of social capital based on a combination of ties, trust, norms of reciprocity, identification, and a shared vision and language, which in turn influences knowledge sharing in the community (Chiu, Hsu and Wang, 2006). The organisational benefits of online communities may be substantial, including boosting profit and sales, generating innovation and improving customer knowledge and relations (Agarawal, Gupta & Kraut, 2008).

Contemporary online communities are not bound to a specific technology and can form on any platform (or across several platforms, cf. Cranefield, Yoong and Huff (2015)) that supports online interaction, ranging from e-mail lists to online discussion boards (e.g. GitHub, StackOverflow) and online social networks (e.g. Facebook, LinkedIn, Yammer, You Tube). Given the value noted above, an important research opportunity lies in the fact that the interactions among online community members leave an enormous amount of complex, digitized and self-documenting records behind (Gleave, Welser, Lento and Smith, 2009; Giles, 2012). This is because every contribution by a community member is logged including its meta-data such as date, time, subject, and name of contributor. This wealth of online data has attracted the attention of a wide range of researchers who have mined online community data to study the interaction patterns and behaviour of users on these platforms (Howison, Inoue and Crowston, 2006; Agarwal, Gupta and Kraut, 2008). Structural mapping of the reply-to network of online communities is of particular interest to research because of the deep insights it allows into how particular communities operate and who are their key influencers. Based on network analysis, it may, for example, be possible to identify issues and design interventions (Cross, Parker, Prusak and Borgatti, 2001; Helms and Buysrogge, 2005; Helms, 2007).

A commonly used approach in this type of research is a social network analysis approach which models the messaging activity between members of an online community as a reply-to network (e.g. Berger, Klier, Klier and Richter, 2014). The data required for this type of research are derived from the logs of online communities. Deriving a reply-to network from this data may be relatively simple or may be a more complex procedure, depending on whether one is dealing with structured or unstructured network data. In the case of structured network data it is generally known who communicates with who; for example, a message on Twitter (i.e. @mention) contains meta-data about the sender and the receiver. However, in many cases it may be necessary to take into account unstructured network data. This occurs, for example, in discussion forums such as StackOverflow or Yammer, where people post to a thread rather than responding to a particular message or person. This makes it more complex to derive reply-to relations from unstructured data (Petrovčič, Vehovar and Žiberna, 2012). A review of research that uses unstructured data to determine the reply-to network (see section 2.2) reveals that researchers use different methods (i.e. different algorithms) to automatically generate network data from unstructured data sources based on the sequence of posts in a thread and assumptions about interaction patterns (Toral, Martínez-Torres and Barrero, 2010; Faraj and Johnson, 2011; Berger et al., 2014). Our review of this literature further revealed that the choice of such algorithms is not always well justified as a methodology, and the effect of using alternative algorithms on the results of the research has not been evaluated. It appears there is a risk that researchers may assume that the generated

networks closely resemble the reply-to relations as originally intended by the poster of the message, which we will refer to as the 'actual' network. If it is found that these assumptions are false, this may have major and serious consequences for the outcomes of this stream of research. Our paper is motivated by this concern: its aim is to explore the problem of creating networks from unstructured network data and to propose a method to increase the rigor of research in this area.

Our research addresses this problem using two years of unstructured data from an online community. We firstly derive and identify the 'actual' reply-to network from this data using content analysis, then compare this network with pseudo reply-to networks generated from the same data by using three different network creation algorithms from the literature. Our comparison is performed by visually comparing the networks (supported by some network metrics) and using Pearson correlation analysis. The goal of the comparison is to study to what extent the generated networks deviate from the 'actual' network. Furthermore, we apply sensitivity analysis to test whether the length of threads or the discussion topic influences the nature of the interaction pattern. Based on the findings we formulate methodological recommendations concerning the processing and use of unstructured data from online communities. We argue that such recommendations are urgently needed as there is currently no consensus on what methods to apply when dealing with unstructured network data from online communities, impacting on the consistency and quality of research, and reducing the ability to undertake higher level aggregative and comparative studies.

The remainder of this paper is structured as follows. The next section provides an overview of research using online community data, emphasizing the network creation algorithms that have been applied. Then section 3 elaborates on the research design and provides details on the sampling and comparison of the networks. Results of the comparison and sensitivity analysis are presented in section 4. Finally, in section 5 these results are discussed and recommendations for using unstructured network data in online community research are formulated.

2 Related work

2.1 Network analysis for studying online communities

An important area of study in the field of online community research is the social structure of communities. This social structure consists of the relationships and interactions between the people in the community, which can be studied using social network analysis (e.g. Falkowski, Bartelheimer and Spiliopoulou, 2007; Trier and Richter, 2014). Social network analysis (SNA) is a research method that stems from the Sociology domain but is also widely applied outside this domain including the domain of information science (Otte and Rousseau, 2002; Hanneman and Riddle, 2005). The concept of SNA is to model a social structure as a network where the nodes represent persons and the edges represent relationships between those persons (Wasserman and Faust, 1994). Visualization of a network provides insights in the structure of a network and reveals how well it is connected and who are central persons in the network for example. Networks can also be analysed more quantitatively using graph theory.

When SNA is applied in online community research it often focuses on the user level, aiming to identify user roles by analysing conversation patterns. For example, the *online leader* role is used to signify the central people in the network. Research by Huffaker (2010) has shown that online leaders acquire their influential role through high communication activity and linguistic variety in messages. Other research has identified the role of *value added users* – users who help others in the community by responding to questions and sharing knowledge (Berger et al., 2014). Typically, these value added users are well-connected and have a central position in the network. Research by Himelboim, Gleave and Smith (2009) identified three social roles by analysing the communication behaviour of Usenet users. Combining network analysis with interpretive analysis of message content they identified the following roles: *answer person*, *discussion person* and *discussion catalyst*.

