SparQs: Visual Analytics for Sparking Creativity in Social Media Exploration

Nan-Chen Chen, Michael Brooks, Rafal Kocielnik, Sungsoo (Ray) Hong, Jeff Smith, Sanny Lin, Zening Qu, and Cecilia Aragon

University of Washington, Seattle WA 98195, USA,
nanchen@uw.edu,
WWW home page: https://depts.washington.edu/hdsl

Abstract. Social media has become a fruitful platform on which to study human behavior and social phenomena. However, social media data are usually messy, disorganized, and noisy, which makes finding patterns in such data a challenging task. Visualization can help with the exploration of such massive data. Researchers studying social media often begin by reviewing related research. In this paper, we consider the idea that information from related research can be incorporated into social media visualization tools in order to spark creativity and guide exploration. To develop an effective overview of social media research with which to seed our tool, we conducted a content analysis of social media related papers and designed *SparQs*, a visual analytics tool to spark creativity in social media exploration. We conducted a pilot evaluation with three social media researchers as well as a participatory design workshop to explore further directions.

Keywords: visualization, visual analytics, social media, exploratory analysis, research questions, social science

1 Introduction

In the past decade, social media has become a useful platform on which to study human behavior and social phenomena. Many fields, among them sociology, communication, and epidemiology, leverage the richness of social media to investigate how different dimensions and elements (e.g., time, hashtags, and network connectivity) relate to their subjects of interest. However, social media data are usually messy, disorganized, and full of noise, which makes finding patterns within the data a challenging task. Visualization can be useful for the exploration of such massive data. Although numerous tools have utilized visualization to study social media data, few systems have focused on giving users an overview of the research field itself and on helping the users generate ideas for exploring the data.

Research is typically informed by or based on previous work. Researchers often begin their studies with a review of related work, so developing visualizations based on an overview of existing research may be able to spark creativity

and guide exploration. To develop an effective overview of social media research with which to seed our tool to inform such guidance, we conducted a content analysis of social media related papers. We collected 75 papers related to social media research and manually extracted research questions, dimensions, visualization type, analysis methods, data sources and scale. Based on the results from content analysis, we designed *SparQs* to present research questions along with the visualization of data distributions of tweets over user-specified dimensions. We then conducted a pilot evaluation with three social media researchers and a participatory design workshop to explore further directions of improvement for SparQs.

The contributions of this paper are three-fold: First, the results from content analysis on social media papers provide an overview of recent progress in social media research. Specifically, the extracted research questions, dimensions, and other properties can inform future system design to support social media research. Second, we present SparQs, a visual analytics tool that incorporates visualization with research questions for exploratory analysis. Last but not least, the outcomes from the pilot evaluation and participatory design indicate many potential research directions that extend the use of research questions in visual exploratory analysis.

2 Related Work

Visual analytics is "the science of analytical reasoning facilitated by interactive visual interfaces" [12]. The goal of visual analytics is to leverage visual channels to deliver synthesized information as a way to support analytical tasks [7]. As visualization is commonly used in exploratory data analysis (EDA) [13], visual analytics can further facilitate exploratory processes through carefully designed support for analytical tasks (e.g., automatically extracting potential points of interest, explicitly displaying commonly-used analysis functions).

Past research has attempted to use visual analytics for studying social media data. For example, Diakopoulos et al. created *Vox Civitas* for journalists to explore topics, sentiment, and keywords among tweets of an event [4], and Marcus et al. built *twitInfo* to automatically detect peaks of stream tweets and highlight important text to use in labeling these peaks [8]. Brooks et al. developed *Agave* for collaborative sentiment analysis among tweets of a specific event [1]. Chae et al. designed a location-based visual analysis system for disaster events using geolocation tweet information [2]. These examples all utilize visualization to display results and information about the analysis targets, but they explore only limited dimensions. Furthermore, none of these works consider social media literature as a medium that can inform design and guide the exploratory process.

To support exploration in early stages, when the dimensions of interests have not yet been decided, SparQs focuses on incorporating dimensions and research questions from previous social media literature. The goals are to discover unknown aspects of a dataset, and to spark creative ideas when examining the dataset.

