Causal Discovery in Social Media Using Quasi-Experimental Designs

Hüseyin Oktay, Brian J. Taylor, David D. Jensen
Knowledge Discovery Laboratory
Department of Computer Science
University of Massachusetts Amherst
140 Governors Dr.
Amherst, MA 01003
{hoktay,btaylor,jensen}@cs.umass.edu

ABSTRACT

Social media systems have become increasingly attractive to both users and companies providing those systems. Efficient management of these systems is essential and requires knowledge of cause-and-effect relationships within the system. Online experimentation can be used to discover causal knowledge; however, this ignores the observational data that is already being collected for operational purposes. Quasi-experimental designs (QEDs) are commonly used in social sciences to discover causal knowledge from observational data, and QEDs can be exploited to discover causal knowledge about social media systems. In this paper, we apply three different QEDs to demonstrate how one can gain a causal understanding of a social media system. The conclusions drawn from using a QED can have threats to their validity, but we show how one can carefully construct sophisticated designs to overcome some of those threats.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithms, Design, Experimentation

Keywords

Causal Discovery, Social Media Systems, Observational Data, Peer-Production Systems, Quasi-Experimental Design

1. INTRODUCTION

Social media systems, specifically collaborative platforms for asking questions and providing answers such as Yahoo! Answers[1] and Stack Overflow[2], can be seen as implicit knowledge discovery systems. By grouping their content around user questions, these platforms organize the collective information of individuals and allow their users to identify useful content and interesting knowledge. Bian and coauthors [5] considered question-answer platforms as a form of information retrieval, and here, we further consider them as a form of knowledge discovery.

Social media systems are widely used and intrinsically interesting to analyze for two main reasons. First, these systems allow us to collect data about human social interaction. This data can be analyzed to identify cause-and-effect relationships in a way that reveals the human behavior within social media. Second, causal analysis of this data can provide a better understanding of the impact of system design, how users collaborate and interact with one another, and how best to incentivize individuals within peer production to increase user contribution and improve content quality. Such understanding is essential for managers of these systems [9] and can provide useful insights for making better design choices for future versions of the system.

Efforts to discover cause-and-effect relationships in social media have taken three basic forms. First, Kohavi and coauthors [14] presented online experimentation to identify causal relationships. This method has the advantage of exploiting random assignment; however, it typically requires full control over the system to perform experiments. Furthermore, an experimental platform is needed for efficient online experimentation [13], which requires a high upfront cost. Even with an experimental platform, online experiments take considerable resources, thus experimenters want to strategically choose the right experiments to perform. The second line of work focuses on surveys among users to gather data for causal conclusions [10, 17]. This is an expensive method and requires careful implementation of the whole process to account for all potential biases. The third main line of work is on game theoretic formulations of social media systems to understand the user behavior [20, 10]. Different incentive designs are compared to understand user motivation.

Much of this previous work requires experimentation or additional data collection to identify causal knowledge, yet none of it exploits the already-collected observational data of social media systems. Observational data exists through standard logging of day-to-day system usage by users and in the database of content within the system. Social scientists use a class of methods known as *quasi-experimental designs*

[1] http://answers.yahoo.com/
[2] http://stackoverflow.com/

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

(QEDs) to discover cause-and-effect relationships [19] from observational data. Discovering these relationships by QEDs is beneficial for two reasons. First, QEDs do not require an experimental platform or additional data collection. Second, performing QED analysis is often cheaper and quicker when compared to performing experiments.

In this paper, we present how causal conclusions can be made about social media systems simply by analyzing existing data through the use of QEDs. We identify three different causal questions about Stack Overflow, a question-and-answer social content website. By performing the analysis suggested by our designs, we are able to understand several interesting behaviors by users within the system. For example, QEDs have allowed us to discover that (1) the existence of a previous high-quality answer does not discourage other users from contributing subsequent answers; (2) when two answers have equally high ratings, users prefer the newer answer regardless of the order in which the answers are provided; and (3) the contributions of highly active users decline shortly after receiving community recognition of that activity level. These conclusions can be made with high confidence due to the use of QEDs.

In section 2, we briefly introduce QEDs and threats to validity for experimental studies in general. In section 3, we overview the specific social media used in our analysis, Stack Overflow. In section 4, we ask three different causal questions and identify potential QEDs to answer each of these questions. We also perform analysis of these designs and discuss threats that are not eliminated in our conclusions for each design. In section 5, we overview existing literature on discovering knowledge from social media. Section 6 concludes this paper and poses future directions.

2. QUASI-EXPERIMENTAL DESIGNS

In experimental settings, random assignment of experimental units to treatment variables [7, 8] is often used. Randomization ensures that, in expectation, the effects due to all non-treatment variables are equivalent. Having probabilistically identical groups is an ideal condition for assessing the effect of treatment on experimental units.

In many social media applications, randomization may not be viable due to economic and experimental integrity concerns. It may be too costly to design an experiment and then deploy it over millions of customers. Businesses are likely to be careful about the design of experiments and hesitant to deploy anything that sours their users' interaction with a system. In many social media systems, collaboration and sharing are an integral component of the experience. Individuals within an experiment may be able to communicate and thereby eliminate the independence between treatment groups, invalidating any causal conclusions.

QEDs are a type of design that is often used in circumstances when random assignment of treatment is either impossible or infeasible. Other than lacking random assignment, QEDs have purposes and attributes similar to those of randomized experiments. Designs generally work by identifying an experimental unit that has undergone treatment and comparing it to another experimental unit that has not undergone treatment but which is similar in other aspects.