Research is also conducted on the thread level and the network level. On network level, Zhongbao and Changshui (2003) have shown that online conversation patterns are very much affected by the personal interests of members. Furthermore, by visualizing the network of certain forums and analysing their structure, researchers identified different types of structure for social support: Q&A-, conversational- and flame newsgroups (Turner, Smith, Fisher and Welser, 2005; Fisher, Smith and Welser, 2006). For instance, a forum with 70% questioners, 20% answers and 10% discussions would be the marker of a Q&A group (Fisher et al., 2006). An example of research on the thread level is the research by Gómez et al. (2008). They found that people are more likely to relate to establish relationships with those people that have outspoken or controversial opinions. To identify those people they developed a "simple measure to evaluate the degree of controversy provoked by a post" (Gómez et al., 2008, p. 1).

2.2 Methods for creating networks from online community logs

To perform the network analysis studies discussed in the previous section, it is necessary to derive reply-to data from the online communities being studied. In some studies this is fairly straightforward: Data logs may indicate, for example, who interacts with who (e.g. Ediger et al., 2010). In these cases, the identity (ID) of both the sender (i.e. poster) and the receiver of messages is known. Hence, the reply-to relations can be derived directly from the log of the online community. In other cases, such as on StackOverflow or Yammer, the communication is recorded at a thread level. A thread is typically a discussion topic that is started by the thread starter. Other users can contribute to the thread by posting messages. In this situation, each message is linked to a thread and has a sender ID. A receiver ID is typically lacking in these situations, making it is impossible to directly derive the reply-to network from the online community logs. This poses a significant challenge to scholars who are interested in utilizing the enormous amount of quantitative data generated by such online communities (Petrovčič et al., 2012).

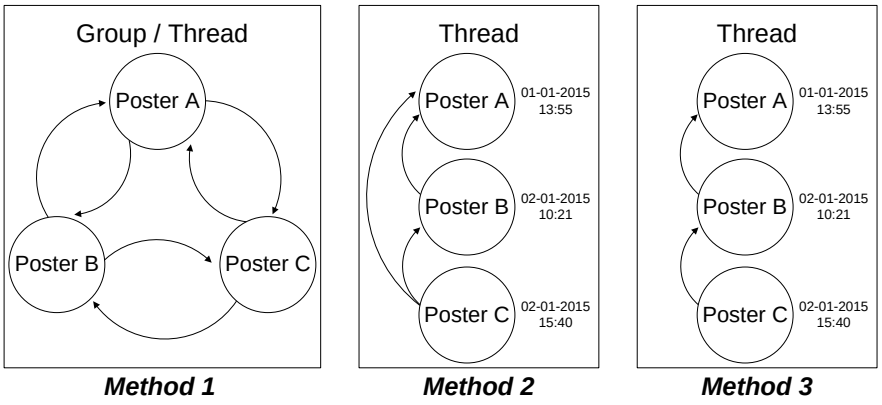

Figure 1. Visualization of three pseudo network creation methods

As it is impossible to directly derive network data from online community logs storing communication at a thread level, researchers working with such data need to make inferences and assumptions about who is talking to whom. A review of the literature reveals that different assumptions have been employed by different researchers working in this area. The first example is Berger et al. (2014) who studied the role and impact of key users in knowledge-intensive online communities regarding internal information and knowledge sharing. These researchers collected network data from a multinational corporation by mining activity data (messages in threads) from the corporation's Yammer platform. Network relations between posters were inferred when these posters posted messages in the same discussion thread, resulting in bi-directional relations between all posters (see method 1 in figure 1). In analysing this approach, we can derive that it is based on the assumption that everyone involved in a discussion (i.e. thread) is contributing equally.

Another approach was developed by Toral, Martínez-Torres and Barrero (2010) for collecting network data from an online community for open source projects. Their research objective was to analyse the knowledge sharing behaviour of active users in this online community with the goal of helping to improve the underlying projects. Once again, the network is based on activity data that takes place in threads on the community platform (see method 2 in figure 1). In contrast to Berger et al. (2014), these authors also use the message timestamp in network creation. Using the timestamp all messages in a thread can be arranged in chronological order, i.e. from oldest to newest. Network relations are then established by creating a directed network relation between the poster of a message and the posters of all previously posted messages in a thread, i.e. a "reply-to preceding posts" approach (method 2 in figure 1). The rationale for this approach is outlined by Toral et al. (2010) who argue that using discussion threads as the basic unit of analysis is "*highly valid, considering that the epistemic interactions in support of OSS [Open Source Software] development often take place in discussion threads where individual postings provide the context to encourage participation (Kuk, 2004).*" They note that it is more cognitively demanding to reply to a threaded discussion than a single message, as the flow of earlier postings needs to be taken into account (Toral et al., 2010, p. 298).

In similar settings to the previous example, Faraj and Johnson (2011) conducted a research project to better understand the mechanisms of how online communities sustain themselves. From studying the communication networks of five large-scale online communities they found that network exchange patterns are characterized by (a) direct reciprocity (i.e. direct interaction between a pair of actors), (b) indirect reciprocity (i.e. indirect interaction between a pair of actors via a third actor) and (c) preferential attachment (i.e. a concentration of communication). Their research illustrates yet another alternative network creation method. According to this network identification method, posters within a thread are again organized in a chronological order based on the timestamp. But in this case a network relation is inferred on the basis of "*inbound links where the presence of a reply represents a directed relationship from its author to the author of the immediately prior message in the message thread*" (Faraj and Johnson, 2011, p. 1471). Following this approach there is only a directional link to the poster of the previous message in a thread (see method 3 in figure 1). The rationale for this approach remains unclear because the authors mention that it follows from their theoretical frame (i.e. including direct reciprocity, indirect reciprocity and preferential attachment) without making it more explicit.