3 Content Analysis on Social Media Papers

3.1 Process

To understand what research questions and dimensions are interesting to social scientists, we collected 75 papers from university library databases by searching on social media-related keywords (e.g., twitter, social media, social network) and filtered them to focus on social science-related papers only. We collectively conducted content analysis on a web interface (Fig. 1) where a paper's PDF file and analysis questions were shown on the interface. The analysis questions included the source and scale of the dataset(s), the research questions explored, variables, as well as the visualization and methods used in the papers.

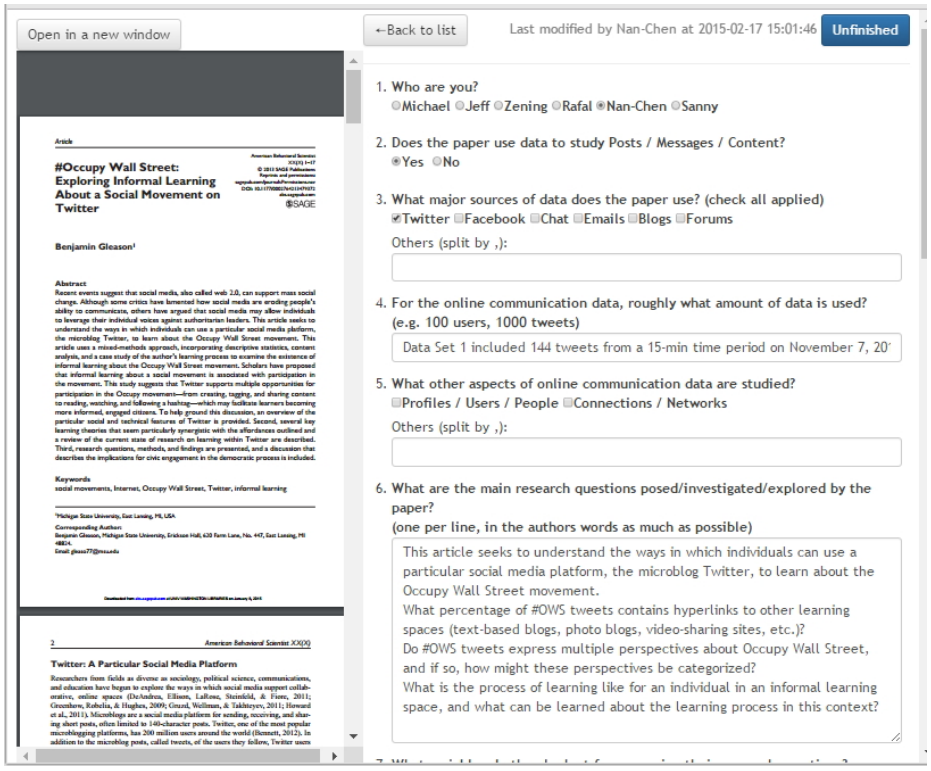

Fig. 1. Content analysis interface

Full list of questions used:

1. Does the paper use data to study Posts / Messages / Content? [Yes/No]
2. What major sources of data does the paper use? (check all that apply) [Twitter / Facebook / Chat / Emails / Blogs / Forums]

3. For online communication data, roughly what amount of data is used?
4. What other aspects of online communication data are studied? [Profiles / Users / People Connections / Networks / Others]
5. What are the main research questions posed/investigated/explored by the paper?
6. What variables do they look at to answer their research questions?
7. Is the paper *primarily* concerned with: (social phenomena includes individual, group, interactional, or otherwise human-related phenomena) [Offline social phenomena / Online social phenomena / Computational data processing technique / Research methodology]
8. In the authors' own words, what methods of analysis are applied to the online communication data? (e.g. manual/auto content analysis, machine learning, some type of modeling, close reading, qualitative analysis, etc.)
9. In your words, what methods of analysis are used? [Modeling (e.g. machine learning models, topic models...) / Statistical analysis (e.g. descriptive statistics, comparing two subgroups) / Social network analysis (e.g. centrality) / Human interpretation (e.g. qualitative coding, close reading)]
10. How are the results presented? [Simple charts and graphs / More complex visualizations / Tables / Quotations or excerpts / Statistical results / Narrative accounts]
11. Should we look at the visualizations? [Yes/No]
12. For each visualization in the paper, what is the primary question they answer?