The simplest design is to compare a unit against itself by evaluating a change in an outcome variable on that unit in a series of pre-tests and post-tests. When the data permits, more sophisticated designs allow the creation of comparison groups that match different units in similarity save for the treatment. QEDs which come closest to true randomized experiments specifically model the reason treatment was applied to one unit and not to another and potentially yield an unbiased analysis.

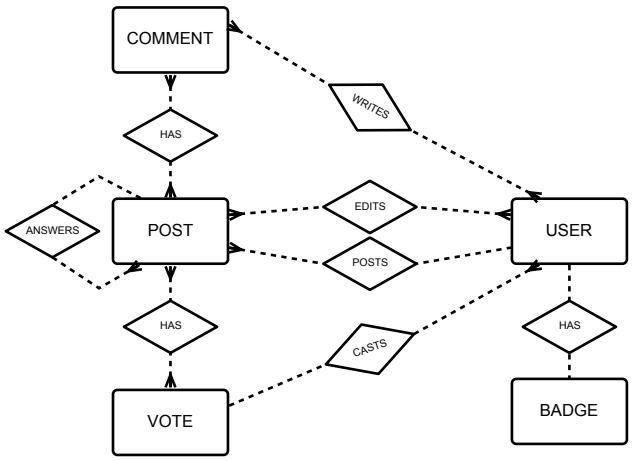

Figure 1: Entities in Stack Overflow

Since the groups under comparison are not necessarily identical, there may be concerns about the conclusions that can be drawn over these experiments. These concerns are called threats to validity. To justify the conclusions, one typically rules out as many of the threats to validity as possible. Since each quasi-experimental design is subject to its own set of threats to validity, one can apply multiple QEDs to eliminate multiple threats. Threats can also be ruled out by utilizing additional strategies, namely, utilizing domain knowledge and employing more sophisticated designs. Shadish, Cook, and Campbell [19] identify threats to four categories of validity:

- **Statistical Conclusion Validity:** The confidence in the statistical methods utilized to identify the correlation between treatment and outcome variables in a unit.
- **Internal Validity:** The confidence of an assertion that observed correlation in a unit is a causal relationship.
- **Construct Validity:** The confidence in the selection of metrics for variables that are measured.
- **External Validity:** The confidence in generalization of inferred relationships to alternative groups of units.

3. STACK OVERFLOW

Stack Overflow is an online platform where users can exchange knowledge related to programming. The content of the platform is completely provided by users. There are three main services in Stack Overflow. First, users can ask questions. The users are restricted to asking questions related to programming, and the moderators on the system are very strict on this policy. Second, users can share their knowledge by providing an answer to a particular question. Third, users can explore questions for which answers are already provided and can discover interesting facts about the topics that they are interested in.

Like other social media systems, Stack Overflow continuously collects observational data from users. This includes the questions users ask, the answers they provide, how users

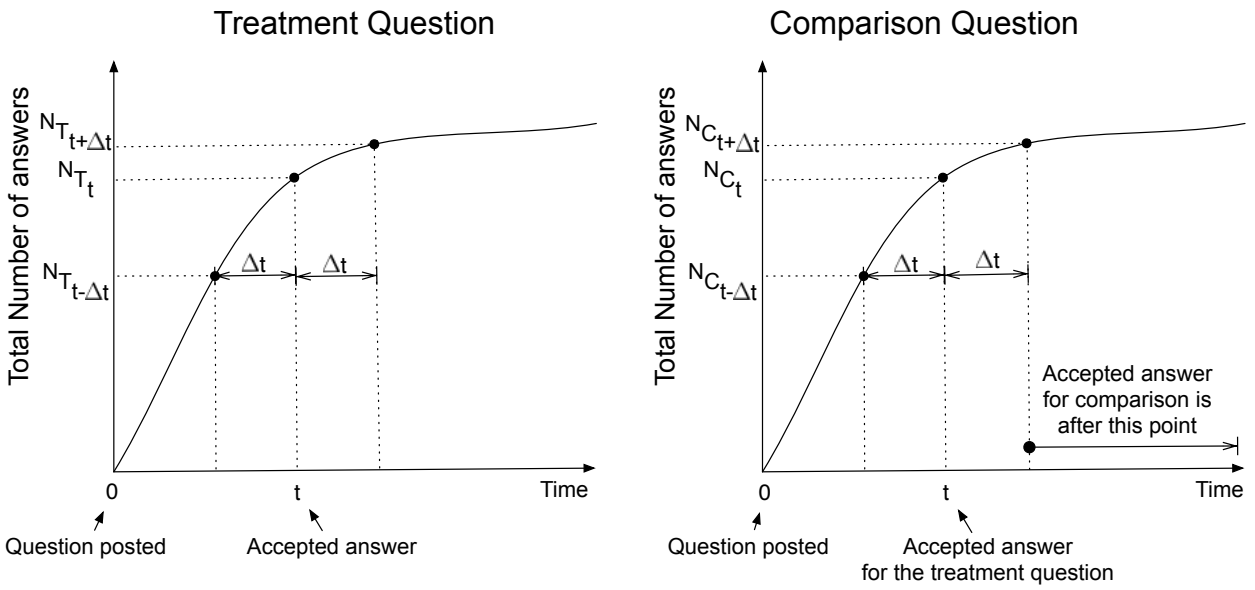

Figure 2: Matching process for treatment and comparison questions

rate each answer or question, and even how user reputations change over time. In the website, many entities interact and affect each other. Understanding the interactions among different entities has valuable implications when designing future versions of the system. Stack Overflow provides an excellent opportunity for exploiting existing data by applying QEDs to discover causal knowledge.