Method	Sender	Grouping attribute	Resulting relations(s)
1	A	group_id or thread_id	A → all posters of messages within the boundary set by grouping attribute
2	A	thread_id and timestamp	A → all posters of messages in thread_id and prior to the timestamp of the poster's own message
3	A	thread_id and timestamp	A → closest message in thread_id and prior to the timestamp of the poster's own message
4	A	thread_id and timestamp	A → all k posters of messages in thread_id and prior to the timestamp (1<k<infinity). Relations receive a weight based on distance and time between messages.

Table 1. Details of four network creation methods

Another approach is proposed by Petrovčič, Vehovar, and Žiberna, (2012) based on a review of the literature on network creation. These authors identify approaches similar to those identified previously in this section but take into account the time passed and messages in between the messages of two posters. In addition they report on network creation approaches that are based on quotations (for example, when an explicit link is made to a previous message by mentioning the name of a user or quoting some text). They note, however, that that such quotations are often missing in messages, making this unsuitable as an approach. Following Sack (2000), they then propose a more sophisticated ap-

proach that is "based upon more and more sensitive readings of post-to-post relationships". The assumption is that contributors in web forums typically respond to k of the recent messages, where k is 1 or higher (for k=1 it resembles method 2 and for k=infinity it resembles method 3). Furthermore, the relations have different weights depending on the number of messages (i.e. distance) and the time elapsed between a pair of messages in a thread. This algorithm for creating pseudo reply-to networks is then evaluated against quotation networks, since the true reply-to network is not known, according the researchers. Based on an evaluation, for different parameter settings for k, time and distance, they finally conclude that the algorithm for creating pseudo reply-to networks results in a good approximation of the quotation network that they used as a benchmark in their study. The best approximation is found when k is set to infinity, an outcome which is similar to method 3 (if the weights of the relations are not taken into account).

For reference purposes the network creation methods identified in the literature are briefly summarized in table 1.

3 Research design

This section describes the research design that was applied to evaluate the pseudo network creation algorithms that were identified in the literature. We collected activity data from an online community then derived a series of pseudo networks using three of the four network creation algorithms presented in the previous section. From the same data we also mapped the 'actual' reply-to network using a content analysis approach. This enabled a comparison of the pseudo networks versus the 'actual' network. The details of how we conducted this evaluation are outlined in the following sub-sections.

3.1 Data set and samples

For our research we used data from the *Hallo! Community*, an online forum of the Dutch Chamber of Commerce. It is a free and public online forum for all entrepreneurs (mostly Dutch oriented) for knowledge sharing, exchanging experiences, asking questions and providing answers to each other regarding any business related activities. It was launched in 2009 and can be characterized as a Topic-oriented Discussion Community according the taxonomy of Stanoevska-slabeva (2002).

Discussions on the Hallo! Community platform are hierarchically structured. The first layer of the Hallo! Community is the main category, which consists of 6 categories. Underneath this main category, we find the sub-categories each having 6 sub-categories (except one). Examples of this sub-categories are: Networking, Promotion, Communication & Marketing, Finance and so forth. Each of the sub-categories consists of a discussion forum where an user of the Hallo! Community can start discussions related to the topic of the forum. A discussion is started by starting a new thread and the post of the thread starter is the first message in the thread. The posts of other users who participate in the discussion are added to the same thread. Hence, a sub-category is a collection of threads that we will refer to as a community.

Number of registered users	35,972
Number of active users	7,662 (21.3% of total)
Number of main categories	6
Number of sub categories	35
Total number of threads	12,776
Total number of posts	45,682
Average reply rate per thread	3.58 replies
Average activity per active user	7.63 threads/replies
Longest thread (max # of replies)	571 replies

Table 2. Descriptive statistics of Hallo! community

For our research we had access to the first two years of data of the Hallo! Community (March 2009-March 2011). From this data we removed threads that did not get any replies because there can only be a discussion if there is reply to a thread by somebody other than the thread starter. Therefore, we also removed threads that only received replies from the thread starter. Finally, we removed threads that started before the official launch date of the community since this involved test data. After cleaning the data, there were almost 36,000 registered users on Hallo! community with 7,662 active users having contributed to the community by starting at least one thread or posting one reply to a thread in this community; 12,776 threads were started (across all 35 sub-categories), and those threads received 45,682 replies (see also table 2 for more statistics).

Since there was limited time available and owing to the fact that we applied content analysis for deriving the 'actual' reply-to network from the data, there was a need to extract a smaller data sample from our total data set. Rather than extracting one data sample, we extracted three data samples to militate against possible effects that might influence our results. First of all, different sub-categories might have different purposes and therefore different discussion patterns (Adamic, Zhang, Bakshy and Ackerman, 2008). Secondly, also the thread length is associated with different discussion patterns (Adamic et al., 2008). Hence, three different data samples with different thread selections have been created. Sample 1 contained 500 threads varying from 1 to 10 replies (83% of the threads ended within 10 replies). For each thread length (i.e. 1 to 10) fifty threads were included in each sample. Sample 2 contained 150 threads from six different discussion categories (randomly selected) with a minimum average of one post each day. Each sub-category in this sample contained 25 threads and the number of replies for each thread varied from 1 to 10. Sample 3 contained only threads with more than 10 replies. In total 43 threads were included. The threads in each of the three samples are unique, meaning that there is no overlap between the samples. An overview of the composition of the samples is shown in table 3.