We not only looked for explicit statements of research questions, typically in the introduction or methods sections, but also uncovered and collected implicit questions indicated in other sections. Dimensions of interest were sometimes explicitly referenced in research questions, but in many cases they also came from sections describing analysis, charts, and visualizations, as well as tables of results.

3.2 Results

About 350 dimensions, 250 research questions, and 140 visualizations were extracted. Selected examples of research questions and dimensions are shown in Table 1. We printed out the dimensions and research questions and sorted them into groups in a collaborative affinity diagramming activity. As a result, we created the dimension topology shown in Table 2. This topology is an effort to organize the dimensions into a structured form, so that we can create visualizations based on these dimensions. Furthermore, two of the authors further analyzed 56 questions in detail to rewrite them in a form less connected to the particular past research. They also extracted words representing specific dimensions of interest explored in the questions. These were later used to link the questions to the dimensions in the visualization. The full set of results can be found on https://github.com/hds-lab/sparqs-data.

Table 1. Example dimensions extracted from the research questions

Social science research question	**Dimensions**
How do Twitter users communicate their involvement with Haiti relief efforts? [10]	Qualitative labels (e.g. connecting, promoting, personalizing)
How do professional athletes use Twitter to communicate with fans and other players? [6]	Qualitative labels (e.g. interactivity, diversion, sharing, promotional, fan-ship)
To what extent does distance determine the informal communication of users from different nations? [5]	RT network, country, external data about countries

Table 2. Dimension topology

High-level Category	**Dimension**	**Open/Closed**	**Variable Type**	**Subtype**	**Range**	**Twitter-Specific**
Time	Time	Open	Quantitative	Time		
Time	Timezone	Closed	Nominal	-		
Contents	Topic (from topic model)	Open	Nominal	-		
Contents	Specific words in the message	Open	Nominal	-		
Contents	Specific hashtags in the message	Open	Nominal	-		
Contents	Contains a hashtag	Closed	Nominal	Boolean	Yes, no	
Contents	Contains a photo	Closed	Nominal	Boolean	Yes, no	
Contents	URL domain	Open	Nominal	-		
Contents	Contains URL	Closed	Nominal	Boolean	Yes, no	
Meta	Language (of a tweet)	Closed	Nominal	-		
Meta	Sentiment	Closed	Nominal	Small set	Positive, neutral, negative	
Interaction	Message type	Closed	Nominal	Small set	Original, retweet, reply	Yes
Interaction	Number of replies	Open	Quantitative	Frequency		
Interaction	Number of shares	Open	Quantitative	Frequency		Yes
Interaction	People mentioned in message (name)	Open	Nominal	People		
Author	Language (of an author's profile)	Closed	Nominal	-		
Author	Author of message (name)	Open	Nominal	People		
Author	Number of messages authored	Open	Quantitative	Frequency		
Author	Number of friends	Open	Quantitative	Count		Yes
Author	Number of followers	Open	Quantitative	Count		Yes
Author	Number of times replied to	Open	Quantitative	Frequency		
Author	Number of times mentioned	Open	Quantitative	Frequency		
Author	Number of times retweeted	Open	Quantitative	Frequency		Yes

4 SparQs

In this section, we describe SparQs, a visual analytics tool to support exploratory analysis on social media data and suggest creative research questions. SparQs leverages the dimension typology and research questions we extracted from the content analysis. The key idea is to enable users to explore the common dimensions and their combinations quickly through visualization, and also to display potentially relevant research questions along these same dimensions.

