SparkClouds: Visualizing Trends in Tag Clouds

Bongshin Lee, Nathalie Henry Riche, Amy K. Karlson, and Sheelagh Carpendale

Abstract—Tag clouds have proliferated over the web over the last decade. They provide a visual summary of a collection of texts by visually depicting the tag frequency by font size. In use, tag clouds can evolve as the associated data source changes over time. Interesting discussions around tag clouds often include a series of tag clouds and consider how they evolve over time. However, since tag clouds do not explicitly represent trends or support comparisons, the cognitive demands placed on the person for perceiving trends in multiple tag clouds are high. In this paper, we introduce SparkClouds, which integrate sparklines [23] into a tag cloud to convey trends between multiple tag clouds. We present results from a controlled study that compares SparkClouds with two traditional trend visualizations—multiple line graphs and stacked bar charts—as well as Parallel Tag Clouds [4]. Results show that SparkClouds' ability to show trends compares favourably to the alternative visualizations.

Index Terms—Tag clouds, trend visualization, multiple line graphs, stacked bar charts, evaluation.

1 Introduction

Tag clouds are a text-based visual depiction of tags (or words), typically used to display the relative tag frequency, popularity, or importance by font size. They can also serve as a visual summary of document content. In the last decade, tag clouds have proliferated over the web. They are now a common visualization in news sites for displaying the most active news story themes [1], photo sharing sites for conveying the distribution of image content [15], and social bookmarking sites for showing popular tags [6]. In fact, several online programs are available that help you create your own tag clouds from different types of text sources [20][21][22][29].

Tag clouds can evolve as the associated data source changes over time. For example, the US Presidential Speeches tag cloud shows the popularity, frequency, and trends in the usages of words within speeches, official documents, declarations, and letters written by the Presidents of the US between 1776 and 2007 [24]. Other sources of highly dynamic content include online news and photo-sharing sites which serve freshly uploaded and tagged material every day. Interesting discussions around tag clouds often include a series of tag clouds and consider how they evolve over time. However, while tag clouds seem to invite exposure of their evolution over time, they do not explicitly represent them. This results in a significant cognitive demand on people who want to understand how a tag cloud evolved.

In this paper, we introduce SparkClouds (Fig. 1), a new breed of tag cloud that incorporates sparklines [23] with more typical tag cloud features to convey evidence of change across multiple tag clouds. We also present a controlled study that we conducted to compare SparkClouds with Parallel Tag Clouds (PTCs) [4] (the only previous tag cloud visualization designed for understanding multiple tag clouds), as well as two traditional trend visualizations—multiple line graphs and stacked bar charts. We compared these four visualizations in terms of speed and accuracy in supporting three types of tasks (specific data, topic trends, and overview). We found that SparkClouds' ability to show trends compares favourably to the alternative visualizations. Participants also preferred SparkClouds to stacked bar charts and PTCs.

We organize this paper as follows. In the next section we outline the related work to provide context for our description of SparkClouds, which follows in Section 3. Section 4 describes the design and results of the controlled study along with the alternative visualizations we used in the study. We then conclude the paper with a discussion of the lessons learned from the study and suggestions for future work.

Fig. 1. SparkClouds showing the top 25 words for the last time point (12th) in a series. 50 additional words that are in the top 25 for the other time points can be (top) filtered out or (bottom) shown in gray at a smaller fixed-size font. (bottom) is used in the study.

2 Related Work

The origin of tag clouds goes back to 1976 when an experiment was carried out by Stanley Milgram [14]. A collective "mental map" of Paris was created using font size to show how often each place was mentioned as a landmark in the city. In 1997, Search Referral Zeitgeist was created by Jim Flanagan as a way to visualize the number of times a term was used to find a given website by font size. Among high-profile websites, Flickr [15] used tag clouds first, followed by other Web 2.0 sites (e.g., Del.icio.us [6]) [3]. For more details about the history of tag clouds, see [25].

Due to their astonishing popularity, there have been many efforts in exploring various properties of tag clouds. Several websites now enable people to create their own tag clouds from different types of text sources [20][21][22][29]. One interesting variation, showing two-word phrases, provides a quite different perspective of the text

- *Bongshin Lee, Nathalie Henry Riche, and Amy K. Karlson are with Microsoft Research, E-Mail: {bongshin, nath, karlson}@microsoft.com.*
- *Sheelagh Carpendale is with the University of Calgary and performed this work while at Microsoft Research, E-Mail: sheelagh@cpsc.ucalgary.ca.*

Manuscript received 31 March 2010; accepted 1 August 2010; posted online 24 October 2010; mailed on 16 October 2010.
For information on obtaining reprints of this article, please send *email to: tvcg@computer.org.*

by revealing themes in the content [13]. There has been considerable research to improve tag cloud layouts. Kaser and Lemire organized tags in nested tables for HTML based sites by using an Electronic Design Automation (EDA) packing algorithm [11]. Seifert et al. proposed a new algorithm to address several issues found in the traditional layouts [18]. It creates compact and clear layouts by reducing whitespace and featuring arbitrary convex polygons to bound the terms. Tree Cloud arranges words on a tree to reflect their semantic proximity according to the text [7]. Tag Maps employs a unique layout based on real geographical space [10]. Wordle provides remarkably distinctive layouts by utilizing typography, color, and composition to balance various aesthetic criteria [26][29].

Research efforts that attempt to understand the effectiveness and utility of tag clouds generally fall into one of two categories; those which investigate the visual features of tag clouds and those which compare tag clouds with different layouts. Bateman et al. compared nine visual properties of tag clouds for their effects on visual search for tags [2]. Their results show that font size and font weight have stronger effects than others such as color intensity, number of characters, or tag area. Rivadeneira et al. conducted two experiments [16]. In the first study, they examined the effect of font size, location, and proximity to the largest tag, asking participants to recall terms (for 60 seconds) that were previously presented in tag clouds (for 20 seconds). In the second study, they investigated the effect of both font size and word layout on users' abilities to form an impression (gist). From both studies, in accordance with previous research, they observed a strong effect of font size.

Lohmann et al. compared four tag cloud layouts for three types of search task; one of them was a standard list using uniform font size with wrapping [12]. All but the traditional tag cloud worked best for one task. For the task of finding a specific tag, the list performed better than the tag cloud. Halvey and Keane compared tag clouds with traditional lists (horizontal and vertical), each with regular vs. alphabetical order by asking participant to find a specific tag [8]. They found that lists perform better than tag clouds and that alphabetical order further accelerates the search speed. Sinclair and Cardew-Hall conducted an experiment to investigate the preference between a tag cloud and a traditional search interface both for general browsing and for information seeking tasks [19]. They found that participants preferred the search interface for specific information retrieval tasks whereas the tag cloud was preferred for more open-ended browsing tasks.