Sample / sub-category	**Number of replies**	**Number of threads**	**Number of replies**
Set 1	1	50	50
	2	50	100
	3	50	150
	4	50	200
	5	50	250
	6	50	300
	7	50	350
	8	50	400
	9	50	450
	10	50	500 (sum: 2750)
Set 2			
- 101	1-10	25	127
- 301	1-10	25	111
- 303	1-10	25	107
- 306	1-10	25	120
- 601	1-10	25	94
- 605	1-10	25	89 (sum: 648)
Set 3	More than 10	43	932
Grand total		693	4330
% of total data set		13.1%	12.2%

Table 3. Samples for content analysis

3.2 Baseline ('Actual') network creation

To evaluate the accuracy of pseudo network creation algorithms it was firstly necessary to establish as a baseline the 'actual' network: We did this by manually analysing and interpreting the messages that members posted to threads and then mapping this to relationships between those members. Since the online community only registers posts to threads and not direct reply-to relations, these 'actual' reply-to relations were determined manually using content analysis (Hara, Bonk and Angeli, 2000). In qualitative research content analysis is often used to derive meaning from qualitative data, e.g. texts (Yang and Fang, 2004). We used content analysis to derive the receiver of each message (within a thread). Our analysis of the text focused on identifying any information that linked the message content to one of the previous messages in the thread; for example, mentioning the name of the user who had posted a previous message, a quote from the text of a previous message or any other contextual information that could be linked to a previous message. In cases where it was not possible to identify the receivers, the message was interpreted as being a response to all previous messages in the thread (less than 5% of the messages). Content analysis was performed by one of the authors who coded each message twice.

3.3 Pseudo network creation

The pseudo networks were then derived from the same data set using the algorithms presented in section 2. One of the authors wrote a small Python program for each algorithm. The data samples from section 3.2 were put into these programmes and the output was an edge list. An edge list is a standard format for storing social network data. Each row in the edge list contains the data of one relation, indicating the user IDs of the users that have a relation and the direction of the relation. During the analysis the edge lists were imported into UCINET and Gephi, two commonly used social network analysis tools, for further analysis.

One of the pseudo network creation algorithms from section 2 was not implemented in Python code: method 4. The reason for not including this algorithm in the evaluation is that the results from Petrovčič et al. (2012) showed that it is bears much resemblance with method 3 (i.e. best results were obtained were k=infinite). Instead, we added a new method that we identified during content analysis of the data when deriving the 'actual' network.

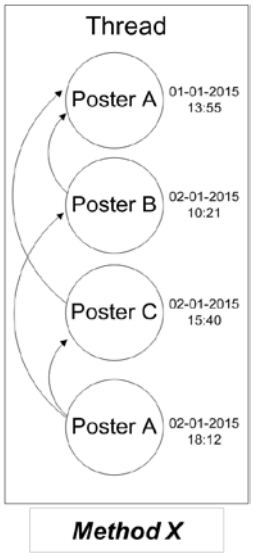

Figure 2. Pseudo network creation method X

Content analysis revealed that the vast majority of replies within the same thread, are addressed toward thread starters. And when thread starters reappear in their own thread, it is most likely that the thread starter posts a reply addressing every other participant who previously has posted within the discussion thread, e.g. giving a summary of the replies or thanking everyone for their responses. Based on these two factors, we propose an alternative pseudo network creation method with the following rules:

1) The initial post (the thread itself) does not build any relation towards others due to its broadcasting nature;
2) A reply posted by a user other than the thread starter himself will result in an relation from this user to the thread starter;
3) If a thread starter replies in its own thread, and this is not the first reply of the thread, then this reply will result in relations being built from this [thread starter] replies' re-entering position to all users who had replied previously in the thread.

A visualized version of this proposed method is displayed in Figure 2, named Method X. The outcome of Method X is that the thread starter has a link with all posters in the thread, while the other posters in the thread do not have relations with each other.

3.4 Evaluation of pseudo networks

Evaluation of the generated pseudo networks was performed in two different ways. First a visual comparison of the networks was performed. These visualizations were generated using Gephi and the applied visualization algorithm was Forced Atlas 2 (Jacomy, Venturini, Heymann and Bastian, 2014). Although a visual inspection is not a very precise comparison, it provides a useful initial impression of how closely the pseudo networks resemble the 'actual' network. Furthermore, the visualizations were supported with network metrics that gave a basic impression of the structure of the networks. This supported the visual inspection by showing whether some structural characteristics were similar. The metrics calculated, using the Gephi software, were: number of nodes, number of relations, density of the network, diameter of the network, and average clustering coefficient.

A more precise comparison of the networks was then performed using Pearson correlation analysis. This is a standard measure for calculating the similarity between networks with the same nodes (Hanneman and Riddle, 2005). For calculating the Pearson correlation the quadratic assignment procedure (QAP) was used to test for the significance of the relationship. This procedure was required because in network analysis one cannot assume independence of the observations (Barnett, 2011). For calculating the Pearson correlations we used the UCINET software package (Freeman, Everett and Borgatti, 2015).