4.1 Visualizing Dimensions

The SparQs interface is shown in Fig. 2. The left panel lists 20 dimensions which are grouped into five high-level categories. Users drag and drop these dimensions to the rounded rectangle boxes in the middle panel to create visualizations (The red color indicates the primary dimension, whereas the blue is a secondary dimension). The visualization types with regard to dimension compositions are

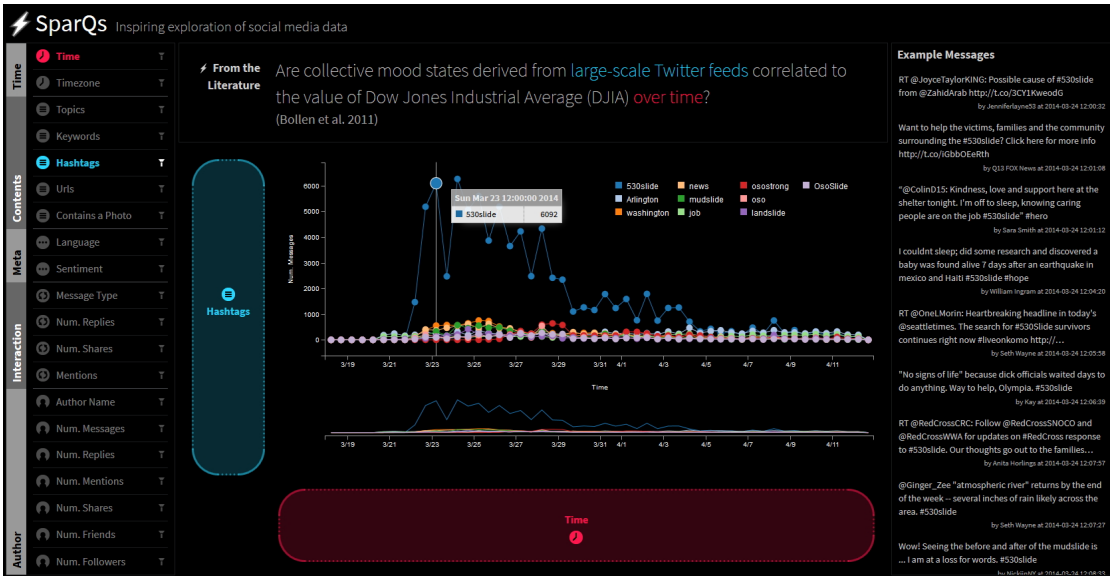

Fig. 2. SparQs. The left panel allows users to filter a list of dimensions from the typology. Visualizations are created and displayed in the middle panel. Users drag and drop variables or dimensions in the red and blue rounded rectangle boxes. The area above the visualization depicts a research question relevant to the current set of filters. Example tweets are displayed on the right.

shown in Table 3. Users also filter on any of the dimensions by opening the filtering box at each dimension (Fig. 3). For time-series plots, a focus+context view [3] is displayed in the middle panel where users brush to focus on a specific range of the quantitative dimensions. When mousing over a data point, a tooltip shows its corresponding values. Example tweets sampled based on the dimension compositions are displayed on the right panel. When clicking on the data points, the list of tweets is updated to tweets that belong to the corresponding point. For the sake of simplicity, SparQs only shows 10 levels of a categorical dimension in the visualization at a time (as in the dimension Hashtag in Fig. 2, which displays only the top 10 most frequent hashtags). Other levels are shown by filtering. For dimension "Topics", we modeled topics using Gensim [9] using the top keywords as topic names; for dimension "Sentiment", we used TextBlob [11] to label sentiments as positive, negative, or neutral.

4.2 Displaying Research Questions

When a user creates a visualization with dimensions in SparQs, a research question shows up in the top of the middle panel. The research question is randomly sampled from the set of research questions that examines the same dimension(s). The words corresponding to the dimension(s) are highlighted in the same color as

Table 3. Visualization types with regard to dimension topology

	Primary Dimension				
Secondary Dimension	*Time*	*Open Quant*	*Open Nominal*	*Closed Nominal*	*Boolean*
Nothing	Time series	Time series	Bar chart	Bar chart	Bar chart
Open Quant	Time series	Scatter-plot	Bar chart	Bar chart	Bar chart
Open Nominal	Multi-series line chart	Multi-series line chart	Grouped bar chart	Grouped bar chart	Grouped bar chart
Closed Nominal	Multi-series line chart	Multi-series line chart	Grouped bar chart	Grouped bar chart	Grouped bar chart
Boolean	Multi-series line chart	Multi-series line chart	Grouped bar chart	Grouped bar chart	Grouped bar chart

