The Intellectual Standard

Volume 2 | Issue 1 Article 4

10-2012

The Issue Of Internet Polling

Nick A. Nichols
Illinois Wesleyan University, nnichols@iwu.edu

October 2012

The Issue Of Internet Polling

The Issue Of Internet Polling

Nick Nichols

Surveys, polls, and focus groups are common phenomena in our daily lives. We live in a world where big data is big business. Large decisions hinge on the accuracy and predicative power of these numbers. Therefore, it should not be surprising that there is a market for the malicious manipulation of data. Extreme care must be taken in the collection, checking, and processing of data to prevent decisions from being made on incorrect assumptions. In order to demonstrate the full potential and possible impact of these attacks, I shall provide the following example:

John Doe is a member of the United States Senate. In recent years, the political pressure to make a preemptive strike against a potential nuclear threat has grown exponentially. In some of the more extreme cases, several senators have begun asking for support to make a motion to the President for military intervention. Eventually, Senator Doe is asked to sign a petition for their cause. Senator Doe decides that he must take the concerns, priorities, and beliefs of the voters in his state into account before he can make a decision as their representative.

In order to accurately and quickly gather the opinions and concerns of his constituents, John opens a polling section on his website which is advertised across his entire state. The poll asks for each participant to express their beliefs about the effectiveness of potential solutions, the immediate threat posed, and ultimately, whether or not he should endorse military action.

Historically, Senator Doe's state has opposed similar legislation and government involvement in foreign affairs; however, in this particular case, the voters have voted in favor of a preemptive military strike.[1] In order to do his job, John must sign the petition. However, if he does not take the proper precautions, he could be making his decision based on malicious behavior instead of the seemingly new political paradigm in his state.

Understandably, this is an extreme case of the problem posed. There would be many other considerations in real-life deliberations, but

1 Assuming that the number of responses was high enough to be an accurate portrayal of the voting population.

data like this carries considerable weight. Many politicians have similar polls on their websites for less pressing matters and make legislative decisions based on that data. Maintaining the integrity of that information is of the utmost importance. While security for Internet applications is an ever-distant goal, we do have tools to help us filter out bad data and limit our susceptibility to such attacks.

To begin, we need to establish a clear standard and goal by which a fair and open Internet poll can be considered successful. That goal is to collect the opinions of a specific population[2] in a way that each person's opinion is considered equally[3] while maintaining a high degree of accessibility. Many checks and methodologies exist to assist in solving the wide array of concerns and vulnerabilities present. This paper will primarily focus on "flooding" attacks, where single users respond many times in order to promote a single response, or set of responses.[4]

Any fair poll should give equal weight to each response, and ensuring this is a trivial matter; however, making sure equal weight is given to each person is a different matter. If a person responds to a poll multiple times, and each vote is counted fairly, then their voice is made louder. Malicious software that continuously votes in polls like these is fairly common. Within seconds, thousands of illegitimate responses can flood the system. How can the real data be sorted out?

In these instances we need to be careful about how we remove data. We need a method that doesn't rely upon our preconception of how popular the response is. If we used a methodology like this, then we would be led to assume that any surprising outcomes were the result of malicious voting. More information is needed to increase the probability that we find illegitimate votes without throwing out any of the legitimate votes.

Since these systems typically vote as fast as they can for a predetermined number of votes or time, it would be a reasonable first step to store a timestamp with each vote. Now if a large number of votes for the same option appear within a narrow time span, we have a clue that a block of

2 This should be a well defined population; i.e. the voting populace of Texas, the students of a specific class, or the visitors of a specific website.

3 This means that the votes shouldn't be weighted by status, repetitive voting is disallowed, etc.

4 By repeatedly voting for the same responses, the chosen option becomes a clear outlier, regardless of how unpopular the option may actually be.

data is probably the result of a flood. This method does have a drawback: it doesn't scale very well.