There are five main entities in Stack Overflow, as shown in a simple entity-relationship diagram in Figure 1: (1) users, (2) posts, which represent the questions and answers in the system, (3) comments, (4) votes, and (5) badges.

Registered users can ask and answer questions and leave comments discussing the different posts. A user is limited to certain capabilities in the system, dependent on the number of reputation points earned. Thus in addition to the intrinsic benefit users get from using the website, they are also motivated through the acquisition of points and status.

In Stack Overflow, users get reputation points when one of their answers or questions receives positive votes from other users. Users lose reputation points when one of their answers or questions receives negative votes. After accumulating a certain number of reputation points, a user's capabilities increase in the system. One interesting capability is that users with enough points can edit other individual's answers and questions to improve their quality. This capability is expected to improve the quality of questions and answers within Stack Overflow, making it an authority on programming topics.

Questions and answers are treated as posts within the system. Each question is asked by a specific user and can be tagged according to the content of the question. A question can have at most five different tags. Tags make it easier for other users to find questions that they are interested in answering. Questions can have multiple answers provided by users other than the asker. A user may provide only one answer for a particular question.

Users can provide comments on both questions and answers. Comments can be used to clarify the question or answer, and they provide a mechanism for a discussion between the users. Users may leave any number of comments.

Votes are used by users to indicate a positive or negative opinion of a post. Questions and answers can be voted up or down by users who have enough reputation points to do so. Users can get answers for their questions anytime after questions are posted, and they can determine the accepted answer anytime after they start getting answers. If no answer is selected, then the system will automatically assign the answer with the highest vote as the accepted answer.

Users can earn a variety of badges. A couple of examples are a *famous question* badge for users who ask a question with 10,000 views, an *autobiographer* badge for users who complete all user profile fields, a *good answer* badge for users who provide an answer that is voted up by 25 users, and a *epic* badge for users who hit the daily reputation point limit on 50 days.

4. USING DESIGNS TO DISCOVER CAUSAL KNOWLEDGE

In this section, we show the full process of applying QEDs to discover causal knowledge by using three example designs to answer different causal questions about Stack Overflow. For each design evaluation we use the following process: First, we identify an interesting causal question from Stack Overflow. Second, we apply a preliminary design (and perhaps additional, more sophisticated designs) and show the results. Finally, we discuss the conclusions drawn from the results as well as the threats to the validity of such conclusions. For all of our analysis, we used the Stack Overflow data dump which is released in February 2010.

4.1 The Matched Design

The first design we consider is a matched design. The matched design identifies pairs of units where one has received a treatment and the other has not but who are otherwise similar.

Time	Experiments		
Δt (in minutes)	No Matching	Random Pairs	Matching
5	-2.46	-2.04	1.14
10	-0.90	-0.57	0.48
15	-0.78	-0.66	NS
20	-0.45	NS	NS
25	-0.52	-0.25	NS
30	-0.36	-0.24	0.18
60	-0.24	-0.12	NS
90	-0.10	-0.07	NS
120	-0.10	NS	NS
150	-0.03	NS	NS
180	-0.05	NS	NS
210	-0.04	NS	NS
240	-0.04	-0.03	NS

Table 1: For *No Matching* experiments, differences in answer rate for each time interval are shown. For *Random Pairs* and *Matching*, differences between the answer rate change for treatment group and the answer rate change for comparison group are shown. NS means *Not Significant*. Values are number of answers per hour.

The validity of the causal conclusions drawn from a matched design improves as the matched pairs become more alike.

A causal question that can be tested with this design is:

> *Does posting a high-quality answer for a particular question cause other users to stop providing answers for that question?*

The user who asks a question can select one answer and this selected answer is called the *accepted answer*. The accepted answer is often selected long after it is initially posted, and this time lag allows us to examine the effect of high-quality answers (as measured by its selection as the accepted answer) on behavior that occurs before it is actually selected as the accepted answer. In this causal analysis, we want to identify whether posting a high-quality answer for a particular question has any effect on the answer rate for that question. In general, this test will analyze whether users stop providing answers for a question once they think a good answer has been provided. We assume that the accepted answer has high quality for that particular question since the user who asks the question chooses the answer.

For comparison purposes, we apply three different designs. In the first experimental setup, we only examine treatment questions (i.e., questions with accepted answer) without any matching. In the second setup, we have treatment-comparison pairs where comparison questions are randomly selected from questions that have the same tag. Here, we make sure that comparison questions do not have an accepted answer prior to the relative time the treatment has an accepted answer. In the third setup, we match treatment and comparison questions by requiring a much stronger similarity within pairs.

The first design is a simple statistical analysis that evaluates the change in answer rate around the time an accepted answer occurs. Time is measured on a relative scale where the creation time of a question is 0 for all questions and t is the time the accepted answer is posted for a question.

The outcome measure in this test is the difference between the answer rate for Δt minutes after an accepted answer is posted and the answer rate for Δt minutes before the accepted answer is posted:

$$\textit{answer rate change} = \frac{N_{T_{t+\Delta t}} - N_{T_t}}{\Delta t} - \frac{N_{T_t} - N_{T_{t-\Delta t}}}{\Delta t} \quad (1)$$

where $N_{T_{t+\Delta t}}$ is the number of answers for the treatment question at time $t + \Delta t$, N_{T_t} is the number of answers for the treatment question at time t and $N_{T_{t-\Delta t}}$ is the number of answers at time $t - \Delta t$ (See Figure 2).