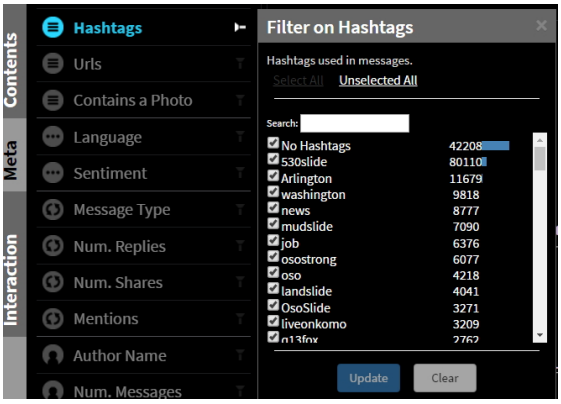

Fig. 3. SparQs filters. The view shows the filter on Hashtag.

the matching dimension(s). When users hover over the citation text, the details of the reference are displayed. In order to make the research questions understandable, two of the authors rewrote the questions based on their original text in the papers. For example, the original text for the research question in Fig. 4 is "we study how information spreads through the social network by measuring how the number of in-network votes a story receives, i.e., votes from fans of the submitter or previous voters, changes in time".

Fig. 4. Example research question for dimension "Time" and "Hashtag"

5 Pilot User Testing

5.1 Study Procedure

To evaluate SparQs, we conducted a pilot user test with three social media researchers. We loaded the tool with a Twitter dataset containing 685,311 tweets about the 2014 Oso mudslide in Washington State, USA. All the social media researchers for the study were familiar with the dataset. We invited them to test SparQs individually in one-hour sessions. We first introduced the interface and then let them use the tool to come up with potential research questions while seeking for interesting or unexpected insights regarding the dataset. All the participants were asked to think aloud during their sessions. The study moderators took notes and audio and screen recordings of the sessions.

5.2 Results

Listed dimensions helped users explore aspects they had not considered All the participants tried to look at all the dimensions SparQs provided, and they were able to discover a few patterns that they did not notice before. One participant raised a question to further look into what types of accounts receive more positive sentiment. Another participant was wondering how hashtags were used between different groups of accounts. The participants liked the ability to combine dimensions to create plots, but two mentioned it would be more useful if they could create customized groups.

Research questions were not directly useful We noticed most participants did not spend much time reading the research questions; according to their explanations, the research questions seemed irrelevant to the dataset. Some questions were not even from the same discipline as theirs, and thus they did not see why those questions were important.

The need to incorporate prior knowledge During the sessions we found that all the participants looked for something that came from their prior knowledge. For example, since the Oso dataset was about a disaster, they wanted to look at specific types of accounts such as governments or non-government organizations. They also wanted to compare the tweeting behavior among people who were or were not victims of the disaster. These inquiries all went beyond dimensions that we could directly derive from the dataset; this indicates the need to enhance our tool to incorporate prior knowledge and other sources of information.

6 Participatory Design Workshop for Future Directions

In order to explore ways to improve the use of the research questions and SparQs, we held a participatory design workshop with four social media researchers to explore potential extensions of the tool. One of the researchers participated in the pilot user test, but the other three were new to SparQs.

6.1 Process and Materials

We invited the four participants to a conference room in our department building, explained the background and goals of the workshop to them, and provided each of them with a stack of ideation resource printouts, including tweets from the Oso mudslide dataset as well as visualizations, titles, abstracts, and research questions from papers. We asked them to use these materials to brainstorm questions and directions that they would want to investigate further. To structure the brainstorming session, we proceeded in 5-minute sprints, and after each sprint we asked them to briefly describe their ideas, and then continue the brainstorming. We ended up running four sprints with a brief discussion after each sprint. The whole brainstorming session continued for about 40 minutes, after which we asked them to reflect on the experience and describe what they found useful during the session. Fig. 5 shows an example set of sketches and notes along with the materials provided from the workshop.