3.5 Sensitivity analysis

As described in section 3.2 we created three different data samples (see also Table 3). The first evaluation of the pseudo network creation algorithms was performed on sample 1. This sample contained a good mix of threads of different lengths (except for very long threads >10 messages) across different sub-categories. The rationale behind doing this was to find results that were applicable to all threads from the Hallo! Community. The other two samples were used for sensitivity analysis, which concerns checking whether the results also hold in different circumstances (Pannell, 1997; Faraj and Johnson, 2011). Sample 2 contains an equal distribution of threads from 6 different discussion categories. This sample was used to analyse whether the outcome of the evaluation varied depending on the discussion category. For example, the type of discussion might vary across the discussion categories and this could therefore result in different reply-to structures (Turner et al., 2005; Fisher et al., 2006). Consequently, there might not be one pseudo network creation algorithm that works equally well for all subcategories. Finally, sample 3 was used to test whether long discussion threads (>10 messages) had a different reply-to structure and therefore required a different pseudo network creation algorithm than shorter discussion threads.

4 Results

4.1 Visual comparison of networks

The first analysis involves a visual comparison of the 'actual' network versus pseudo networks 1, 2, 3 and X, which are based respectively on methods 1, 2, 3 and X. To enable a comparison of the networks we generated the network visualisation for the 'actual' network using Gephi (Forced Atlas2 algorithm) and then fixed the position of the nodes. The same position of the nodes was then used for generating the visualisation for the pseudo networks (see Figure 3). Fixing the nodes ensures that the networks have the same orientation which makes comparison easier.

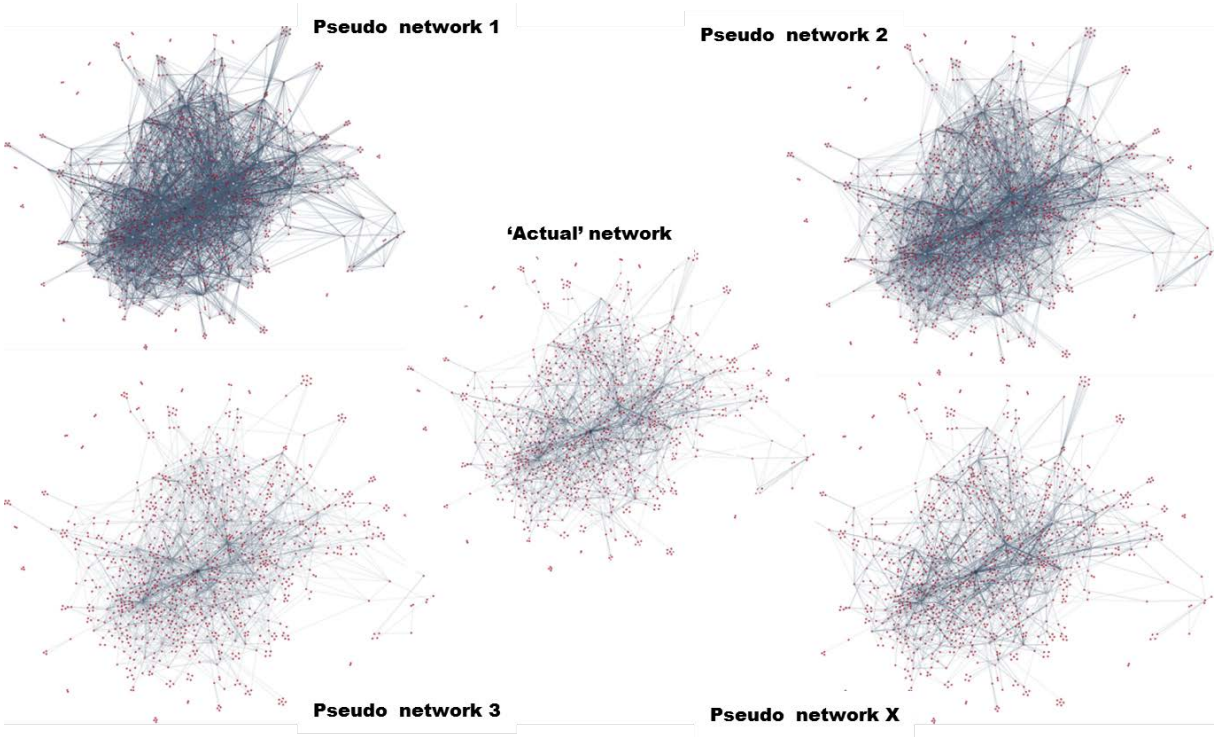

Figure 3. Visualization of various networks from content analysis sample set 1

Visual inspection of the networks in Figure 3 shows the networks have similar network patterns (same kind of concentration of relations at the same place in the network) . Furthermore, all networks have a central node in the middle of the network that has many relations (darker than other nodes in the network, although difficult to see for pseudo network 1). Two further observations that can be made are that (a) pseudo networks 1 and 2 are much denser (i.e. they show more relations) than the 'actual' network and (b) pseudo networks 3 and X look very similar to the 'actual' network. Since a visual inspection alone is not sufficient, we supported it by calculating common network metrics describing the basic properties of each of the networks (see Table 4).

The number of nodes (1,046) is the same for all networks because the same users are included in all networks. However, the number of edges (i.e. relations) determined by the network creation algorithms varies significantly. Method 1 generates four times the number of relations that appear in the 'actual' network while Method X is most accurate in showing the number of relations (deviation of 0.4%). However, this apparent match does not guarantee that the specific relations shown in the 'actual' network and pseudo network X are identical. The graphical density is based on the number of edges and therefore shows a similar pattern as the number of edges. The measure of network diameter indicates the longest distance between two nodes in the network, which depends on the pattern of how the nodes are connected. Once again, there we find major variations in the values for this metric (see table 4). Method X again provides the best fit with the 'actual' network, but Method 2 is not far off

with a diameter of 9. The last metric used for comparison is the average clustering coefficient, which indicates the level of clustering in the network (i.e. grouping of nodes based on dense relations among group members). Once again the results vary with Method X producing the closest fit with the 'actual' network (0.03 vs. 0.05). In summary, the analysis shows that of the four techniques, Method X produced a pseudo network (Network X) with structural network properties that most closely resembles the network properties of the 'actual' network.