We apply this design using Δt values ranging from 5 minutes to 240 minutes. The results are shown in Table 1 under the column *No Matching*. The difference in the answer rate Δt minutes after an accepted answer is posted (i.e., our treatment in this design) and Δt minutes before an accepted answer is posted is a negative number. This suggests that the answer rate is greater before the treatment and that there is a decrease in the number of answers provided after a high-quality answer is posted. Although this simple analysis identifies a decrease in answer rate after the treatment, it is not clear whether the cause of this change is the appearance of the eventually accepted answer (i.e., treatment). What if the decrease in the answer rate is caused by the amount of elapsed time rather than the presence of the accepted answer?

To eliminate the effect of some other unobserved variables (e.g., time), we can use the second design, a basic matching design. We pair each treatment question with a random question to better compare the difference in the behavior. The matching goal is to factor out some of the unobserved variables by capturing the trend around time t with the comparison group.

To find a matched pair, we first select a random question with an accepted answer posted at some time t. Then, we find a comparison question by selecting another question that has the same tag as the treatment question and for which an accepted answer occurs after time $t + \Delta t$. We assume that questions having the same tag are similar in topic to one another and have a similar pattern to their overall answer rates. By choosing comparison questions with accepted answers occurring at least $t + \Delta t$, we assume they do not have a treatment applied to them.

The outcome measure for this design is the difference between the answer rate change of a treatment question (i.e., T_{arc}) and the answer rate change of its comparison question (i.e., C_{arc}) within the matched pair:

$$\textit{outcome} \;=\; T_{arc} - C_{arc} \quad (2)$$

T_{arc} and C_{arc} are calculated according to Equation 1 for both the treatment and comparison questions, respectively. In Figure 2, N_{C_t} is the number of answers for the comparison question at time t, $N_{C_{t+\Delta t}}$ is the number of answers for the comparison question at time $t + \Delta t$, and $N_{C_{t-\Delta t}}$ is the number of answers for the comparison question at time $t - \Delta t$.

A positive answer rate change indicates that more answers occur after time t within the Δt window while a negative answer rate change means fewer answers show up after time t. If the outcome measure is positive, it means that the treatment question experienced more answers occurring after a high-quality answer is posted, $t + \Delta t$, than did the comparison. When the outcome is negative, it means that

the treatment question experienced fewer answers occurring after a high-quality answer is posted than the comparison question.

As shown in the *Random Pairs* column of Table 1, with this more sophisticated design, we conclude that at least in some of the cases the difference between the treatment and comparison questions is not significant when compared to the simple design without matching. We conclude, in those cases, the accepted answer has no effect on answer rate. We can also observe that the real effect is not as large as was observed for the former design without pairs. This shows that we can have a more thorough analysis by using designs that rule out larger number of threats to the validity of conclusions.

Although this design forms a pair of questions to compare, we are not guaranteed to find perfectly matched pairs. In the ideal case, we would like the questions in our pairs to be identical except for the fact that one has a treatment and the other does not. However, since we do not have control over assigning treatment, we may pair questions with different answer rates. In such cases, any difference we observe may be due to the inherent difference in their answer rate rather than the high-quality answer provided.

To create better matched pairs in the third design, we use two criteria: First, we want the treatment question and the comparison question to have almost the same number of answers provided before time t. Second, we want the matched pair to have a similar answer rate for the specified time interval, $t - \Delta t$. From these two metrics, we formulated a similarity metric which is calculated for treatment questions as follows:

$$similarity\ metric\ =\ \frac{\frac{N_{T_t}}{t} + \frac{N_{T_t} - N_{T_{t-\Delta t}}}{\Delta t}}{2} \qquad (3)$$

For each treatment question, a comparison question is chosen that has the same tag as the treatment question and that is the closest in terms of the similarity metric. As before, we also require that comparison questions have no accepted answer before time $t + \Delta t$. Using the procedure described above, we identified 200 matched pairs.

This design uses the same outcome measure as the second design. With this matching design, we are able to identify a positive effect for only the 5- and 10-minute time intervals. This suggests that there is much less decrease in the answer rate when compared to the comparison question after time t when a high-quality answer is provided. In other words, having a high-quality answer actually increases the answer rate when compared to the comparison group.

A close observation reveals that posting an answer may take on average 5 to 10 minutes, and there may often be one other user who is in the process of typing an answer around the time the accepted answer is posted—accepted answers are usually posted not too long after the question is posted. We may conclude that it is not that high-quality answers increase the answer rate, but that users type their answers for the question without knowing about other users who are also providing answers for the same question, especially right after a question is posted. We suspect that the significant result for the 30-minute time interval is a false positive error given its small magnitude and the overall set of non-significant results.

As shown in Table 1, we can reach different conclusions depending on the design we use. The most sophisticated matching design shows that having a high-quality answer has no effect on answer rate whereas the previous designs do show an effect. Even in some cases where the results of the random pair design suggest that a high-quality answer has an effect, the matching design is able to conclude that the effect is not significant.

Vote Number	Number of Instances	P-Value	χ^2-Statistic	Frequency
1	87405	< 2.2e-16	2684.1770	0.41
2	45899	< 2.2e-16	2207.9890	0.39
3	24908	< 2.2e-16	1042.6050	0.40
4	14260	< 2.2e-16	807.8006	0.38
5	8555	< 2.2e-16	532.8141	0.38
6-8	12291	< 2.2e-16	638.3208	0.39
9-12	6434	< 2.2e-16	337.6867	0.38
13-17	3748	< 2.2e-16	145.3159	0.40
18-26	3593	< 2.2e-16	109.4153	0.41
26-665	9728	< 2.2e-16	160.6188	0.44

Table 2: The results of a chi-square test on the frequency of votes for the older answer before the policy change. Degree of freedom is 1 for each stratum.