In use, tag clouds can evolve as the associated data source changes over time. Interesting discussions around tag clouds often include a series of tag clouds and consider trends of their tags over time. This desire to study trends and understand how text content or topics evolve over time has been the purpose of other visualizations such as the commonly used line graphs and bar charts. However, despite the significant amount of research on tag clouds, there has not been much research on how to visualize trends in tag clouds. Many Eyes allows people to compare two texts in a single tag cloud [13]. It uses two colors (one per tag cloud) and pairs the tags that appear in both clouds. While it enables easy comparison between two tag clouds, this technique does not offer help in understanding trends over time as it is limited in the number of tag clouds visualized. Tagline Generator allows people to generate a sequence of tag clouds that are associated with time, from a collection of documents [21]; a dynamic slider control is used to navigate the time points, but only one tag cloud is shown at a time. Parallel Tag Clouds (PTCs) is designed to provide an overview of a document collection by incorporating graphical elements of parallel coordinates with the text size encoding of traditional tag clouds [4]. While PTCs do show multiple clouds simultaneously, they do not explicitly represent trends, and thus comparing multiple tag clouds to ascertain trends places the cognitive demands on the person.

Cloudalicious is an online tool specifically designed to visualize how tag clouds develop over time [17]. For a given website, it downloads the tagging data from del.icio.us [6] and then graphs the collective users' tagging activities over time using multiple line graphs. While Cloudalicious clearly shows some trends, such as decay of the collective usage of tags, it may suffer from overlapping of lines and does not retain the visual appearance of tag clouds. Dubinko et al. presented a new approach to visualize the evolution of tags in Flickr using an animation via Flash in a web browser [5]. While they allow people to observe and interact with the tags, their main contributions were not focused on the visualization but rather on algorithms and data structures to generate the list of "interesting" tags for a specific time period.

To better convey the evidence of change across multiple tag clouds, we developed a new breed of tag cloud called SparkClouds that integrates sparklines [23] into a tag cloud. We also conducted a controlled study to explore the efficacy of SparkClouds by comparing it with two traditional trend visualizations, multiple line graphs and stacked bar charts, as well as with PTCs.

3 Designing SparkClouds

The basic idea behind SparkClouds is to retain the advantages of tag clouds while incorporating minimal but sufficient indication of trends for a reasonable number of related tag clouds. In particular, we focused on these advantages of tag clouds:

- Compact use of space that can be flexibly reorganized into different aspect ratios without negatively impacting the readability of the cloud as a whole.
- Tag (or word) readability in that the importance or frequency of a tag is encoded directly in the size of the word.

Since we based our design of SparkClouds on two usage scenarios, we first describe these scenarios to provide the setting for our design development.

3.1 Scenarios

3.1.1 Keeping track of different non-familiar tag clouds

Jason is a stock market analyst. Every day, he has the same morning ritual. He spends about an hour on the web to absorb the information required to keep up-to-date with the market. Is the new product everyone is talking about out already? What is the next big application that mobile users are talking about? What are the reviews on the latest phone? Jason has dozens of websites bookmarked, but in fact, the majority of them are irrelevant today. Indeed, he already knows a great deal of information and may already have consulted them in recent days. The real challenge for him is to identify and select which sites to dive into, to find where the new information is, and to locate any deeper information he may need. Currently, Jason's strategy is to read RSS feed titles and browse a dozen or so websites, many of which conveniently present the current topics in a tag cloud form. Jason likes these simple representations as they give him a gist of the content of the website. He often remembers the large tags, but he still has trouble spotting the new tags and topics, as well as keeping track of the less popular ones, especially because he sees many different tag clouds every day.

3.1.2 Monitoring familiar tag clouds

Laura is a researcher working as part of a 16-person team that is rapidly increasing in size. Right now, she works closely with 5 of her team members and is relatively well-aware of their activities. But, as she juggles increasing numbers of projects, she realizes that she cannot keep up-to-date with the whole team. For example, it is quite challenging for her to keep track of what is happening from the monthly meetings and weekly status reports. Indeed, she recently learned from her manager that she spent several days surveying a topic that one of the team members had already been working on for the past few weeks. To maintain awareness of the team activity, she generates tag clouds from the weekly status reports. While they help her remember the key projects, she still has a difficult time noticing what has changed week-to-week, much less over longer spans of time. Being able to compare these project tag clouds as they evolved might have helped highlight her colleague's shift in their work focus.

3.2 Design Rationale Summary

A tag cloud is a visual representation that a broad range of people can easily decode, and makes effective use of display space. Our primary goal with SparkClouds is to preserve these two characteristics, while incorporating the ability to convey trends. We also aim at supporting the two usage scenarios described above.

- Jason is dealing with a large number of tag clouds; these may change radically and Jason may not remember all topics in previous tag clouds. He needs to have an *overview* of the trends to understand the market at a high level.
- Laura is familiar with the data. She has a good memory of the previous tag clouds. However she may need to dive into previous tag clouds *time-point-by-time-point* and compare them to get a deeper understanding of what each person did over time.

A SparkCloud encodes the popularity of tags by font size, as do standard tag clouds. To show the trends in popularity of each tag over time, we introduce a second visual element adjacent to each tag: a sparkline, *.i.e.* a minimal simplified line chart.

3.2.1 Tag Clouds as Used in SparkClouds

The tag cloud aspect of SparkClouds has two design parameters: the font size encoding and layout.

Font Size Encoding: In traditional single tag cloud representations, the font size used in depicting a word is typically determined as a function of its frequency, using either a linear or square root transformation [26]. When considering more than one tag cloud, such as those depicting multiple points in time, trends can occur across the tag clouds. However, because different tag clouds often have different word frequency ranges, font size encoding must be considered carefully. To support both overview and time-point-by-time-point scenarios, we define two methods of computing the font size encoding for SparkClouds.