Fig. 5. A photo of the sketches and notes from the participatory design workshop

6.2 Results

Use and comments on the provided materials During the brainstorming session, the four participants approached the materials with very different strategies. One

participant primarily focused on reading research questions and sometimes the tweets, whereas another participant used many tweets and only some research questions. One other participant only flipped through the provided materials, and he later explained that he was very familiar with the tweets and he was thinking about some directions on his own. Other findings include that the abstracts were not used much due to limited time, and tweets with photos got more attention where the participants described them as "attractive". Based on the comments, we found it is important to let users of exploration tools like SparQs directly read the text and images during exploration.

Integration with qualitative coding and other types of analysis Two of the participants mentioned the desire to do qualitative coding. From their perspectives, qualitative coding is a common task for them during the exploration stage. Some of the research questions we provided had manually coded categories as targeted research dimensions, and our participants pointed out that they were also interested in categories that were not standard and emerged from the data. Another participant with experience in network analysis suggested an interface which combined tweets, a follower-followee network, and code (bottom-right sketch in Fig. 5). As a result, incorporating both qualitative coding and other types of analysis with SparQs is a valuable direction for future research.

Diverse context In the final reflection section, the four participants agreed that the research questions were not very helpful because they were high-level and not exactly relevant to disaster-related research. However, one participant pointed out that some of the research questions were necessary to examine because they were basic. Another participant commented that it was still fascinating to read these research questions that came from very different research contexts. These points indicated that research questions may be useful during exploration, but we need better ways to draw research questions that are closer to the user's research context. Therefore, we suggest that future research should focus on building a system that can automatically identify research questions that are relevant to a user's research interest, and adaptively take into account exploration logs for better recommendations.

7 Conclusion

In this paper, we presented SparQs, a visual analytic tool for exploratory analysis on social media data which lays out research dimensions and questions from social media literature. We conducted a pilot user test as well as a participatory design workshop to examine the tool. The results showed that incorporating information from literature can be valuable, but more study is required to effectively use extracted questions from past research. Future work should explore in-depth automatic analysis on structuring the information and incorporation with other methods such as qualitative coding and network analysis. Additionally, the dimension topology we constructed from the literature can be useful to inform the design of exploratory tools for social media.

8 Acknowledgments

We would like to thank everyone who participated in the user testing and the participatory design workshop, as well as the research group who shared the Twitter dataset with us for testing the tool.

References

1. Brooks, M., Robinson, J.J., Torkildson, M.K., Aragon, C.R., et al.: Collaborative visual analysis of sentiment in twitter events. In: International Conference on Cooperative Design, Visualization and Engineering. pp. 1–8. Springer (2014)
2. Chae, J., Thom, D., Jang, Y., Kim, S., Ertl, T., Ebert, D.S.: Public behavior response analysis in disaster events utilizing visual analytics of microblog data. Computers & Graphics 38, 51–60 (2014)
3. Cockburn, A., Karlson, A., Bederson, B.B.: A review of overview+ detail, zooming, and focus+ context interfaces. ACM Computing Surveys (CSUR) 41(1), 2 (2009)
4. Diakopoulos, N., Naaman, M., Kivran-Swaine, F.: Diamonds in the rough: Social media visual analytics for journalistic inquiry. In: Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on. pp. 115–122. IEEE (2010)
5. García-Gavilanes, R., Mejova, Y., Quercia, D.: Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. pp. 1511–1522. ACM (2014)
6. Hambrick, M.E., Simmons, J.M., Greenhalgh, G.P., Greenwell, T.C.: Understanding professional athletes use of twitter: A content analysis of athlete tweets. International Journal of Sport Communication 3(4), 454–471 (2010)
7. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. In: Information visualization, pp. 154–175. Springer (2008)
8. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: Twitinfo: Aggregating and visualizing microblogs for event exploration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 227–236. CHI '11, ACM, New York, NY, USA (2011), http://doi.acm.org.offcampus.lib.washington.edu/10.1145/1978942.1978975
9. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
10. Smith, B.G.: Socially distributing public relations: Twitter, haiti, and interactivity in social media. Public Relations Review 36(4), 329–335 (2010)
11. Textblob (2016), https://github.com/sloria/TextBlob
12. Thomas, J.J.: Illuminating the path:[the research and development agenda for visual analytics]. IEEE Computer Society (2005)
13. Tukey, J.W.: Exploratory data analysis (1977)