Public Data and Visualizations: How are Many Eyes and Tableau Public Used for Collaborative Analytics?

Kristi Morton[1], Magdalena Balazinska[1], Dan Grossman[1],
Robert Kosara[2], and Jock Mackinlay[2]

[1]University of Washington
{kmorton,magda,djg}@cs.washington.edu

[2]Tableau Software
{rkosara,jmackinlay}@tableausoftware.com

ABSTRACT

Recently, online visual analytics systems have emerged as popular tools for data analysis and sharing. The database community has an important role to play in shaping the design and implementation of these new types of systems. Little, however, is known about how these systems are used today. In this paper, we address this shortcoming by presenting an analysis of usage patterns of Many Eyes and Tableau Public, two popular Web-based, collaborative visual analytics systems.

1. INTRODUCTION

As data has become more publicly available on the Web, for example, through local and national government initiatives such as the Open Data movement [6], a broader audience has emerged as data consumers and knowledge-seekers (referred to as *data enthusiasts*) [11]. These people are not statisticians or programmers, yet want to use data to answer a question or solve a problem. A typical example is a news reporter who wants to use data and visualizations to illustrate a story and make it available online (*e.g.,* on her blog).

Over the past few years, increasingly many online data visualization systems have appeared to meet the demands of such users for data analysis and sharing [2, 1, 4, 18, 5, 9, 10]. The core functionality of these systems is threefold: (1) They enable users to *visually* explore their data: users have access to a graphical user interface through which they can easily create various charts and graphs. Importantly, through these interfaces, users are basically executing queries to filter, group by, and aggregate datasets. (2) These systems also facilitate the integration and study of *multiple* datasets. (3) Finally, they support *collaboration* between users through sharing visualizations and data *online* for both viewing and editing by others [12, 20]. These services are thus a new type of easy-to-use data management and analytics systems.

While different systems have different architectures, several are based on the integration of a visualization front-end with a database management system back-end [19, 9]. For example, Tableau supports analysis across a variety of structured, heterogeneous data sources (*e.g.,* delimited text files, cubes, data marts, and databases), and issues live queries to these sources to obtain the necessary data to render each visualization. With live query support, interactive analysis is possible: visualizations can be altered on-the-fly and multiple data sources can be joined together.

Unlike Tableau Public, Many Eyes visualizations are created and published through a Web browser. Either structured (*i.e.,* tables) or unstructured (*i.e.,* bag of words) data is ingested through a browser using cut-and-paste operations from a text file up to 5 MB in size. Once a visualization is chosen for a given data set it cannot be arbitrarily altered nor combined with other data. Moreover, while both systems share many of the same visualization types (*i.e.,* bar, line, text, pie, area, scatter, and maps), Many Eyes includes a number of unique text analysis techniques that are not available in other systems. Such visualizations include word clouds, phrase nets, and word trees.

Despite their growing popularity, little is known about how these systems are being used. Even basic statistics such as the number of users are often not published (*e.g.,* Fusion Tables [9]), let alone any details of user activity. The most prominent system, Many Eyes, started in early 2007, and initial studies [8] indicated a significant uptake, as well as collaboration between users; but there have been no follow-up studies on usage, nor have there been comparable studies of other web-based or web-centric visual data analysis systems. Shortly before Many Eyes, in December 2006, Swivel.com was launched. Swivel was much simpler and less academically ambitious than Many Eyes, but run as a start-up rather than an experiment. It shut down in summer 2010, casting doubt on whether there was a market for web-based visual data analysis systems. At the same time, there is clearly broad interest in data integration, analysis, and visualization. *The New York Times*, *The Washington Post*, *The Guardian*, and other news media are not only increasingly using visual data analysis as part

System	Start Date	# Visualizations	# Workbooks	# Datasets	Users
Many Eyes	January 1, 2007	149,395 (3.2/user)	n/a	358,880 (7.8/user)	46,048
Tableau Public	February 10, 2010	269,609 (11/user)	73,404 (3/user)	107,596 (4.4/user)	24,563

Table 1: Summary of the collected data from Many Eyes and Tableau Public, from each system's inception until December 31, 2012.

of news stories, but also experimenting with more sophisticated types of visualizations.