The first method encodes the font size of each tag with the frequency of the tag over the entire time period. This provides an overview of the aggregated changes across time. Using this view, Jason is able to identify at a glance the most popular topics for the entire period (i.e., the largest tags). However, this method does not fully support comparisons of the frequency of tags within a given time point. For instance, the tag cloud for a given time point might contain tags of predominately similar sizes. This can make comparisons within a time point difficult. Furthermore, as tag clouds for new time points are added to the visualization, the previous size encodings may need to be recomputed. Thus the appearance of a given time point tag cloud may change, confusing Laura and interfering with her memory of prior clouds.

In the second method, the font size is used to encode the frequency of tags at a given time point. In this view, Laura is able to identify at a glance her team members' most popular topics for each time point. This view is also more appropriate for routine review, as it most clearly depicts the activity at a given time point. With this method, the font size is stable for each tag in that it does not change whether the visualization shows either one or several tag clouds. However, this encoding makes tags between tag clouds comparable only in rank not in frequency.

While SparkClouds, by default, use per time point font size encoding, they allow people to select a more appropriate encoding (or normalization) method according to their task as PTCs do.

Layout: As described above, there has been much research on laying out tag clouds [10][11][12][18]. Even though we present the design of SparkClouds as having an alphabetical order throughout the paper, it can support any existing tag cloud layout because SparkClouds retains the same structure as conventional tag clouds.

3.2.2 Sparklines

The most common visual encodings of trends over time are the traditional line graphs [17] and more recently stacked graphs [9][27]. We selected the simpler one of the two, sparklines, for use in visualizing trends over time in SparkClouds. Sparklines are simplified line graphs in the sense that axes are implied rather than explicitly drawn or labelled. They can be very compact and still provide an indication of a trend. As was also shown in [28], this property made them attractive to use as they can be inserted adjacent to each tag without cluttering the entire presentation.

Data Encoding: In SparkClouds, a sparkline depicts the popularity of the tag (vertical axis) over time (horizontal axis). To maintain consistency in the representation, the vertical axis of sparklines does not encode the raw popularity of tags, but instead uses a linear transformation function based on the relative popularity of tag. A potential refinement to this approach could be to take advantage of the extra precision offered by sparklines—if two words functionally map to the same font size, the sparkline can indicate whether one is slightly more frequent than the other. This may be useful when precisely comparing trends but we thought it was not needed for our current usage scenarios.

Representing Zero: To help people make comparisons between trends, we depict the horizontal axis of each sparkline. This transforms the sparkline into a sparkarea by filling the space between the sparkline and the horizontal axis using a light gradient blue color (Fig. 1). This visual encoding helps in comparing sparklines that are not horizontally aligned with each other and gives additional visual assistance in identifying the periods during which a tag was not present. By glancing at the sparkline below a given tag, Jason and Laura can assess if it is new, if its popularity has been stable or when it experienced a spike in popularity.

3.2.3 Unifying Tag and Sparkline

Given that we introduced an additional visual element for each tag, we broke the "homogeneity" of the tag cloud. To help people perceive the two visual elements (tag and sparkline) as a single unit of information, we considered several possibilities (Fig. 2).

Alignment: Sparklines can be placed before (Fig. 2a), after (Fig. 2b), above (Fig. 2c) or below the tag (Fig. 1). We experimented with using a script (handwriting) font to provide visual continuity, but we thought that the legibility of the tag was compromised (Figs. 2a-b). We also tried using an explicit baseline and aligning both text and sparkline on the baseline, but decided against this idea as it introduced too much clutter.

Overlays: We explored overlaying the tag over the sparkline (Fig. 2d). These representations appeared too cluttered and did not work well in practice since words vary widely in length; this made it difficult to compare points in time between two words whose sparklines spread across different scales. In addition, we thought that text overlays, even with light font and white outlines, made the words less readable (especially when the font size is small).

Mirror: To avoid the clutter caused by the overlays, we tried to mirror the line chart and place it under the word (Fig. 2e). We decided against this option as we believed it may compromise the decoding of the sparkline and thus be misleading. Indeed, even with various background and line-color options, we tended to interpret the mirrored sparkline as standard non-mirrored ones.

Foreground and background colors: We first explored using two foreground colors, which alternated between adjacent terms (but coloring the tags and associated sparkline with the same color). This solution worked quite well, but it looked more cluttered than the one with a single color. We also thought that we could use color more effectively for other purposes. In Laura's scenario, for example, we

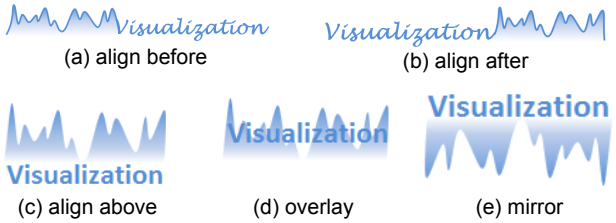

Fig. 2. Design alternatives for unifying tag clouds and sparklines.

believed that different colors could be used to identify specific individuals on the team in the aggregated tag cloud view. Thus, rather than focusing on color, we investigated solutions which take advantage of the background of the representation. We divided the tag clouds in a grid and colored the background of the cells with different colors. However, the resulting visualization looked too cluttered. To reduce the clutter, we left white space behind the tag and the sparkline, leaving only the SparkCloud's "skeleton" colored. While this solution provided better results, the background still looked too "overwhelming." Finally, we generated the opposite visual effect: we placed a circle with a faint gradient (from a colored center to white periphery) in the background of the tag and sparkline (Fig. 1). We thought this effect gave the display the best sense of both unity and texture. To make sure that the text was readable, we outlined the tag with a thin white border.

4 CONTROLLED EXPERIMENT

The goal of the controlled experiment was to explore how efficiently people could learn and use SparkClouds as compared to three alternative visualizations (two traditional trend visualizations and the existing PTCs) for different types of tasks.

4.1 Participants

Seventeen (7 female) volunteers from the Greater Puget Sound region in Washington State participated in the study. We screened participants to ensure all were familiar with traditional tag clouds. We also screened them for color blindness and required normal or corrected-to-normal vision. The average age of participants was 34.4 (36.4 for males, 31.4 for females), ranging from 21 to 47 years of age. Participants were given a software gratuity for their participation.

4.2 Tasks and Data

We tested the four visualizations across six different tasks. The first two tasks focus on finding specific data within the dataset, either one time point or one tag (specific data tasks). The next two focus on understanding topic trends for two or more continuous time points (topic trends tasks). The last two focus on perceiving trends across all time points (overview tasks). Table 1 summarizes the six tasks we used in the study. The specific data tasks, which are relatively simpler than the others, are open-ended questions that require a text-based response. For other types of real-world tasks, people may need to review all tags to understand overall trends and get an overview of the dataset. However, we suspected that the time to perform an exhaustive exploration (e.g., move the mouse over all individual tags) would dwarf the time needed to perceive trends. Therefore, we decided to provide four possible choices and ask participants to select the one which answers the task question most accurately.