In an ideal matching design, good matching criteria should eliminate all alternative explanations of the observed effect. Depending on the causal question, an analyst can identify matching criteria to control for most of the variables that can influence the observed effect. For example, matching criteria could control for the tag of the question as well as some property of the users posting the questions, like their badges or reputation points. Regardless of the matching criteria used, however, the analyst should continue to think about alternative explanations at the end of the analysis.

4.2 The Natural Experiment Design

A natural experiment is a condition within the observed dataset which approximates a randomized experiment. Such a condition can occur if a social media system changes a single aspect, like a user interface, and has data collected both before and after the change. While the system change was never intended to be a treatment used in an experiment, a quasi-experiment can look at the data as if it was.

A causal question in Stack Overflow for this design is:

> *For a particular answer, does being displayed above other answers cause it to get more vote-ups?*

To answer this question, a policy change in Stack Overflow can serve as a natural experiment. In Stack Overflow, answers for a question are sorted in descending order in terms of their net number of votes. To break ties, two different approaches are taken in the system. Before August 2009, ties were broken in terms of the creation date of the answers. Older posts got higher priority and were listed higher on the page when there was a tie in the number of votes. After August 2009, Stack Overflow managers changed their policy and decided to break ties randomly, which removed bias in ordering answers for older posts. Answers are sorted randomly when they have the same net number of votes.

To exploit this implicit natural experiment, we consider *tie-breaking votes* before and after the policy change as our data instances. A tie-breaking vote is a vote-up that is cast for an answer that is in a tie with exactly one other answer.

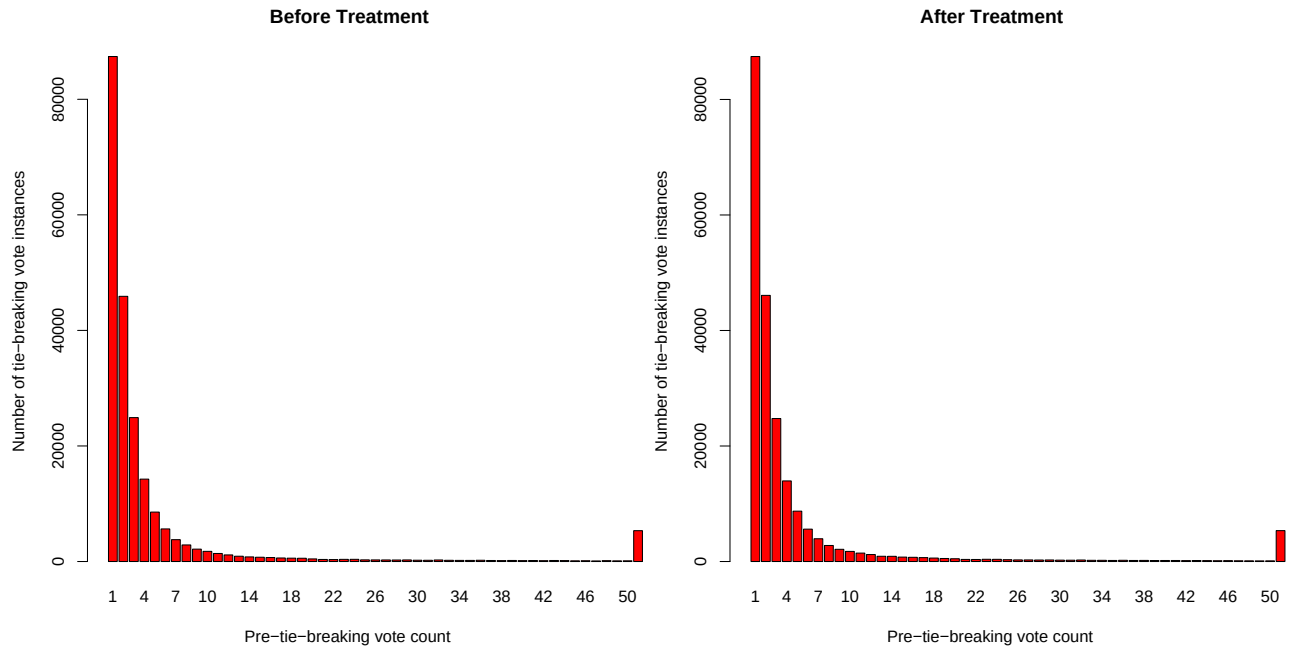

Figure 3: Histogram for tie-breaking votes before and after the policy change. The last bar is the aggregate of all the tie-breaking votes of answers for which the pre-tie-breaking vote count is greater than 50.

Our treatment variable is the way the answers are ordered in a tie situation (i.e., either from oldest to newest or randomly). Our outcome variable is the frequency of voting for the older answer when there is a tie. If the order of answers has an effect on voting, we would expect to see a significant difference between the frequency of voting for the older answer before and after the policy change.

We randomly choose more than 200,000 tie-breaking votes, both before and after the policy change. For each vote, we assign a binary value that is 0 if the vote is for the newer answer in the tie situation or 1 if the vote is for the older answer. Then we do a chi-square test to determine if those values are significantly different than a binomial distribution where the success probability is 0.5. In such a binomial distribution, we would expect to have equal numbers of votes for newer answers and for older answers. We do the same test for the votes before and after the policy change.