As our society continues to become "data-enabled", it is important that we continue to improve data management and analysis tools. If we are to build better online data visualization and sharing systems, the first step is to understand how they are being used today. The key contribution of this paper is to shed light on this exact question: *How are online visual data analysis and sharing systems being used?*

We take a first step toward answering this question through a longitudinal measurement study of two popular online data visualization and analysis systems: Tableau Public [4] and Many Eyes [18, 3]. Both systems allow users to create visualizations online, and both are free to use. Tableau Public requires the download of a Windows-only client, while Many Eyes is used entirely in the browser. Both systems provide a variety of different visualization techniques, which not only generate static images, but which the viewer can interact with in the browser. The data used in visualizations can be downloaded in both systems.

We tackle the question of how both of these systems are being used from the perspective of the database community. Through our study, we thus focus on the following core set of questions: (1) How popular are these systems? How many users do they attract and how active are these users? (2) How heavily do users leverage the collaborative features of these tools? (3) What do users actually do with the data? How do they analyze it? How much data (in terms of relation cardinality and degree) do users choose to visualize at any given time? And finally (4) Do users integrate multiple data sources in their visualizations? And how do they perform these integrations? To the best of our knowledge, this is the first formal study of these types of systems.

2. METHOD

Our study is based on traces of Many Eyes and Tableau Public as summarized in Table 1. The traces span six and three years respectively and include detailed information about the data and visualizations that are published to each system. The Tableau Public trace also includes detailed traffic and impression data for each visualization. For Tableau Public, each workbook specifies the data sources analyzed (including all schema metadata), the types of visualizations produced, and all of the specific VizQL[1] definitions [17] that produce each visualization. For Many Eyes, metadata was collected from the visualization types used and inferred from data sets uploaded.

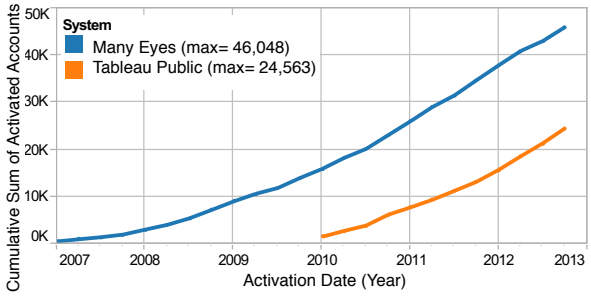

Figure 1: Cumulative growth of Many Eyes and Tableau Public activated user accounts.

3. RESULTS

We present the key results of our analysis, organized around our four core questions related to the user-base, collaborations, single-dataset analytics, and analytics of integrated datasets.

3.1 User-base

The first question that we ask is whether web-based visual analytics systems are at all popular. To answer this question, we measure the size of the user-base for each system. Figure 1 shows how the systems are growing over time in terms of the number of opened accounts. As the figure shows, since its inception in January 2007, Many Eyes, has grown to over 46,000 authors who have published over 358,000 data sources and more than 149,000 visualizations. For Tableau Public, its user-base includes 24,500 authors who have contributed over 73,000 workbooks, 107,500 datasets, and 269,000 visualizations (Table 1). We define authors to be users who have published at least one data set or visualization. **These systems thus have moderate numbers of users today, but their popularity is continuing to grow significantly each year.**

The natural next question is how active are these authors over time. Interestingly, as Table 2 shows, **half the users (or almost half) are *one-time* users who publish only one dataset or visualization. The remaining users are mostly *light* users who publish two to four**

[1]Visual Query Language (VizQL) a formal declarative language for describing visualizations

	Number of Data Sources Published				
	1	≤2	≤3	≤4	≤5
Many Eyes	44%	65%	76%	83%	86%
Tabeau Public	45%	63%	73%	79%	83%
	Number of Visualizations Published				
	1	≤2	≤3	≤4	≤5
Many Eyes	52%	72%	82%	87%	90%
Tabeau Public	53%	71%	80%	85%	88%

Table 2: Cumulative fraction of users who publish up to a given number of data sources or visualizations (e.g., 80% of Tableau users publish 3 visualizations or less).

visualizations. Only 10% to 17% are *prolific* users who publish five or more data sets or visualizations. Despite the significant age difference between the two systems, it is still interesting to note the most prolific author who is not directly affiliated with either system: on Many Eyes such a user contributed 1,617 data sets and visualizations; and on Tableau Public 2,927 data sets and visualizations were published by one user.