Table 1. Six Study Tasks

#	Type	Task
1	Specific Data	What topic was ranked as the most frequent topic (top-ranked) during month **X**?
2	Specific Data	During which month did topic "**A**" occur most frequently?
3	Topic Trends	Which of the following topics experienced the longest continuous **increase**/**decrease** in rank during the year (e.g., for the most months in a row)?
4	Topic Trends	Which of the following topics experienced the largest **increase**/**decrease** in rank from month Y to month Z?
5	Overview	Which of the following topics was the most frequent topic (top-ranked) for the greatest number of months during the year?
6	Overview	Which of the following topics was ranked among the top 25 topics most consistently over the year (e.g., was ranked ***at all*** during the ***most*** months)?

When preparing datasets for a study, there are often trade-offs between using real-world datasets (ensuring ecological validity) and preparing datasets that allow task difficulty and other properties, such as task isomorphism, to be controlled. Since the primary goal of the study was to compare SparkClouds with other visualizations, we created the tag clouds in a systematic way to achieve tasks of similar difficulties across datasets. We generated 5 sets of tag clouds (one for each of the four visualizations and one for the practice), each with 12 time periods representing months. We first selected the top 75 most frequent words from the first chapter of 5 well known books; stop words are excluded. We then randomly generated the number of occurrences (ranging from 1 to 100) of each word for 12 months. To ensure that each trial has a unique answer, we manually tweaked some of the automatically generated frequencies. For the tasks with alternative choices, the four options were manually chosen from the top 25 words for the selected month. Across six tasks, each trial was initially presented with a different default month selected.

4.3 Visualizations

We compared four visualizations: SparkCloud, Multiple Line Graph, Parallel Tag Cloud, and Stacked Bar Chart (see Figs 1, 3, and 4). We chose multiple line graphs and stacked bar charts because they are commonly used in many tools to show trends (e.g., [9][17][27]). We chose Parallel Tag Clouds because they are the only tag cloud specifically designed to support cross tag cloud comparison for more than two tag clouds.

While each of these visualizations is inherently different from one another, when possible we implemented the visualizations based on the same visual guidelines. For example, words are presented in an alphabetical order in all four visualizations. SparkClouds and Parallel Tag Clouds share the same range of font sizes (from 10 to 34), and Multiple Line Graphs and Stacked Bar Charts use the same font size (15). For all visualizations, all 75 top words were shown, clearly indicating which were the top 25 words for each month. In Parallel Tag Clouds, since the top 25 words for each month are displayed simultaneously in the column for their month, all words are shown and some words are shown in multiple columns. The three other visualizations display all the top 75 most frequent words and highlight only the words that were included in the top 25 for the selected month.

4.3.1 Multiple Line Graph (MultiLine)

A line graph explicitly shows trends (i.e., how a value changes) by connecting a series of successive data points; usually the x-axis represents time. While a multiple line graph helps people compare trends between multiple variables, it often suffers from overlapping when many lines are displayed at once. To alleviate this problem, we implemented the multiple line graph in the following way. While all the top 75 most frequent words are displayed, only the words that were included in the top 25 for the selected month are highlighted (Fig. 3a). Participants can select a time point by clicking on the label shown at the bottom of the visualization. The currently selected time point is marked with blue background and a thick border. When participants move the mouse over a line or a word in the legend list box at the top, we highlight only the focused word and the line. When the values overlap, we slightly shift the data point diagonally (2 pixels to the right and below). We used 5 distinctive colors to help participants better differentiate the lines.

4.3.2 Parallel Tag Cloud (ParallelCloud)

Parallel Tag Clouds provide a visualization for comparing a document collection by incorporating graphical elements of parallel coordinates, but using the font size encoding of traditional tag clouds (Fig. 4). Each column represents a list style tag cloud for one time point. We implemented Parallel Tag Clouds' basic visualizations and interactions techniques required to complete the tasks used in the study, based on the description in [4]. We draw small truncated links (stabs) that point to the next occurrence of the word, in the color blue to indicate the existence and location of the same word in other tag clouds for different time points. When participants move the mouse over a word, we display a gradient line that links the same word occurring in multiple tag clouds (Fig. 4). Note that we remapped a

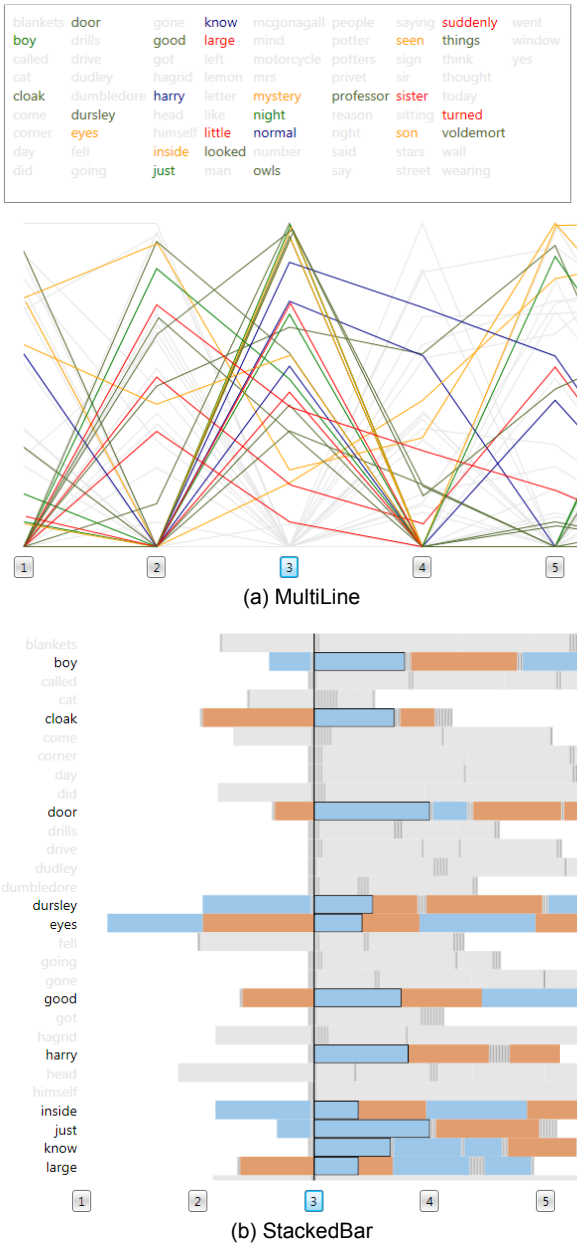

Fig. 3. MultiLine and StackedBar highlight only the words that are included in the top 25 for the selected time point, marked with blue background. They are cropped to fit in the paper.