If a website is trying to collect the opinions of every person who comes across it, similar to CNN.com, then their large amount of traffic could easily account for behavior like this. The method also begins to fail if the attacker tunes down the speed of the flood and allows for a vote or two to interrupt a few of theirs. Some of these programs are also trained to select different responses with each vote, while still making their intended selection a clear statistical outlier. Clearly, we need a better detection strategy.

Instead of focusing on the votes themselves, it can be helpful to think about identifying the voter. A poll in which only account holding members are allowed to vote is an obvious choice. Then we can link the username to the vote. After voting once, a user cannot vote again, and with reasonable security measures we can prevent fake accounts from being quickly set up. With this option more information about the voter is available. Naturally, anonymity is a major concern. Further, requiring an account will be a deterrent to a large number of voters, which may skew polling data.

A less intrusive requirement is to store the IP address of the source of the vote. If we collect this alone, we have a means of identifying a computer without storing any intrusive information. Thus, votes that have matching IP addresses can be discarded; however, not every machine has a globally unique IP address. The number of devices connected to the Internet is far greater than the number of addresses. In order to continue adding devices without breaking the underlying structure that the Internet operates upon, Network Address Translation (NAT) and the Dynamic Host Configuration Protocol (DHCP) were created.

These two technologies allow a single address to be shared by multiple devices in the case of NAT, or a set of addresses to be leased out as needed with DHCP. So we cannot assume that each address directly corresponds to a unique individual, but this does help us narrow our search window. Now the attacker would have to vote at a reduced speed, spread the results across several categories, and occasionally swap IP addresses[5] to completely avoid suspicion. The added complexity will dampen the impact

5 Proxy services and Tor allow users to do this for anonymity's sake.

significantly, considering that each step requires a more complex program and consumes more time.

Naturally storing more information makes things more expensive on our end. It would be nice if the information could be stored on a computer after it had been used to vote. This common practice is accomplished with cookies, a datum that servers give to browsers for communicative purposes. If we check for cookies before any voting occurs, and make sure we distribute them after voting we can deny users that try to vote again. As with the other approaches, this has flaws as well. Cookies can be silently denied by browsers, and can also be deleted by users. There is no way to ensure that the data stored outside of our computer has not been tampered with or removed.[6]

Lastly, we can add a captcha to our site. Captchas are tests that are more easily solved by humans than computers. The most common example of this is extracting a word or two from a blurred, unclear, chaotic background. The human brain is able to recognize the patterns fairly well and see the real text. Since computer vision is still a budding area, they will typically fail these tests. Again, this method only works with some probability, and research into artificial intelligence reduces the effectiveness of such solutions.

No one of these methods is an absolute solution. Each gives us another screen to filter the data, but bad information can still get through. They are intended to reduce the potential impact of an attack by making the process more complex. These hurdles make it far more difficult and time consuming for a human to manually vote multiple times, and also vastly increase the complexity and time required to build an autonomous voter.

With this information in hand, it should now be clear why extra caution should be used in conjunction with an online poll. Simple attacks like these are one of many concerns associated with digital voting, and far more complex attacks are also quite common. Voting software, built by the Federal Government, to tally absentee ballots electronically were made public to security analysts for a test election. Researchers at the University

6 Technically speaking, through the use of a rootkit or another exploit, the cookie could be forcefully and secretly stored; however, this method is illegal and unethical.

of Michigan were able to crack the system in a matter of hours, and pushed Futurama's Bender to a winning position. During this attack, they were also able to remove their digital footprint from the server.[7]

Significant checks need to be added to make polls as reliable as possible, and plenty of real world examples demonstrate the absence of such systems. More research into this problem and the defense mechanisms associated with it is needed. The content creators of the Internet should use their due diligence when selecting and implementing any sort of software on their websites, especially when critical decisions are made based on the data that they collect. Researchers, planning committees, corporations, and politicians are currently using insecure polls and the data they collect every day. The data that can be gathered is powerful, and the ability to control it is even more impressive.

7 Slashtot.org, "Voting System Test Hack Elects Futurama's Bender To School Board," March 2, 2012.