Since the majority of users are one-time or light users, does it mean that most of the activity in the systems is due to novices? Figure 2 confirms this hypothesis. In the figure, we group users into cohorts based on the quarter in which they publish their first visualization (workbook on Tableau Public). For each quarter, the figure shows the fraction of active accounts that come from each *returning* cohort. For example, in the third quarter, 8% of active accounts in Many Eyes belong to the second cohort and 11% of active accounts belong to the first cohort. The remaining 81% active accounts (not shown) belong to users who joined the system that quarter. On average only 17% of active accounts in Many Eyes belong to returning users from any cohort. The average is 31% for Tableau Public. **Hence, web-based data analysis systems today need to provide good support for novices.**

One hypothesis for high user churn is bad system performance.According to a study published on Web users' tolerable waiting time [15], two seconds is considered an acceptable waiting time for loading Web pages. We measure, however, that 84% of all visualizations on Tableau Public take less than two seconds to load (includes both query and rendering time) and 98% are under ten seconds (the accepted limit for keeping a user's attention focused on a given task [14]). Although attitudes and expectations change over time, the basic capability of human attention has not changed over the decades [7, 14]. Thus, our results indicate that the majority of load times should not negatively impact Tableau Public's users. Performance alone thus cannot explain the high degree of user churn. It could, however, simply be that the systems do not offer the visualization capabilities users want (*e.g.,* limited to no support for unstructured data). Users explore the systems but walk away when they find them unsuitable for their needs.

If we frame these retention results in the context of other free, web-based services such as Twitter, we see that low retention after initial use is common. According to a 2009 Nielsen report [13] only 40% of Twitter users returned to use the site after the first month.

Bottom Line: Web-based visual analytics systems continue to attract thousands of users, but most of them use these systems lightly right after registering and then stop. Only a small fraction of users grows into power users. The implication for the database community is that Web-based visual analytics systems must be geared toward supporting novice users and there is significant room for improvement in retaining these users.

3.2 User Interaction and Collaboration

Since both systems are designed for sharing visualizations and collaboratively analyzing data, we explore the frequency of viewership, collaboration, and sharing in this section.

3.2.1 Viewership

Based on a distinct count of user cookies, we found that there are approximately 52 million unique visitors to Tableau Public. **Visitors are thus several orders of magnitude more numerous than the authors** (only ≈ 24,500 authors). Additionally, we found that the top 50% of all Tableau Public traffic is attributed to 244 distinct workbooks (or 0.3% of all workbooks). For Many Eyes, however, we did not have access to the equivalent traffic and viewership information.

3.2.2 Collaboration

On Tableau Public and Many Eyes, users can download, edit, and republish any visualization and supporting data set. To get a sense of the degree of such collaborative activities among authors, we explore how often authors take existing content and evolve it for their own analytical needs (*e.g.,* by changing the visualization content to explore some other dimension or measure) and then republish it with their insights. In our approach, we trace the provenance of Tableau Public workbooks that were created by one author and edited and republished by a different author (called a *derivation*). Figure 3(top) shows that a workbook is about five times more likely to be derived if it is not the author's first publication. Nevertheless, the probability of derivation remains small at only 6%. Very few workbooks are derived more than once. Only 28 workbooks were derived > 4 times.

In Figure 3(bottom) we observe two types of collaborative behaviors. Some workbooks are derived multiple times by alternating between the same two authors as in a *Direct Collaboration* while others are derived by a dif-

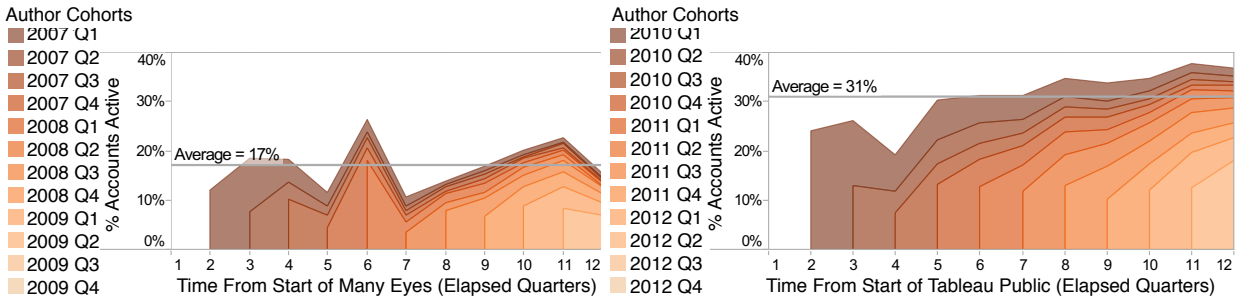

Figure 2: Many Eyes (left) and Tableau Public (right) author cohorts for the first 12 quarters (3 years). Authors are grouped into cohorts based on the quarter in which they published their first visualization or workbook. The fraction of authors that returned to the site to publish a dataset or visualization is shown as the *percent of accounts still active* for each quarter.

