Statistics Hacking - Exploiting Vulnerabilities in News Websites

Amrinder Arora

Department of Computer Science, George Washington University, Washington DC, 20052 USA

Summary
We analyze and discuss a vulnerability in leading news websites, that can lead to modification of the system statistics by malicious users (statistics hacking). We outline two broad categories of methods that can counter statistics hacking. The first category consists of many methods already available to distinguish human users from computers. We compare the different methods within this category on basis of security and accessibility. The second category uses well known clustering techniques applied to the web statistics. We compare the different methods within this category on basis of security.

Key words:
Statistics Hacking, Click Fraud, Link Spamming, Image Verification, Turing Test, CAPTCHA, Human Verification

1. Introduction

News websites often attempt to be interactive. In order to present an enhanced user experience, news websites attempt to measure user interest indirectly, as well as allow readers to give feedback, or allow them to vote on popularity of news, etc. This feedback feature can lead to vulnerability in their system. In this paper, we analyze the vulnerabilities and present mechanisms to combat them.

Example: BBC is one of the leading news websites and is often ranked in top 10 most visited websites. Due to this, it wields an enormous amount of influence. As a user feedback mechanism, BBC also monitors the popularity of a news item on at least two counts (i) Number of times the news item was read, and (ii) Number of times the news item was emailed. These correspond to the "Most E-mailed" and "Most Read" segments of the BBC website. This seemingly innocuous feature is often exploited by statistics hacker to affect the news (or their order) displayed by BBC. The very same feature also leads to traffic surges on the Internet, as statistics hackers attempt to outperform each other, in the process impacting the accessibility of website to other users.

1.1 Motivation

As Internet continues to gain market share from television, radio and print media as a primary news provider, the news websites are currently experiencing tremendous growth. One advantage of Internet news is that it allows more interactive behavior as compared to traditional media. Trustworthy news websites attempt to perform fair and accurate reporting. The goal is to report the news, without creating any unnecessary hype or pursuing any special agenda. However, this is a delicate balance that requires strong editorial control. The configurability and user-interactive feature of Internet news websites however makes this objective even harder.

To further complicate the matters, websites are also visited by software programs and scripts (popularly known as spiders, robots, or just "bots"). Thus, websites need an adequate mechanism to distinguish a human reader from a non-human reader.

Yet another motivation for this problem is that Statistics Hacking is very similar to Click Fraud. Advertisers need to be able to distinguish clicks by human users from clicks by non-human readers, and also be able to distinguish clicks by normal users from clicks by malicious users.

1.2 Scope and Structure

This article is based upon the vulnerability in the news websites, as it allows hackers to modify the statistics 1 of the system quite easily. As we will show briefly in this article, news websites (for example BBC) employ an extremely basic protection scheme against statistics hackers.
This paper is organized as follows. The current section introduces and motivates the problem. In Section 2, we define the statistics hacking and analyze the extent to which the statistics hacking is pervasive in the current news systems. In Section 3, we present in detail the vulnerability in BBC, a leading news website, and how the vulnerability can be exploited by statistics hacker. In Section 4, we discuss methods for protection against

Manuscript received February 5, 2007.
Manuscript revised March 25, 2007.

statistics hacking by distinguishing human users from computers, and present our analysis of these methods. In Section 5, we discuss a different category of methods for protection against statistics hacking. These methods involve adjusting counting principles to ignore the duplicate actions from malicious users. Finally, our conclusions presented in Section 6 complete the paper.

2. What is Statistics Hacking?

We define Statistics Hacking to be a process in which a malicious user manages to modify the system usage statistics. Statistics Hacking explicitly refers to the situation where the resource (website or system) is available to the malicious user for acceptable usage, but the user is able to modify the system usage statistics using some unacceptable methodology.

Difference from other forms of hacking: Statistics hacking is distinctly different from some other forms of hacking as it does not try to gain unauthorized access to the system or the equipment. It only affects the system usage, even acting in the same way a valid user would act. There is also the possibility that a malicious user gains an unauthorized access to the host system and manually modifies the statistics (by increasing the hit counts database tables or files). However, we do not consider that form of hacking as Statistics Hacking.

Relation to click fraud: Statistics Hacking is in practice quite similar to Click Fraud, which as a term used mostly in conjunction with search engines and pay per click advertising. The primary difference between Click Fraud and Statistics Hacking is the motive. Operationally, the acts are similar and much of the discussion in this paper also applies to Click Fraud.

Relation to link spamming: Statistics Hacking is also similar to the practice of Link Spamming, in which the motive is to obtain free advertising or spread a propaganda by posting the page onto free public boards or forums on the Internet.

Not necessarily programmer's hacking: It is also worth pointing out that Statistics Hacking may not even require any software to operate. For a low usage system, a user may simply be able to modify the statistics by using the program manually sufficient number of times. In this article, we focus on the form of statistics hacking that can work against BBC, that is, a very high usage system.