The frequency of tie-breaking votes for the older answer is 0.40 before the policy change, and the chi-square test reveals that this frequency is significantly different than 0.5 with χ^2=8487.76 and p<0.001. Similarly, after the policy change, the frequency of tie-breaking votes for the older answer is 0.40, and the chi-square test reveals that this frequency is also significantly different than 0.5 with a χ^2=8424.14 and p<0.001. These results show that users are more likely to vote for the newer answer than for the older answer regardless of the policy change. The results could also imply that the newer answers are often of a higher quality than older answers.

To assess the effect of the policy change, we compare the frequency of tie-breaking votes for the older answer before and after the policy change. We do a *two-sample unpaired t-test with two-tail analysis*, and the results reveal no significant difference between the frequencies of tie-breaking votes for the older answer before the change and after the change with t=0.35, degrees of freedom = 433640, and p=0.73. We can conclude that the data does not support the hypothesis that the policy change had an effect on voting behavior.

Recall that in our dataset of tie-breaking votes, we have votes that correspond to a vote-up for an answer for which the pre-tie-breaking vote count (i.e., the number of existing vote-ups of this answer right before this tie-breaking vote) is in a tie situation with exactly one other answer. This vote-up in our data set breaks this tie between these two answers (i.e., answer pairs) either by voting for the older answer or the newer answer. A closer observation of this vote distribution for the tie-breaking votes in our dataset is shown in Figure 3. In these figures, the x-axis shows the number of pre-tie-breaking votes and the y-axis shows the frequency of tie-breaking votes. For example, from Figure 3, we can see that in our dataset there are more than 20,000 tie-breaking votes that break the tie between answer pairs for which the pre-tie-breaking vote count is three. Both answers in the pair already have three vote-ups, and the tie-breaking vote in our dataset breaks this tie by increasing the vote number for one of the answers (either the older answer or the newer answer) by one.

The histograms in Figure 3 show that many of the tie-breaking votes are cast for answer pairs when there is exactly one pre-tie-breaking vote for each answer in these pairs. The number of qualifying-vote instances decreases rapidly as the number of pre-tie-breaking votes goes up for two reasons. First, since vote-ups are counted cumulatively for an answer, there are more vote-up instances that correspond to answers for which the pre-tie-breaking vote counts are small than instances that correspond to answers for which the pre-

tie-breaking vote counts are large. Second, being in a tie situation with another answer when the pre-tie-breaking vote count is small is more likely than when the pre-tie-breaking vote count is large. For these two reasons, we get more tie-breaking votes that correspond to answers for which the pre-tie-breaking vote counts are small when we randomly sample from vote-up instances (i.e., there is a bias towards small pre-tie-breaking vote counts for data instances both before and after the treatment as shown in Figure 3).

To better account for the difference in the distributions seen in Figure 3, we stratify our data points into 10 strata, as shown in the vote number column of Table 2. For each stratum, we perform the chi-square test to see if the frequency of tie-breaking votes for older answers is different than the binomial distribution where probability is 0.5. Tables 2 and 3 show the results for the data points in each stratum before and after the policy change, respectively. Those tables show that, for each stratum, the frequency is significantly different than 0.5, both before and after the policy change.

We also performed an two-sample unpaired t-test with two-tail analysis to see whether, in each stratum, the frequency of voting for the older answer before the policy change is different than the frequency after the policy change. Table 4 summarizes the results of these tests. Except for strata 2 and 3, the results were not significant, indicating that we cannot accept the hypothesis that the frequency of the vote for the older answer differs before and after the policy change. For strata 2 and 3, although we get a significant result, the difference in the frequency values is extremely small, suggesting a very small effect, if any.

Another recent policy change in the Stack Overflow system sets up another natural experiment to assess whether there is a causal relationship between reputation points and asking questions. Users get reputation points when their questions or answers get a vote-up. Before the recent policy change, users received 10 points for a vote-up, both for a question and an answer. However, after this change, users get 5 reputation points after a vote-up for their questions and still get 10 reputation points after a vote-up for their answers. The goal of this change is to decrease the number of questions a user asks and increase the number of answers that a user provides.

This natural experiment can be exploited to identify the effect of the change in vote-up reputation points on user behavior. The treatment variable is the change in vote-up reward for questions. The outcome variable is the number of questions provided by a particular user. We seek to answer the following question: *Will users provide more answers and fewer questions after this policy change?* Although this is an interesting natural experiment, we could not apply this design because the policy change is recent and data collected after the treatment event is not yet sufficient to perform the analysis.

4.3 The Interrupted Time-Series Design

In the *interrupted time series* design, we observe an outcome variable for a certain time interval, Δt, before a treatment and after the treatment. This observation over Δt lets us identify intrinsic trends within the time-series and therefore rule out some threats to validity. The causal question we consider that uses this design is:

Vote Number	Number of Instances	P-Value	χ^2-Statistic	Frequency
1	87436	< 2.2e-16	2652.1340	0.41
2	46079	< 2.2e-16	1915.5370	0.40
3	24756	< 2.2e-16	1253.2280	0.39
4	13954	< 2.2e-16	884.9216	0.37
5	8718	< 2.2e-16	527.2696	0.38
6-8	12323	< 2.2e-16	670.7478	0.38
9-12	6530	< 2.2e-16	271.7035	0.40
13-17	3970	< 2.2e-16	118.5380	0.41
18-26	3668	< 2.2e-16	139.7644	0.40
26-665	9810	< 2.2e-16	177.5423	0.43

Table 3: The results of a chi-Square test on the frequency of votes for the older answer after the policy change. Degree of freedom is 1 for each stratum.