	'Actual'	**Method 1**	**Method 2**	**Method 3**	**Method X**
Number of nodes	1,046	1,046	1,046	1,046	1,046
Number of edges	2,852	11,112	6,772	2,445	2,863
Graph density	0.003	0.010	0.006	0.002	0.003
Network diameter	10	6	9	18	10
Average clustering coefficient	0.05	0.79	0.51	0.10	0.03

Table 4. Descriptive network metrics (sample set 1)

4.2 Results of network correlation analysis

Results from the previous step show that pseudo network X has similar structural properties as the 'actual' network, but that does not necessarily mean that the networks are entirely similar. For example, it might be that the edges in pseudo network 1 connect different nodes from those in the 'actual' network. To check the extent of network similarity we performed a Pearson correlation analysis (based on QAP) to compare the networks. The results of this analysis are shown in Table 5.

Method	**Observed value**	**Significance**	**Average**	**Std. Deviation**
'Actual', Method 1	0.69	0.0002	-0.0000	0.0015
'Actual', Method 2	0.70	0.0002	0.0000	0.0014
'Actual', Method 3	0.54	0.0002	0.0000	0.0012
'Actual', Method x	0.89	0.0002	0.0000	0.0012
Method 1, 2	0.87	0.0002	0.0000	0.0020
Method 1, 3	0.64	0.0002	-0.0000	0.0015
Method 1, x	0.70	0.0002	-0.0000	0.0015
Method 2, 3	0.68	0.0002	-0.0000	0.0014
Method 2, x	0.71	0.0002	-0.0000	0.0013
Method 3, x	0.51	0.0002	0.0000	0.0012

Table 5. Pearson correlations using QAP (sample set 1)

The column labelled 'Significance' shows that all Pearson correlations were significant (using QAP). The most interesting results are in the column labelled 'Observed value' showing the Pearson correlations between the networks. In the first four rows the Pearson correlations of all pseudo networks with the actual network are presented, followed by the Pearson correlations among the pseudo networks. The results show that pseudo network X is most similar to the 'actual' network with a Pearson correlation of 0.89, but also Method 1 and 2 also score well with 0.69 and 0.70. Results of the Pearson correlation analysis, comparing the pseudo networks amongst each other, shows that pseudo networks 1 and 2 are most similar (see also Figure 3) and pseudo networks 3 and X are least similar.

4.3 Results of sensitivity analysis

Finally, to test whether the Pearson correlation results are stable and apply under different circumstances we applied sensitivity analysis. Firstly, we tested whether the results were influenced by indi-

vidual sub-categories having different discussion patterns. For this analysis we used sample 2 and applied the same Pearson correlation analysis to data from six different sub-categories. The Pearson correlation results for one of those sub-categories is shown in Table 6. Results for the other sub-categories show the same pattern: pseudo network X is most similar to the 'actual' network.

Sub category 101, Start-ups				
Method	**Observed value**	**Significance**	**Average**	**Std. Deviation**
'Actual', Method 1	0.73	0.0002	-0.0002	0.0135
'Actual', Method 2	0.66	0.0002	-0.0001	0.0134
'Actual', Method 3	0.49	0.0002	-0.0004	0.0133
'Actual', Method X	0.85	0.0002	0.0001	0.0140

Table 6. Pearson correlations using QAP (sample set 2)

Secondly, we tested whether the results were influenced by the length of the threads. For this analysis we used sample 3 that contains only threads with more than 10 messages and re-applied Pearson correlation analysis. The results of this analysis are shown in Table 7 and show that pseudo network 1 is now most similar to the 'actual' network. But the similarity, i.e. Pearson correlation score, is not as high as in the previous analyses (0.65 versus 0.85 and 0.89). Method X that performed best for threads with length 1-10 (sample 1) is now the worst performing method and is not suitable for communities where discussions are characterised by long threads.

Method	**Observed value**	**Significance**	**Average**	**Std. Deviation**
'Actual', Method 1	0.65	0.0002	0.0001	0.0060
'Actual', Method 2	0.66	0.0002	0.0001	0.0056
'Actual', Method 3	0.42	0.0002	0.0001	0.0045
'Actual', Method X	0.49	0.0002	-0.0000	0.0033

Table 7. Pearson correlations using QAP (sample set 3)

5 Discussion and implications for research

In this study, we examined whether different pseudo network creation algorithms found in the literature were able to generate reply-to networks that approximate the 'actual' reply-to networks of online communities and compared their accuracy. In total, four pseudo network generation algorithms were tested on data from an online community. The 'actual' reply-to network was derived from this data using content analysis. From the visual comparison and comparison of network metrics we found that Method 3 and Method X produced pseudo networks that most closely resembled the 'actual' network. The Pearson correlation analysis showed that all networks were significantly correlated with the 'actual' network but Method X showed a substantially higher correlation than the other methods.

The most important finding from this analysis is that it demonstrates that different pseudo network generation algorithms yield different results. Although there was some similarity between the structural properties of all pseudo reply-to networks and the 'actual' network, they were not identical and the amount of edges varied hugely. In particular Method 1, which assumes that everyone in a thread is (bi-directionally) connected with everyone else in the thread, results in many more edges. Accordingly, it assumes many edges that are not there in reality. Hence, method 1 favours people who post responses in many different threads; especially long threads, by rewarding them with too many relations. In research that focuses on out-degree centrality of persons (i.e. influential persons), those persons may be seemingly very central in the network while in reality they are not. Other research that is potentially affected by this method includes studies concerning the level of clustering in networks. Results in Ta-

ble 4 suggest that there is hardly any clustering in the 'actual' network, while the result for pseudo network 1 suggests the opposite.