small number of the words from the original text sources to shorter variations to avoid the need for horizontal scrollbars; this was necessary because the mouse-over interaction fails in Parallel Tag Clouds when scrolling is required. Furthermore, during the trials for topic trends and overview tasks—those that offered participants four possible options from which to choose—the label of the month which contained all four terms was highlighted with a thick dark blue border so that participants did not have to perform a tedious visual search over the entire set of tag clouds to find the four tags to compare (Fig. 4).

4.3.3 Stacked Bar Chart (StackedBar)

Stacked bar charts, in which bars are divided into nominal variables, are commonly used to show trends. ThemeRiver [9] is a timeline indicating the flow of document themes. It uses width of the river to show the number of documents and the river is sub-divided by topics, which ebb and flow over time. The Name Voyager visualizes a graph of the popularity of baby names over the past century [27]. Both of these examples use smoothly connecting lines between the data points. In our implementation we kept the adjacent bars discrete to make individual data values more readable.

While stacked bar charts are particularly good for conveying at cumulative and overall trends, we identified two major issues with standard stacked bars. First, they may be deceiving for evaluating tends across time for a single term because at any given time point the placement of neighboring terms interferes with (or rather, has a displacement effect on) the placement of the series of interest. Second, label placement is non-trivial because each tag requires a position that is large enough to ensure the label is readable, which is not always possible if a tag generally has a low popularity; furthermore, labels are not guaranteed to be vertically or horizontally aligned, making scanning difficult. Both of these issues are problematic for tag clouds since readability is of critical importance to investigate trends in tags over time. To address these issues, we modified the standard stacked bar in the following way. The tags are shown as a vertical list on the left, each of which is accompanied with a (horizontally) stacked bar (Fig. 3b). This stacked bar for each tag is created by horizontally stacking individual bars for each month. To help people compare bars for a particular month, stacked bars can be interactively aligned to the left side of the bars for the selected month. We also drew a baseline starting from the time point label.

As with the multiple line graph, only the words that were included in the top 25 for the selected month are highlighted, even though all the top 75 most frequent words are displayed. Participants can select a time point by clicking on the label shown at the bottom of the visualization. The currently selected time point is marked with blue background and a thick border. When participants move the mouse over a word or a stacked bar, we highlight only the focused word and the bar. To help people count or locate the time point in a stacked bar, we alternate two colors between consecutive time points and draw a small tick mark for any time point when the word was not included in the top 25. Since the top 25 words for the selected time point are often scattered, we support vertical scrolling with mouse wheel for easy access.

4.4 Study Design and Procedure

We ran the study as a 4 (*Visualization*: SparkCloud, MultiLine, ParallelCloud, StackedBar) × 6 (*Task*) within-subjects design, with each participant performing all the tasks using all the visualizations. To avoid the learning effect, we counterbalanced the order of visualizations using Latin Square Design and used 5 sets of tag clouds (one for each of the four visualizations and one for the practice for all visualizations). We kept constant the order of datasets and tasks (from T1 to T6). Each of the 6 tasks was performed twice using the same set of tag clouds but with a different month selected by default, to alleviate learning effects over the data. All twelve trials were always presented in the same order. We measured task completion time and accuracy to estimate the efficacy of each visualization, and collected participants' subjective preferences for each visualization.

The study began with the administrator describing the dataset and explaining the goals of the study. She explained to the participants that the tag clouds represented the top 25 most frequent words that occurred in some text over the course of a year, grouped by month. Participants were asked to answer questions as quickly as possible without sacrificing accuracy. Before beginning the trials for a particular visualization, participants received instruction specific to that visualization and performed twelve practice trials (2 for each task) in order to familiarize themselves both with the task and with the visualization. While participants could spend as much time as they wanted for the practice trials, each timed trial had a 1-minute time limit. After completing all twelve timed trials for the visualization, participants filled out a satisfaction questionnaire. The same procedure was repeated with the remaining three visualizations. At the end of the experiment, participants were asked to rank the visualizations in order of preference, from most to least preferred. The experiment lasted approximately two hours.

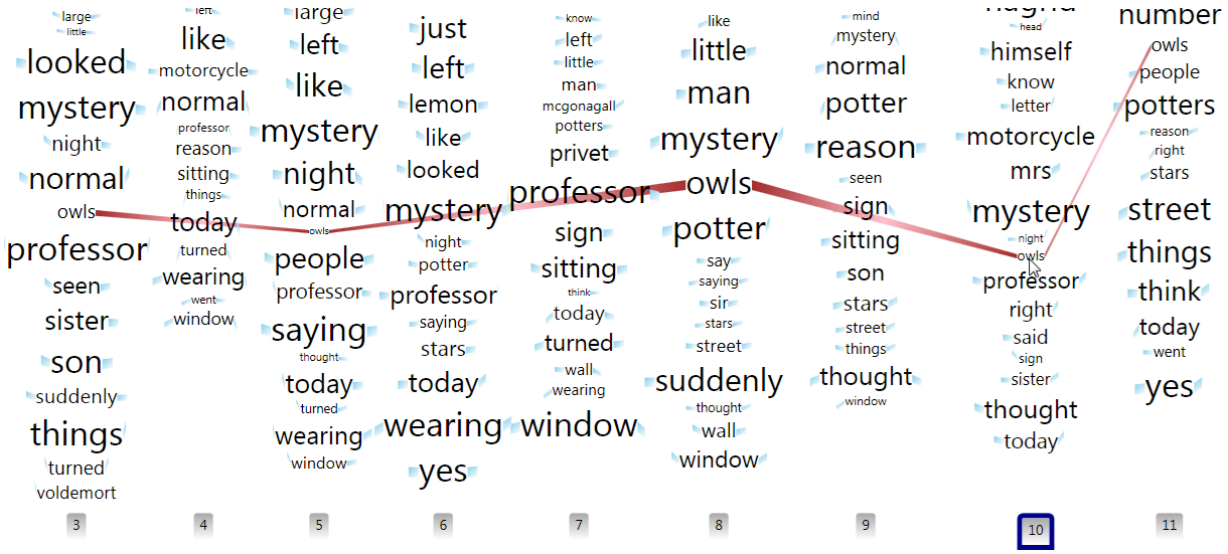

Fig. 4. ParallelCloud displays a gradient line that links the same word occurring in multiple tag clouds when people move the cursor over a word.