Vote Number	P-Value	t-Statistic	Frequency Before	Frequency After
1	0.81	0.2332	0.41	0.41
2	0.02	2.3235	0.39	0.40
3	0.02	-2.3376	0.40	0.39
4	0.25	-1.1484	0.38	0.37
5	0.90	0.1263	0.38	0.38
6-8	0.66	-0.4326	0.39	0.38
9-12	0.15	1.4448	0.38	0.40
13-17	0.19	1.3155	0.40	0.42
18-26	0.41	-0.8243	0.41	0.40
26-665	0.69	-0.3994	0.44	0.43

Table 4: The results of a two-sample unpaired t-test with two-tail analysis for comparing the frequency of voting for the older answer before and after the policy change.

> *Does receiving an epic badge cause an increase in participation for that user?*

This causal question can be formulated across many different badges. However, we focus on only the *epic* badge, which is given to users who hit the daily reputation cap 50 times. For this design, the treatment is a user getting the epic badge. The time-series will be the number of posts per day by the user with the outcome measure being the change in the number of posts by that particular user before and after treatment. We use number of posts provided by a user instead of number of points obtained by a user as our outcome metric because number of posts is only influenced by the corresponding user, whereas number of points is influenced both by the corresponding user and the external events (e.g., another user can vote-up a question/answer). Hence, number of posts is a better metric for user behavior.

Threats to validity of the interrupted time-series include historical effects. For example, users may actually have a decreasing trend in their number of posts at the same time as treatment occurs. Without getting multiple observations of the outcome measure, we may falsely conclude that this decrease is due to the treatment effect; however, it may be actually due to an intrinsic decreasing trend.

There are 54 users with the epic badge in our dataset. For each user, we determine the relative time at which they

get the epic badge. Then we calculate the number of posts corresponding to each user for 30 days before they get the epic badge and 30 days after they get the epic badge. To make the analysis clearer, we calculate the average number of posts for the 30-day period before the treatment for each user, and we normalize the daily number of posts for each user by subtracting that average. We then calculate the average number of posts from those normalized values among 54 users for each day and plot those values in Figure 4. We fit linear models to the data points before and after the treatment. If we observe a significant slope change in these two fitted linear models, we can conclude that treatment has an effect. If the slope of the post-treatment line is smaller than the pre-treatment line, we can conclude that users start to contribute less after getting the epic badge and vice versa.

In Figure 4, we show the results of the interrupted time series design. The vertical dashed line represents the 30-day mark for each user at which the epic tag was granted. As mentioned above, we fit two linear models: The first linear model is for the average number of posts before the badge is granted and the second linear model is for the average number of posts after the badge is granted (note that values for average number of posts are normalized). The slope of the first line is -0.001 and this slope is not significantly different than 0 ($p=0.94$). The slope of the second line is -0.100 and this slope is significantly different than 0 ($p<0.01$). We observe a decrease in the slope of these lines. We may conclude that getting the epic badge reduces the number of posts provided.

In our interrupted time series design, we do not have a comparison group which would allow us to account for other unobserved effects. Other events occurring at the same time as our treatment might explain the observed behavior. For example, the system may go down so that users cannot interact with the system, or the time of the establishment of the badge may be right before another event that causes them not to participate. However, we have more than 50 users who received this badge, and all measurements were done in a relative time scale for each user. Thus, we expect that these effects should average out. In other words, the time that a user obtains the epic badge can be treated as random.

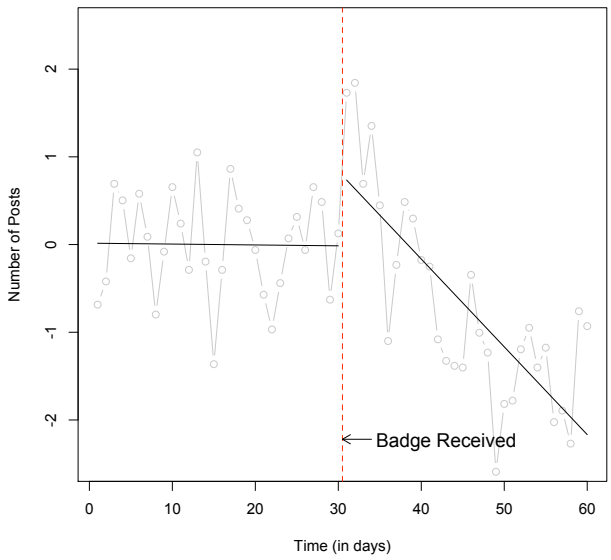

Figure 4: Average post count for users with an epic badge. Values are normalized for each user by subtracting the corresponding average number of posts before treatment.

5. RELATED WORK

Discovering cause-and-effect relationships from observational data by using QEDs has been widely studied in the social sciences [6, 19]. Social scientists use QEDs in circumstances for which experimentation is not feasible for various reasons (e.g., ethical, economic). However, these methods have not been extensively used by researchers studying social media. Jensen and coauthors [12] showed that these methods can be automatically identified, and Aral, Muchnik and Sundararajan [4] used propensity score matching, a type of QED, to distinguish homophily and social influence. Some researchers have compared different social media systems that have similar purposes, which is a type of QED with non-equivalent control groups [9, 11].

Computer scientists have recently focused on designing algorithms for causal discovery. Spirtes, Glymour and Scheines [21] proposed the PC algorithm for identifying causal structure from data in a propositional domain under certain assumptions. Maier and coauthors [16] extended the PC algorithm to relational domains, pushing causal discovery into more complex domains. Pearl [18] also provided a more theoretical framework for algorithmic causal discovery.