Although Method X provided the best results in our analysis in approximating the 'actual' network, we cannot conclude that Method X is the best method in all situations. The analysis was performed using data from one online community, so results are difficult to generalize. It is important to note that the aim of our study was never to find a universal algorithm. Rather, our study aims to highlight how the results of applying different pseudo network creation algorithms vary, and to demonstrate why the selection of an algorithm for analysis is a such a critical decision in studies that rely on the analysis of unstructured data from online communities. In selecting which algorithm to use, we suggest that researchers should determine the type of community; for example Fisher et al. (2006) have demonstrated that a Discussion oriented and Q&A oriented community are characterised by different interaction patterns. This also follows from our own research where we introduced Method X after analysing and understanding the interaction pattern in the data from our online community. Furthermore, we found that Method X was not applicable to long threads. Those results show that the right choice for a pseudo network creation algorithm is highly dependent on the specific context.

Based on our observations we propose two important guidelines for research that involves unstructured data from online communities. Guideline 1 is that preliminary analysis of the data should inform the decision for a pseudo network creation algorithm. Preliminary analysis should reveal whether a particular interaction pattern can be found in the data. This pattern might depend, for example, on the type of online community (i.e. discussion or Q&A oriented) or the length of the threads. The analysis should result in the selection (or development) of an algorithm or a combination of algorithms, e.g. one for short and one for long threads. Guideline 2 is that researchers should compare the extent and quality of fit of the selected algorithm against the 'actual' network. For deriving the 'actual' network we suggest applying content analysis to a sample of the data. Furthermore, we recommend comparing the selected algorithm against other algorithms, for example those presented in Figure 1. This will reveal its performance in relation to more standard algorithms. The comparison will never result in a perfect fit but the researcher should be able to argue, given their particular research question, whether and why the chosen algorithm is good enough.

The proposed guidelines are based on the assumption that an algorithm should be selected that approximates the 'actual' network. In other words, it is not possible to determine the 'actual' network and an algorithm should be selected that generates a network that best resembles the 'actual' network. An alternative solution is to use Natural Language Processing (NLP) techniques to improve the inference of the reply-to relations between posters in an online community (Gruzd and Haythornthwaite, 2008). Using NLP one can try to inference a reply-to relation between posters based on the actual content of the message. This is similar to how we determined the 'actual' network, but we did it manually and with NLP this can be done automatically. The use of NLP is an interesting direction for future research.

In summary, our study and the resulting guidelines make a methodological contribution to studies of online communities in which pseudo network creation algorithms are used to generate reply-to networks from unstructured data (i.e. receiver of message is not explicitly known). Following the guidelines will add more rigor to such studies in our opinion because this will result in a more careful evaluation of the pseudo network creation algorithm selected. Guidelines are needed since our results show that different algorithms result in networks that have fundamentally different structural characteristics. Using an unsuitable algorithm might therefore lead to unjustified results, based on flawed reply-to network data, in an otherwise robust study.

References

Adamic, L. A., J. Zhang, E. Bakshy and M. S. Ackerman. (2008). "Knowledge sharing and yahoo answers." In: *Proceeding of the 17th international conference on World Wide Web - WWW '08* (p. 665).

Agarwal, R., A. K. Gupta and R. Kraut. (2008). "The Interplay Between Digital and Social Networks." *Information Systems Research*, *19*(3), 243–252.

Barnett, G. A. (2011). *Encyclopedia of social networks*. Thousand Oaks, CA: SAGE Publications, Inc.

Berger, K., J. Klier, M. Klier and A. Richter. (2014). "'Who is key...?' - Value adding users in enterprise social networks." In: *Proceedings of the 22nd European Conference on Information Systems (ECIS2014)* (pp. 1–16). Tel-Aviv, Israel.

Blanchard, A. L. and M. L. Markus. (2004). "The experienced 'sense' of a virtual community." *ACM SIGMIS Database*, *35*(1), 64.

Chiu, C.-M., M.-H. Hsu and E. T. G. Wang. (2006). "Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories." *Decision Support Systems*, *42*(3), 1872–1888.

Cranefield, J., P. Yoong and S. L. Huff. (2015). "Rethinking Lurking: Invisible Leading and Following in a Knowledge Transfer Ecosystem." *Journal of the Association for Information Systems*, *16*(4), 213–247.

Cross, R., A. Parker, L. Prusak and S. P. Borgatti. (2001). "Knowing what we know: Supporting knowledge creation and sharing in social networks." *Organizational Dynamics*, *30*(2), 100–120.

De Souza, C. S. and J. Preece. (2004). "A framework for analyzing and understanding online communities." *Interacting with Computers*, *16*(3), 579–610.

Ediger, D., K. Jiang, J. Riedy, D. A. Bader and C. Corley. (2010). "Massive Social Network Analysis: Mining Twitter for Social Good" (pp. 583–593). IEEE.

Falkowski, T., J. Bartelheimer and M. Spiliopoulou. (2007). "Mining and visualizing the evolution of subgroups in social networks." In: *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 52–58).

Faraj, S. and S. Johnson. (2011). "Network exchange patterns in online communities." *Organization Science*, *22*(6), 1464–1480.

Fisher, D., M. Smith and H. T. Welser. (2006). "You Are Who You Talk To: Detecting Roles in Usenet Newsgroups." In: *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*.