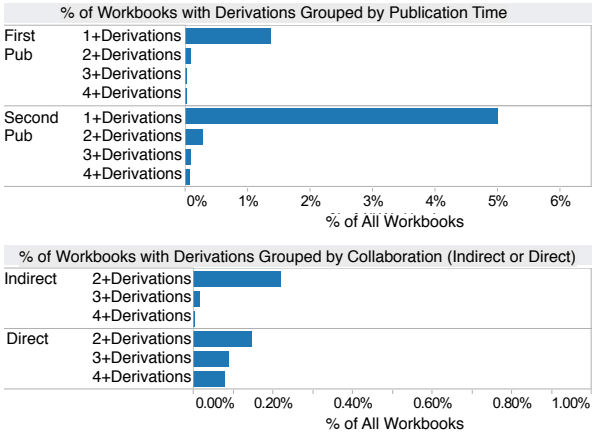

Figure 3: Derived Tableau Public workbooks partitioned by publication time (top) and nature of collaboration (bottom).

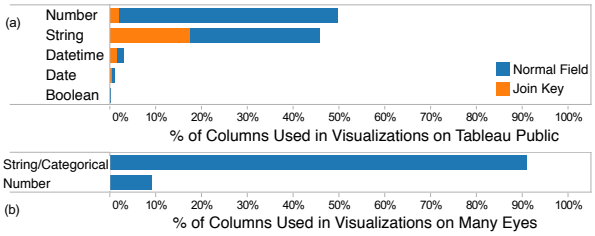

Figure 4: Data types in visualizations on Tableau Public (a) and Many Eyes (b). Tableau Public is split between numbers and strings, and word-oriented Many Eyes is heavily skewed toward strings.

ferent author each time as in an *Indirect Collaboration*.

No equivalent derivation information is available for Many Eyes. However, in order to get a sense of the degree of influence one author's contributions have on other authors, we measure how often authors reuse data uploaded and shared by others for their visual analysis in Many Eyes. We find that only 6% of datasets are used by multiple authors, which is consistent with Tableau's workbook derivation statistics, and that 20% of datasets are used in multiple visualizations. We cannot compute this statistic for Tableau Public because published workbooks make a copy of the data being visualized.

3.2.3 *New Content Published: Data Types*

We next consider what are the predominant data types that are being visualized in both systems. First, in Figure 4(a), we see that `Number` (51%) and `String` (44%) are the most common data types in visualizations on Tableau Public. It is interesting that their use is fairly balanced, while intuition would indicate that numbers might be more common due to the quantitative nature of business analytics. The `Number` data type includes both integers and reals. Finally, we see fewer specialized types such as `Datetime` and `Date`, which indicates that visualizations of time-based data are less prevalent.

Many Eyes, however, has a skewed distribution of `String/Categorical` types. In Figure 4(b), we see that 91% of columns on Many Eyes are of this type. This finding is consistent with Many Eyes's greater emphasis on text-based visualizations (*i.e.,* word clouds, phrase nets, and word trees) that are not available elsewhere.

Bottom Line: Online visual analytics systems are **read-heavy today: Orders of magnitude more people are viewers compared to authors**. Additionally, as is typically the case for database access patterns, viewership is skewed toward a **small fraction of *hot* visualizations**. Interestingly, while publishing visualizations is common, **collaborations among users remain infrequent**. Incentivizing and supporting collaborations thus remain critical challenges for these systems.

3.3 Single-Dataset Analytics

Today's online visual analytics systems are designed for small data. Most of these systems put a bound on the size of datasets that can be processed. On Many Eyes, data sizes are limited to 5MB, while on Tableau Public, each user gets a 50MB account and a visualization can operate on at most 100,000 rows. Interestingly, we find

System	Number of Rows in Visualizations				
	≤100	≤1K	≤10K	≤50K	≤100K
Many Eyes	63%	90%	98%	99%	100%
Tabeau Public	28%	53%	84%	95%	100%

Table 3: Cardinality of visualized relations.