4.5 Equipment

We ran two participants performing tasks independently in each session of the study. Each participant worked on a 3.16 GHz Dell Precision T5400 computer with 8GB RAM and a 24" Samsung SyncMaster T240HD LCD display running at a 1920×1200 pixel resolution. For all the trials, task time and answers were logged by management software running on the computer. The display size was 1600×1200 pixels. Since questions were displayed at the top of the screen for each trial, the space for the visualization was 1600×1000 pixels. While MultiLine, ParallelCloud and StackedBar required all the available space, SparkCloud used less than a quarter of the visualization space (570×570 pixel). For the specific data retrieval tasks that required a text-based response (T1 and T2), a text box was provided for the participants to enter their answers. For the remaining tasks, four possible radio-button based choices were displayed for participants to choose an answer from.

Time on task was defined from the moment a participant pressed the trial *Start* button to the moment the participant committed to a response. So as not to include the time spent reading questions, participants were asked to click on the Start button to indicate that they were ready to begin, after reading the questions. A visualization would appear on the display under the question. For the trials that required a text-based response, the trial timer stopped when the participant began typing a response in order to ensure that individual differences in typing speed would not affect the timing result. The participant could start typing without clicking in the text box to type an answer. The response was submitted when the participant pressed the *Enter* key. To discourage participants from stopping the timer prematurely for these tasks, the visualization disappeared from view once the timer stopped. For the remaining tasks, the timer stopped when participants confirmed their choice by pressing a *Done* button, which was enabled only after a response was registered.

4.6 Results

We present the results from the study in three parts; task time, error, and subjective preferences.

4.6.1 Task Time

Because we instructed participants to perform trials as quickly as possible without sacrificing accuracy, we interpret errors in the result set to be related to characteristics of the visualizations, rather than deliberate trade-offs of accuracy for speed. Under this reasoning, we included all trials in the time analysis, even those that were answered incorrectly. Participant trials that exceeded one minute time limit (4 of 816 trials, or 0.5%) were recorded in the dataset as an incorrect trial that took 60 seconds. During analysis, all post-hoc analyses were performed using Holm's Sequential Bonferroni correction.

A 4 (*Visualization*: SparkCloud, MultiLine, ParallelCloud, StackedBar) × 6 (*Task*) repeated measures analysis of variance (RM-ANOVA) was performed on the logarithm of the mean task completion time for each participant—a standard transformation which corrects for the non-normal distribution of such data.

We found a significant main effect of *Visualization* ($F_{3,48}$=50.9, p<0.001), with SparkCloud and MultiLine each supporting significantly faster overall task times than both ParallelCloud and StackedBar (Fig. 5). The analysis also yielded a significant main effect of *Task* ($F_{5,80}$=221.6, p<0.001) and a significant interaction of *Visualization* × *Task* ($F_{15,240}$=25.2, p<0.001).

Although the main effect of *Task* is unsurprising given that we intentionally designed the tasks to explore different types and complexities of relationships within the dataset, we performed six follow-up one-way RM-ANOVAs to test performance differences among the four *Visualizations* within each task type to better understand how and where the Visualizations diverged from one another according to *Task*. Task completion times differed significantly across the four *Visualizations* within each *Task*: T1 ($F_{3,48}$=68.0), T2 ($F_{3,48}$=45.9), T3 ($F_{3,48}$=17.7), T4 ($F_{3,48}$=10.8), T5 ($F_{3,48}$=9.1), and T6 ($F_{3,48}$=6.4), all at the p<.001 level (see Fig. 6).

Post-hoc comparisons found that SparkCloud was the most consistently competitive visualization, being outperformed by a different visualization in only one task: T2, where it was significantly slower than MultiLine (12.4s v 5.4s). In three of the tasks (T1, T3, T6) SparkCloud performed at least as well or significantly faster than the other visualizations, and in T5 (comparing topics to identify the one that was most often ranked highest) SparkCloud stood out as fastest among all visualizations.

MultiLine was the second-most effective visualization, performing among the fastest visualizations for half of the tasks (T1-T3). MultiLine performed notably well for T2 (identifying the month

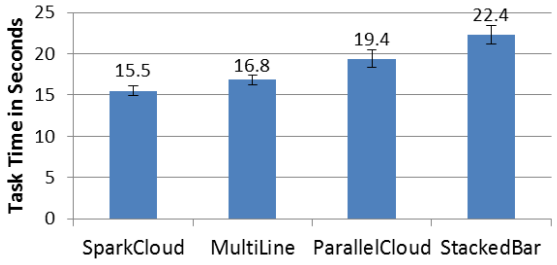

Fig. 5. Mean overall task time for each Visualization. Error bars represent standard error.

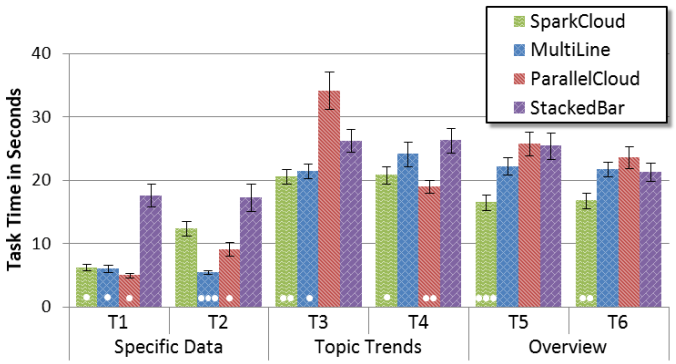

Fig. 6. Mean task completion time by visualization for each task. White dots in columns indicate the total number of other visualizations that were significantly slower within the same task. Error bars represent standard error.

in which a topic occurred most frequently) where it supported significantly faster task completion times than all other visualizations.

ParallelCloud was outperformed by another visualization in four of the six tasks (T2, T3, T5, T6); in T1, ParallelCloud performed comparably to SparkCloud and MultiLine, but in T4 (comparing the magnitude of increase/decrease of four topics in two consecutive months), ParallelCloud was significantly faster than all other visualizations except for SparkCloud.

