Preserving Anonymity in Small Datasets
Brent Reeves
Abilene Christian University
ACU Box 29319
+1.325.674.2385
brent.reeves@acu.edu

Kumiyo Nakakoji
Research Center for Advanced Science and Technology
University of Tokyo
kumiyo@kid.rcast.u-tokyo.ac.jp

ABSTRACT
Small datasets of confidential information make it difficult to provide rich visualizations because they are susceptible to simple deductive strategies for discovering the information source. There are similarities in small team peer evaluation graphics and SQL queries against aggregate data. Building better visualizations competes with the goal of collecting the data in the first place. We use a project team peer evaluation system to illustrate these challenges.

Categories and Subject Descriptors
K.4.1 [Public Policy Issues]: Privacy.

General Terms
Human Factors.

Keywords
Visualizing anonymous data.

1. INTRODUCTION
In spite of the advantages of team learning, evaluation of individual contributions is still a major challenge [3]. Peer evaluations have been studied in many business situations. Paswan & Gollakota [4] list studies of the armed forces, marketing and sales, managerial performance, and future job advancement. The primary focus in academic settings has been on how to determine individual grades based on one’s contribution to a team [2]. The primary motivation behind most peer-evaluation systems is to prevent “free-riding” [1]. To this end, both zero-sum and non-zero-sum systems have been devised [5].

2. THE SYSTEM
Zing’Em is a web-based system for project teams to provide peer evaluation feedback. The student currently has three interactions with the system. First, the student receives an email invitation to participate in the peer evaluation. The email contains a link to a web-page containing a form as shown in Figure 1.

Previous evaluation instruments were viewed as very tedious, so one of the design criteria was ease of use in this data-entry part of the system. For each criteria, there is a list of team members and a Likert scale showing the student’s agreement with the criteria. Professors report that it takes their students “a minute, or two” to complete the form.

The third interaction with the system comes in the form of an email sent to each student showing the results of their evaluation. Figure 2 is an example. The data each student sees is the minimum, maximum, and average for each of the criteria. Figure 2 shows an example of a team with only 3 members. If the recipient has high trust with one of the other two members, she can deduce the third person’s evaluation. By factoring out the “known” numbers, she can know exactly what the third person voted on each criteria.

Now compare this to the problems in supporting SQL queries against confidential data. One approach is to make the data available only in the aggregate. But the following will show a simple method to reveal confidential data, even when it is available only in aggregated form. Assume one knows a 40 year old professor in MIS who is single and earned their degree at CU. The following four queries could reveal his salary [6]:

Query 1: Select count(*) from faculty where dept='MIS' and age >= 40.
Answer: 10 rows found

Query 2: Select count(*) from faculty where dept='MIS' and age>=40 and degree-from = 'CU'
Answer: 2 rows found
Query 3: Select count(*) from faculty where dept='MIS' and age>=40 and degree-from = 'CU' and marital-status = 'S'
Answer: 1 row found
Query 4: Select average(salary) from faculty where…
Answer: 112,000

Figure 2. Evaluation results received by each student

The parallel we see between this approach to SQL queries and our team eval data is that depending on team size and makeup of the team, the data is not as confidential as it needs to be.

Figure 3 shows our first attempt to display team data visually to the professor. It shows a four person team, each team member has three incoming arrows representing the evaluations the others gave. Green is positive, blue is negative.

At first, we thought that removing the names would be enough to ensure confidentiality of individual evaluations. But taking Figure 3 as example, if I am Moe, I know that I gave 2 positive votes and one negative. Thus I know that the person to whom I gave a negative vote, in turn gave me a positive vote. I also know that one of the other two people gave me the negative vote. So with teams of 4, it turned out to be much too easy to deduce the origin of data that was supposed to be kept confidential.

One approach to the problem in Figure 2 would be to remove the min and max, only showing the average. But this does not always suffice to protect the other team member’s identity.

Another approach is to average all the criteria into one number. Indeed, professors often do this to arrive at a grade. But an examination of the data shows that students appear to try to distinguish one criteria from another. Thus lumping them together results in loss of information, defeating the whole purpose of carefully crafting the wording of each criteria.

3. DISCUSSION

A system which was designed from the beginning to deal with confidential data turned out to have exposures in too many real circumstances. Often, more (visual) data is better, but in our setting, the very existence of the system is due to the promise of confidentiality. Successively adding graphics attributes parallels increasingly specific SQL queries against aggregated data.

Our current visualizations do not scale in team size and this is not yet considered a problem, because so many undergraduate courses have teams of size 4. But we are working towards visualizations that scale as well in team-size as they do in confidentiality.

4. REFERENCES