System	Number of Columns in Data Source				
	≤2	≤10	≤20	≤100	≤300
Many Eyes	49%	84%	93%	99%	100%
Tabeau Public	2%	28%	52%	90%	99%

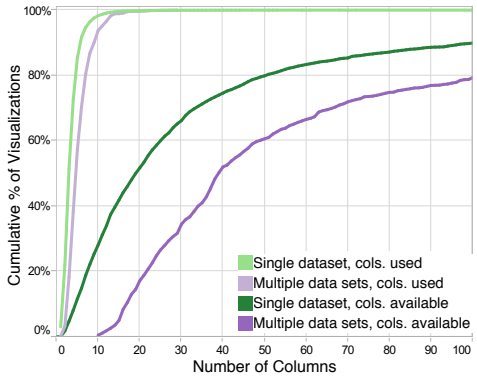

Figure 5: Degree of input relations (top). CDF of the number of columns in visualizations with one or multiple (joined) data sets in Tableau Public (bottom).

that 90% of user accounts in Tableau Public use less than half of their 50MB quotas. Hence most users do not push the pre-set limit. The focus on small-size datasets affects the size of the visualized relations as shown in Table 3: **the majority of visualized relations stays well below the pre-set cap of 100K rows.**

Tableau Public also offers a paid (*a.k.a.*, *Premium*) tier, which allows accounts to go beyond the 100,000 rows limit. These accounts (along with some accounts on Many Eyes) visualize more than an order of magnitude more data, which seems to **imply the need for the online visualization of bigger data too**.

Interestingly, datasets in both systems contain on average a large number of attributes, especially in Tableau Public, where the median dataset has 20 attributes and the top 10% have more than 100 attributes. Only a small fraction of these attributes, however, is visualized simultaneously as shown in Figure 5: 52% of visualizations with a single data source use at most 3 columns and 90% use at most 6. A similar trend appears for visualizations over integrated data sources as we discuss further in Section 3.4. The figure shows results for Tableau Public. No equivalent information was available for Many Eyes.

Table 4 shows the breakdown of the most common visualization types used for a given number of columns. The values denoted with a '⋆' in Table 4 show that a second visualization type was within 5% from the top choice for that given number of columns. For single data sources, we see that the text table is the most common

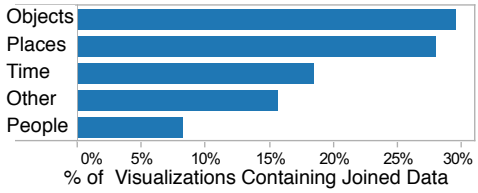

Figure 6: Common semantic entities of join keys in visualizations with multiple (joined) data sources.

type when there is only one data column present in the visualization. As the number of columns increases, we see a shift in visualization techniques used: bar views become the dominant technique for 2–4 columns and maps are the most popular for 5–8 columns.This behavior is not too surprising since map views always include two virtual columns that represent the latitude and longitude coordinates.

Bottom Line: Most visualizations have modest data sizes, well below the limits set in existing systems although some users with special privileges visualize datasets with more than one million rows. Visualizations also focus on a small number of attributes at any one time, even though many more attributes are available. Finally, as the number of columns used increases, so does the complexity of the visualization type (*e.g.,* maps require more columns than other types like bar views.)

3.4 Integrating Multiple Data Sets

In this section, we study the trends in data and visualization on Tableau Public in the context of data integration from multiple data sources. We omit Many Eyes from this section because the platform currently does not support data integration.

3.4.1 Semantic Entities for Data Integration

On Tableau Public, there are 5,532 visualizations that were created by joining multiple data sets. Of these visualizations, we ask *how do authors combine data sets for their analysis?* To answer this question we manually categorized all of the join keys for the 5,532 visualizations (2%) that have integrated data to get a sense of the most popular semantic entities. This process entailed inspecting the column name, data type, and data values of each join key. In the case where the column name was in a foreign language, we used Google Translate on the name and (in some cases) values of that column. If we were still unsure, we opened the workbook to inspect the visualization that was associated with that join key. Figure 6 summarizes the semantic entities of the join keys in five different categories: people, places, time, objects, and other. The people category contains any information pertaining to people, including names and demographics. The places category is restricted to

Number of Data Sets	Number of Columns in Visualization 1	2	3	4	5	6	7	8	9	10	11	12	13	14
One	Text (68%)	Bar (53%)	Bar (47%)	Bar (32%)	Map* (27%) Bar	Map (32%)	Map* (27%) Bar	Map* (26%) Bar	Bar* (25%) Map	Line (32%)	Map* (24%) Bar	Circle (34%)	Bar (30%)	Bar (20%)
Multiple	Text (75%)	Bar (48%)	Bar (50%)	Bar (41%)	Map* (22%) Bar	Map (35%)	Map (40%)	Text (46%)	Map (49%)	Map (56%)	Scatter* (38%) Map	Scatter* (36%) Bar	Circle (64%)	Map (42%)

Table 4: Most common visualization types for different numbers of columns in the visualization.

geolocations and other identifying characteristics such as zip codes, regions, states, countries, continents, etc. As expected, the time category refers to dates and date times and objects refer to any physical entity that is not a person, place, or time. Objects consist mainly of opaque identifiers like alphanumeric product codes as well as more well-known, descriptive entities such as "university", "department", or "team". As the figure shows, users integrate data most often by adding attributes to the same object identifier (30%), or by bringing together items located in the same places (28%) or occurring at the same time (18%).