Freeman, L., M. Everett and S. Borgatti. (2015). "UCINET 6." Retrieved from https://sites.google.com/site/ucinetsoftware/home

Füller, J., G. Jawecki and H. Mühlbacher. (2007). "Innovation creation by online basketball communities." *Journal of Business Research*, *60*(1), 60–71.

Giles, J. (2012). "Making the Links." *Nature*, *488*(7412), 448–450.

Gleave, E., H. T. Welser, T. M. Lento and M. A. Smith. (2009). "A Conceptual and Operational Definition of ' Social Role ' in Online Community." In: *Proceedings of the 42nd Hawaii International Conference on System Sciences* (pp. 1–11).

Gómez, V., V. Gómez, A. Kaltenbrunner, A. Kaltenbrunner, V. López and V. López. (2008). "Statistical analysis of the social network and discussion threads in slashdot." In: *Proceeding of the 17th international conference on World Wide Web - WWW '08* (p. 645).

Gruzd, A. and C. Haythornthwaite. (2008). "The Analysis of Online Communities using Interactive Content-based Social Networks." *American Society for Information Science and Technology Conference (ASIS&T)*, *45*(1), 523–527.

Hanneman, R. and M. Riddle. (2005). "Introduction to social network analysis." Retrieved from http://faculty.ucr.edu/~hanneman/nettext/

Hara, N., C. J. Bonk and C. Angeli. (2000). "Content analysis of online discussion in an applied edu-

cational psychology course." *Instructional Science*, *28*, 115–152.
Helms, R. W. (2007). "Redesigning communities of practice using knowledge network analysis." In: A. S. Kazi, L. Wohlfart, & P. Wolf (Eds.), *Hands-On Knowledge Co-Creation and Sharing: Practical Methods and Techniques* (pp. 251–274). Knowledgeboard.
Helms, R. W. and C. M. Buysrogge. (2005). "Knowledge Network Analysis: a technique to analyze knowledge management bottlenecks in organizations." In: *Proceedings 6th International Workshop on Theory and Applications of Knowledge Management* (pp. 410–414). Copenhagen, Denmark.
Himelboim, I., E. Gleave and M. Smith. (2009). "Discussion catalysts in online political discussions: Content importers and conversation starters." *Journal of Computer-Mediated Communication*, *14*(4), 771–789.
Howison, J., K. Inoue and K. Crowston. (2006). "Social dynamics of free and open source team communications." *Proceedings of the IFIP Second International Conference on Open Source Software (Lake Como, Italy)*, *203*, 319–330.
Huffaker, D. (2010). "Dimensions of Leadership and Social Influence in Online Communities." *Human Communication Research*, *36*(4), 593–617.
Iriberri, A. and G. Leroy. (2009). "A life-cycle perspective on online community success." *ACM Computing Surveys*, *41*(2), 1–29.
Jacomy, M., T. Venturini, S. Heymann and M. Bastian. (2014). "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software." *PLoS ONE*, *9*(6), 1–12.
Kuk, G. (2004). "Selection, Cliques and Knowledge Sharing in Open Source Software Development Communities." In: *Proceedings of the IADIS International Conference Web Based Communities 2004* (pp. 140–148). IADIS Press.
Otte, E. and R. Rousseau. (2002). "Social network analysis: a powerful strategy, also for the information sciences." *Journal of Information Science*, *28*(6), 441–453.
Pannell, D. J. (1997). "Sensitivity analysis: strategies, methods, concepts, examples." *Agricultural Economics*, *16*, 139–152.
Petrovčič, A., V. Vehovar and A. Žiberna. (2012). "Posting, quoting, and replying: A comparison of methodological approaches to measure communication ties in web forums." *Quality and Quantity*, *46*, 829–854.
Phang, C. W., A. Kankanhalli and R. Sabherwal. (2009). "Usability and Sociability in Online Communities : A Comparative Study of Knowledge Seeking and Contribution." *Journal of the Association for Information Systems*, *10*(10), 721–747.
Preece, J. and B. Schneiderman. (2009). "The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation." *AIS Transactions on Human-Computer Interaction*, *1*(1), 13–31.
Ridings, C. M. and D. Gefen. (2004). "Virtual Community Attraction: Why People Hang Out Online." *Journal of Computer-Mediated Communication*, *10*(1).
Sack, W. (2000). "Conversation Map : An Interface for Very Large- Scale Conversations." *Journal of Management Information Systems*, *17*(3), 73–92.
Stanoevska-slabeva, K. (2002). "Toward a Community-Oriented Design of Internet Platforms." *International Journal of Electronic Commerce*, *6*(3), 71–95.
Toral, S. L., M. R. Martínez-Torres and F. Barrero. (2010). "Analysis of virtual communities supporting OSS projects using social network analysis." *Information and Software Technology*, *52*(3), 296–303.
Trier, M. and A. Richter. (2014). "The deep structure of organizational online networking - an actor-oriented case study." *Information Systems Journal*, n/a–n/a.
Turner, T. C., M. A. Smith, D. Fisher and H. T. Welser. (2005). "Picturing Usenet: Mapping Computer-Mediated Collective Action." *Journal of Computer-Mediated Communication*, *10*(4), article 7.
Wasserman, S. and K. Faust. (1994). *Social Network Analysis: Methods and Applications* (1st ed.). Cambridge University Press.

Yang, Z. and X. Fang. (2004). "Online service quality dimensions and their relationships with satisfaction." *International Journal of Service Industry Management*, *15*(3), 302–326.
Zhongbao, K. and Z. Changshui. (2003). "Reply networks on a bulletin board system." *Physical Review E*, *67*(036117), 1–6.