StackedBar was the only visualization that never significantly outperformed another visualization in any of the tasks. This trend was particularly apparent for the tasks that asked for specific data responses (T1 and T2), where StackedBar performed significantly slower than most of the other visualizations.

4.6.2 Error

The overall selection error rate was relatively low at 3.8%. A non-parametric Friedman test was conducted to evaluate differences in median error rate across the four visualizations, but the test did not yield significance, $\chi^2(3,N=17)=3.7$, $p=0.3$. Given that our instructions to participants emphasized the importance of correct answers, it is unsurprising that overall error rate was too low to uncover any significant differences between the visualizations. However, considering only the raw error rates, shown in Fig. 7, we see that the error rates trend similarly to the task speed data, with SparkCloud having the least number of errors (4) across the fewest tasks (2), and StackedBar having the most errors (12) across the greatest number of tasks (6).

4.6.3 Subjective Preferences

Participants rated each of the four visualizations across eight satisfaction criteria including ease of learning, ease of use, effectiveness for use in performing the task, and appearance. were Responses were on a 1-7 Likert scale, with 1=Strongly Disagree and 7=Strongly Agree, and analyzed using Friedman tests, and post-hoc analyses were conducted with Wilcoxon tests.

Participant responses did not differ significantly across any of the visualizations for the three questions: ease of use, readability of the topic terms, and liking the interface. StackedBar was rated significantly lower than at least one of the other visualizations for four questions: learnability, efficiency of use, comparing topics

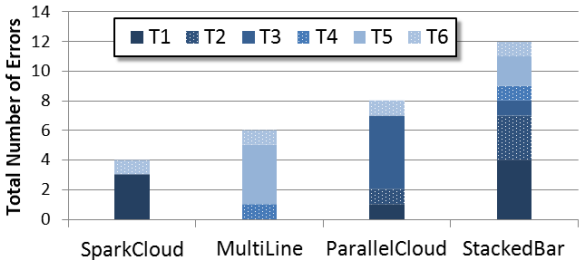

Fig. 7. Total number of errors for each visualization, by task type.

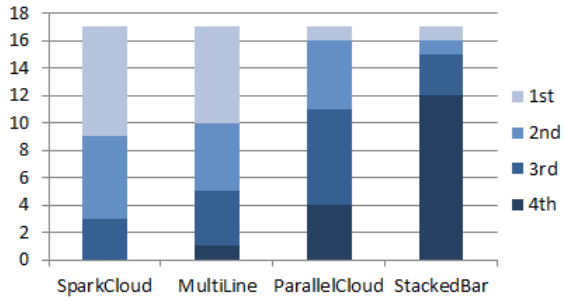

Fig. 8. The number of participants who ranked each visualization 1st through 4th in overall preference.

within a month, comparing topics between months; however SparkCloud, MultiLine, and ParallelCloud were statistically indistinguishable from one another for these same questions. All visualizations received their lowest mean and median ratings for the question regarding overall appearance ("The visualization looked cluttered."), although post-hoc Wilcoxon tests found that SparkCloud was rated significantly higher with a median of 4, than the other visualizations, which all had a median rating of 2. Note that the scale of this last question was reversed for analysis so that higher scores always indicate better ratings.

A final question that asked participants to provide an overall preference ranking (1=most preferred visualization, 4=least preferred visualization) was analysed with a Friedman test to evaluate differences in median rank across the four visualizations. The test was significant, $\chi^2(3,N=17)=21.5$, $p<0.001$. Follow-up pairwise Wilcoxon tests found that SparkCloud was ranked significantly higher than both ParallelCloud ($p=0.003$) and StackedBar ($p=0.001$), and that MultiLine was ranked significantly higher than StackedBar ($p=0.003$). Otherwise, SparkCloud and MultiLine were statistically indistinguishable from one another, as were ParallelCloud and StackedBar (Fig. 8).

5 Discussion and Future Work

The encouraging results of the controlled study support our hypothesis that the SparkCloud design, which integrates sparklines into a tag cloud to convey trends over time, is effective. SparkCloud is not only most consistently competitive overall but also preferred by participants compared to StackedBar and ParallelCloud.

We believe this is because SparkCloud successfully inherits the benefits from both sparklines and tag clouds. As simplified and compact line charts, sparklines can convey trends as effectively as multiple line graphs do. However, to achieve this simplicity and compactness they lack details by not integrating labels for time axis (e.g., months), causing participants to estimate these values. This explains why SparkCloud did not perform as quickly as MultiLine or ParallelCloud for T2 (identifying the month in which a topic occurred most frequently). While participants could easily notice the peak from the sparkline, they still had to make a visual estimate of the month during which the peak occurred and then manually verify their estimation by selecting the month. Furthermore, if this guess was not correct, participants had to iteratively choose months until the correct one was found. To provide more details on the sparkline without sacrificing its simplicity and compactness, we envision providing a tooltip displaying a regular line graph with axis labels.

We argue that it is to SparkCloud's advantage that it retains the form of a tag cloud. Even though SparkCloud used less space (smaller than one quarter of the space used by other visualizations), participants rated SparkCloud as looking significantly less cluttered than others. Furthermore, in addition to being aesthetically pleasing, the compact and space efficient layout makes it possible to replace traditional tag clouds with SparkCloud. However, it is important to note that SparkCloud also inherit the weaknesses of tag clouds. For example, longer words receive more user attention because they occupy more space. This problem is further amplified when they are assigned larger fonts.

To our surprise, ParallelCloud performed better than SparkCloud for T4 (comparing the magnitude of increase/decrease of four topics in two consecutive months). We expected ParallelCloud to do worse than other visualizations because we thought that the slope of lines (MultiLine) or length of the bars (StackedBar) would be easier to compare than word sizes. We were also surprised to observe that StackedBar never significantly outperformed other visualizations in any of the tasks. We hypothesized that StackedBar would perform well for T4 because it allows people to compare two bars side by side. We suspect this is due to the fact that StackedBar required participants to scroll vertically to check all four potential answers. It seems that the scrolling had a more significant effect than we expected. Given that the main focus of our study was to understand how people perceive different visual representations, we provided minimal interactivity. Another reason for StackedBar's poor performance could be the limited interaction supported by the current implementation of StackedBar.