3.4.2 Data Columns per Visualization

Figure 5 shows the number of available and visualized columns for single and integrated datasets. We see that visualizations on top of integrated datasets are significantly richer as they display a larger number of columns than visualizations over a single data source. For example, 43% of visualizations over integrated data use 5 or more columns, compared to only 15% of visualizations over a single dataset. Furthermore, we see a familiar trend as with single data sources: there is a sizable gulf between the number of columns used and the number of columns available in the integrated data sources.

Bottom Line: Data integration often occurs by combining multiple attributes about the same uniquely identified entities from different data sources. This type of data integration is more common than simply placing multiple entities at the same location or at the same point in time, although the latter two dominate when considered together. This finding is especially interesting for data integration tools. For example, a recent tool provides recommendations of potentially useful data to integrate with a given database [16]. This tool does not consider joining on place or time. It only considers extending semantic entities with additional attributes, which covers less than half of all integration scenarios. **Additionally, integrated visualizations tend to be more complex (*i.e.,* use more columns and have more columns available) than single-source ones. Interestingly, the distribution of the most common visualization types for a given number of columns is similar for visualizations over one or more datasets.**

4. CONCLUSIONS

We studied four primary dimensions of two popular online visual analytics systems: (1) what types of users are leveraging these systems and how heavily, (2) how are users collaborating and interacting with the published content, (3) how do users analyze a single-dataset and (4) how they integrate data sources. We find that web-based visual analytics systems have much room for growth: they attract large numbers of users but most users do not push the limit of what these tools can do.

Acknowledgments. This work is partially supported by NSF CDI grant IIA-1028195.

5. REFERENCES

[1] Gapminder. `http://www.gapminder.org/`, 2012.
[2] iCharts. `http://www.icharts.net/`, 2012.
[3] Many Eyes. `http://many-eyes.com/`, 2012.
[4] Tableau Public. `http://www.tableaupublic.com/`, 2012.
[5] ViewShare: Interfaces to our Heritage. `http://viewshare.org/`, 2012.
[6] T. Berners-Lee. The year open data went worldwide. TED Talk, `http://on.ted.com/eU5`, 2009.
[7] S. Card et al. The information visualizer, an information workspace. In *CHI*, 1991.
[8] C. Danis. et al. Your place or mine?: visualization as a community component. In *CHI*, 2008.
[9] H. Gonzalez et al. Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud. In *SOCC*, 2010.
[10] H. Gonzalez et al. Google fusion tables: Web-centered data management and collaboration. In *SIGMOD*, 2010.
[11] P. Hanrahan. Analytic database technologies for a new kind of user: the data enthusiast. In *SIGMOD*, 2012.
[12] J. Heer et al. Design considerations for collaborative visual analytics. In *IEEE VAST*, 2007.
[13] D. Martin. Twitter Quitters Post Roadblock to Long-Term Growth. `http://blog.nielsen.com/nielsenwire/online_mobile/twitter-quitters-post-roadblock-to-long-term-growth/`, 2009.
[14] R. Miller. Response Time in Man-Computer Conversational Transactions. In *Proc. of the AFIPS Fall Join Computer Conference*, volume 33, 1968.
[15] F. Nah. A Study on Tolerable Waiting Time: How Long Are Web Users Willing to Wait? In *Behaviour and Information Technology*, volume 23, 2004.
[16] A. D. Sarma et al. Finding Related Tables. In *SIGMOD*, 2012.
[17] C. Stolte et al. Polaris: a system for query, analysis, and visualization of multidimensional databases. *IEEE TVCG*, 2002.
[18] F. B. Viegas et al. Many eyes: A site for visualization at internet scale. *IEEE TVCG*, 13(6), 2007.
[19] K. Wesley et al. An Analytic Data Engine for Visualization in Tableau. In *SIGMOD*, 2011.
[20] W. Willett et al. Commentspace: Structured support for collaborative visual analytics. In *CHI*, 2011.