Knowledge discovery in social media systems has often been based on associational discovery. Researchers have tried to identify high-quality content in social media [2], and to identify attributes that predict answer quality in question and answer platforms [1, 9]. Maia and coauthors [15] identified discriminator attributes in a social network setting to cluster users by their behavior. These studies are almost all based on identifying correlations between variables, and they do not try to identify cause-and-effect relationships.

As outlined in section 1, existing research on causal discovery from social media focuses on three main approaches. The first is online experimentation, which is widely used [22, 14], and is an ideal tool for causal discovery. However, it often requires full control over the system, necessitating an experimental platform [13] to perform online experimentation cost effectively. Another approach used for causal discovery is surveying users of the system to understand their behavior and motivation [10, 17, 3]. Researchers have also used game-theoretic platforms to understand user behavior in social media [20, 10]. Although their work shares the goal of causal discovery in social media systems, it requires additional experimentation and data collection rather than exploiting observational data.

6. CONCLUSION AND FUTURE WORK

In this work, we show that QEDs can be utilized to discover causal knowledge about social media systems. We identify and carry out three different designs in Stack Overflow. First, we show a matching design and demonstrate how an analyst can reach false conclusions without using a matching design. Second, we show a natural experiment

that is hidden in the system and analyze this implicit existing experiment. Third, we present an interrupted time series design to observe user participation after a particular event (i.e., obtaining a badge in our scenario). For all three designs, we point out the threats to the validity of the conclusions.

Manual application of QEDs can be time consuming and involve a process of successively eliminating threats to a causal conclusion. A better approach is one where an algorithm automatically identifies all of the applicable designs, executes those designs, and then eliminates the threats. One of our goals is to eventually develop algorithms capable of this automated discovery.

QEDs can be combined with online experimentation. Since performing all possible experiments is often infeasible, QEDs can be used to filter out some of the potential experiments that correspond to alternative hypothesis. For a causal question, first a quick QED analysis can be performed, and then for the alternative threats that a particular QED is unable to eliminate, an online experimentation can be formulated.

7. ACKNOWLEDGEMENTS

Discussions with Matthew J. Rattigan contributed to the paper. Assistance with programming and inspiring discussions were provided by Marc E. Maier. Helpful comments and patient editing were made by Cynthia Loiselle and Stephen M. Constantine. This work was generously supported by a gift from Yahoo! Research. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of Yahoo! Research.

8. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proc. of WWW '08*, pages 665–674, New York, NY, USA, 2008. ACM.

[2] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM '08: Proceedings of the International Conference on Web Search and Web Data Mining*, pages 183–194, New York, NY, USA, 2008. ACM.

[3] M. Antikainen and H. Väätäjä. Rewarding in open innovation communities – how to motivate members? In *Proc. ISPIM Conference*, 2008.

[4] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[5] J. Bian, Y. Liu, E. Agichtein, and H. Zha. Finding the right facts in the crowd: Factoid question answering over social media. In *WWW '08: Proceeding of the 17th International Conference on the World Wide Web*, pages 467–476, New York, NY, USA, 2008. ACM.

[6] D. Campbell and J. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago, IL, 1966.

[7] R. A. Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:505–513, 1926.

[8] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver Boyd, Edinburgh, 1950.

[9] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan. Predictors of answer quality in online Q&A sites. In *CHI '08: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 865–874, New York, NY, USA, 2008. ACM.

[10] S. Jain, Y. Chen, and D. C. Parkes. Designing incentives for online question and answer forums. In *EC '09: Proceedings of the Tenth ACM Conference on Electronic Commerce*, pages 129–138, New York, NY, USA, 2009. ACM.

[11] J. Janes, C. Hill, and A. Rolfe. Ask-an-expert services analysis. *J. Am. Soc. Inf. Sci. Technol.*, 52(13):1106–1121, 2001.

[12] D. D. Jensen, A. S. Fast, B. J. Taylor, and M. E. Maier. Automatic identification of quasi-experimental designs for discovering causal knowledge. In *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 372–380, New York, NY, USA, 2008. ACM.

[13] R. Kohavi, T. Crook, and R. Longbotham. Online experimentation at Microsoft. In *Proc. of the Third Workshop on Data Mining Case Studies*, 2009.

[14] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Min. Knowl. Discov.*, 18(1):140–181, 2009.

[15] M. Maia, J. Almeida, and V. Almeida. Identifying user behavior in online social networks. In *SocialNets '08: Proceedings of the 1st Workshop on Social Network Systems*, pages 1–6, New York, NY, USA, 2008. ACM.

[16] M. Maier, B. Taylor, H. Oktay, and D. Jensen. Learning causal models of relational domains. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[17] K. K. Nam, M. S. Ackerman, and L. A. Adamic. Questions in, knowledge in? A study of Naver's question answering community. In *Proc. of the 27th International Conference on Human Factors in Computing Systems*, pages 779–788, Boston, MA, 09/2008 2009.

[18] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[19] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, MA, 2002.

[20] V. K. Singh, R. Jain, and M. S. Kankanhalli. Motivating contributors in social media networks. In *WSM '09: Proceedings of the First SIGMM Workshop on Social Media*, pages 11–18, New York, NY, USA, 2009. ACM.

[21] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 2nd edition, 2000.

[22] S. Thomke. Enlightened experimentation: The new imperative for innovation. *Harvard Business Review*, 79(2), February, 2001.