In fact, the issue of scrolling reveals an important scalability issue in terms of both the number of time points (i.e., tag clouds) and the length of words that can be supported by each visualization, especially for ParallelCloud. As mentioned above, to preserve the utility of the mouse-over interaction in ParallelCloud, we prepared the dataset so that horizontal scrolling would not be required. Note that we could not supress vertical scrolling in StackedBar because reducing the number of tags for each time point would have made our experiment unrealistic. While ParallelCloud performed better than we expected, we would argue that it would suffer significantly when scrolling is necessary. Even though SparkCloud and MultiLine are not completely free from the scalability issue, they would scale better than other visualizations.

Future work we envision is to study the importance of stability and to assess the compromise between stability and visual clutter. In this study, for all visualizations except for ParallelCloud, we displayed all the top 75 most frequent words and highlighted only top 25 words for the selected time point. We made this design choice primarily to support a fair comparison with ParallelCloud in terms of visual distraction since ParallelCloud enforces that all words are shown at all times (and some were duplicated multiple times); certainly the fewer words we show, the less visual clutter a visualization will endure. Another reason for this choice was to ensure that all visualizations were as visually stable as possible. By displaying all the words the visualizations did not change dramatically from one time point to the other. This made it easier to track the trend of a particular tag over time. However, StackedBar suffered significantly from this decision because additional bars introduced the need for scrolling. Since there are trade-offs between stability and visual clutter, it would be interesting to investigate these trade-offs more formally in future work.

6 Conclusion

In this paper, we have described SparkClouds, a novel visualization that incorporates sparklines into a tag cloud to represent trends across a series of tag clouds; SparkClouds inherit advantages from both sparklines and tag clouds. First, with sparklines, SparkClouds effectively provide people with an overview of trends using very little additional space. Second, because it is still in the form of tag clouds, SparkClouds offer a compact and aesthetically pleasing layout and can be used in place of traditional tag clouds. We have also described the design of SparkClouds along with usage scenarios.

We then presented results from a lab study, in which we explored how efficiently people could learn and use SparkClouds as compared to three alternative visualizations (two traditional trend visualizations and existing Parallel Tag Clouds) for three different types of tasks: specific data tasks; topic trends tasks; and overview tasks. Results suggest that participants are more efficient with SparkClouds and also that they liked SparkClouds more than two other visualizations. Finally, we have discussed potential future work that could shed more light on the benefits of SparkClouds.

Acknowledgments

We would like to thank George Robertson for his helpful comments.

References

[1] ABC News – Tag Cloud – News topics organized to show what's in today's headlines, http://www.abc.net.au/news/tag/cloud.htm

[2] S. Bateman, C. Gutwin, and M. Nacenta, "Seeing things in the clouds: the effect of visual features on tag cloud selections," *Proc. HYPERTEXT '08*, pp. 193-202, 2008.

[3] P. Bausch and J. Bumgardner, *Flickr Hacks*. O'Reilly Press, pp. 82-86, 2006.

[4] C. Collins, F.B. Viégas, and M. Wattenberg, "Parallel Tag Clouds to Explore Faceted Text Corpora," *Proc. VAST '09*, pp. 91-98, 2009.

[5] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins, "Visualizing Tags over Time," *Proc. WWW '06*, pp. 193-202, 2006.

[6] Explore tags on Delicious, http://delicious.com/tag

[7] P. Gambette and J. Véronis, "Visualising a Text with a Tree Cloud," *Proc. 11th IFCS Biennial Conference*, 2009.

[8] M.J. Halvey and M.T. Keane, "An assessment of tag presentation techniques," *Proc. WWW '07*, pp. 1313-1314, 2007.

[9] S. Havre, H. Elizabeth, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *IEEE TVCG*, vol. 8, no. 1, pp. 9-20, 2002.

[10] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating Summaries and Visualization for Large Collections of Geo-Referenced Photographs," *Proc. MIR '06*, pp. 89-98, 2006.

[11] O. Kaser and D. Lemire, "Tag-Cloud Drawing: Algorithms for Cloud Visualization," *Proc. WWW '07 Workshop on Tagging and Metadata for Social Information Organization*, 2007.

[12] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration," *Proc. Interact '09*, LNCS 5726, pp. 392-404, 2009.

[13] Many Eyes: Tag Cloud, http://manyeyes.alphaworks.ibm.com/manyeyes/page/Tag_Cloud.html

[14] S. Milgram and D. Jodelet, "Psychological maps of Paris," W.I.H. Proshansky and L. Rivlin, eds., Environmental psychology, New York: Holt, Rinehart, and Winston, pp. 104-124, 1976.

[15] Popular Tags on Flickr, http://www.flickr.com/photos/tags

[16] A.W. Rivadeneira, D.M. Gruen, M.J. Muller, and D.R. Millen, "Getting our head in the clouds: toward evaluation studies of tagclouds," *Proc. CHI '07*, pp. 995-998, 2007.

[17] T. Russell, "Tag decay: A view into aging folksonomies," *Proc. ASIS&T '07*, pp. 1-5, 2007.

[18] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer, "On the Beauty and Usability of Tag Clouds," *Proc. IV '08*, pp. 17-25, 2008.

[19] J. Sinclair and M. Cardew-Hall, "The Folksonomy Tag Cloud: When is it Useful?" *J. of Information Science*, vol. 34, no. 1, pp. 15–29, 2008.

[20] TagCrowd, http://www.tagcrowd.com

[21] Tagline Generator, http://chir.ag/projects/tagline

[22] Tag Cloud Generator, http://www.tagcloudgenerator.com

[23] E.R. Tufte, Beautiful Evidence, Graphics Press.

[24] US Presidential Speeches Tag Cloud, http://chir.ag/projects/preztags

[25] F.B. Viégas and M. Wattenberg, "Tag Clouds and the Case for Vernacular Visualization," *Interactions,* vol. 15, no. 4, pp. 49-52, July/August 2008.

[26] F.B. Viégas, M. Wattenberg, and J. Feinberg, "Participatory Visualization with Wordle," *IEEE TVCG (InfoVis '09),* vol. 15, no. 6, pp. 1137-1144, 2009.

[27] M. Wattenberg, "Baby names, visualization, and social data analysis," *Proc. InfoVis '05*, pp, 1-7, 2005.

[28] W. Willett, J. Heer, M. Agrawala, "Scented Widgets: Improving Navigation Cues with Embedded Visualizations," *IEEE TVCG (InfoVis '07)*, vol. 13, no. 6, pp. 1129-1136, 2007.

[29] Wordle – Beautiful Word Clouds, http://www.wordle.net