2.1 Scope of Statistics Hacking Problem

To understand the scope of statistics hacking, we considered the following leading news websites as a representative sample.
- BBC
- CNN
- MSNBC
- BusinesWeek
- ABC News

It was quite surprising to note that all these leading websites contained significant vulnerabilities to statistics hacking.

2.2 Scope of Problem at BBC

To better understand the scope of statistics hacking and how it affects a news website, we performed detailed analysis of the BBC website [1]. We observed the most popular and the most emailed news sections.

From these sections, we observed the news and attempted to separate them into genuine candidate and non-candidate news. To define non-candidate news, we used the following criteria, all of which must be met for the news to be defined as non-candidate. (i) News must be at least three days old, (ii) No subsequent follow up on BBC posted on that news, (iii) Not listed in the other popular section. Such news represent a very high chance of statistics hacking, as normal reader behavior would not elevate such news to high rank. Between November 1st and November 10th, 2006, we sampled and reviewed the website 50 times over a period of ten days. Our observations include the following:

1. On a total of 37 times, at least one of the news in the 'Most Emailed' section was a non-candidate news.
2. On each of the ten days, at least one of the 5 observations showed a non-candidate news. The worst observations for each of the 10 days are shown in Table 1.

Table 1: Measuring extent of Statistics Hacking at BBC: Data collected for 10 days and separated between candidate and non-candidate news

Day	Most Emailed		Most Read	
	Candidate	Not Candidate	Candidate	Not Candidate
1	4	1	5	0
2	3	2	5	0
3	4	1	5	0
4	4	1	5	0
5	4	1	4	1
6	4	1	5	0
7	4	1	5	0
8	4	1	5	0
9	3	2	5	0
10	4	1	4	1

3. Basic Vulnerability in the BBC's System

This section highlights in detail the vulnerability in the BBC's "Email this to a friend" system, and how it can be exploited by Statistics Hackers.

BBC does not employ any of the advanced methods that we present in Section 4. Instead, it only uses a small hash value, which is hard coded inside the HTML form. It is not clear if this hash is intended as a security mechanism at all. In either case, it has no impact on security.

3.1 Hacker's Code

In the scenario below, we assume that the Statistics Hacker is a dedicated health services professional, who wants to highlight the story "HIV home screening kit launched", a news story carried by BBC at http://news.bbc.co.uk/2/hi/health/6212467.stm.

Hacker begins by opening that page manually in a web browser, and then manually clicking on the "Email this to a friend" link. When the smaller window with email form opens, the hacker views the source of that page (using the browser's 'View Source' functionality). The source of the page reveals most of the information that the hacker requires to submit that form.

Hacker's code involves a relatively simple Java program, in which a URL connection is obtained to the URL of the "Email this to a friend" page. Using the hidden variables and their values obtained from the form source, hacker creates the content that is then written to the output stream of the URLConnection object. Detailed code exploiting this vulnerability is available in Table 2.

We considered a scenario of a relatively benign hacker. It is easy to observe that this technique can also be used to highlight stories that are of embarrassment to individuals and communities, or incite physical or emotional violence. Due to the nature of the news, many news items do need to be reported, and a Statistics Hacker can then use the BBC as a propaganda tool by constantly keeping a story on the top of list of most emailed stories. As has been well documented by FAIR (Fairness and Accuracy in Reporting), a non-profit media watchdog agency, sensationalism is one of the largest problems in the news today.

4. Protecting against Statistics Hacking by Distinguishing Humans and Computers

The first broad category of protection against statistics hacking consists of distinguishing human readers from computers (or software programs). As presented in the Section 3, a software program can be made to repeatedly use a feature. If the website is able to distinguish human and computer users, then this vulnerability can be contained.

Table 2: Hacker's code exploiting vulnerability in BBC's "E-mail this to a friend"

```
// URL of the "Email this news page"
URL url = new URL ("http://newsvote.bbc.co.uk/mpapps/pagetools/" +
"email/news.bbc.co.uk/2/hi/south asia/5404256.stm");
URLConnection urlConn = url.openConnection();
urlConn.setDoInput (true);
urlConn.setDoOutput (true);
urlConn.setUseCaches (false);
urlConn.setRequestProperty ("Content-Type",
"application/x-www-form-urlencoded");
DataOutputStream printout = new DataOutputStream
(urlConn.getOutputStream ());

// Prepares the content
String content = "submit=" + URLEncoder.encode ("send") + "&storyURL="
+ URLEncoder.encode("http://news.bbc.co.uk/2/hi/health/6212467.stm") +
"&summary=" + URLEncoder.enclode("The first home test which says " + "it
can reassure patients they are free of HIV is launched in the UK.") +
"&headline=" + URLEncoder.enclode ("HIV home screening kit launched") +
"&hash=" + URLEncoder.encode("JPpVtKp+mt2ce2S621RMSQ") +
"&emailsString=" + URLEncoder.encode ("fakeReceiver@email.com") +
"&fromName=" + URLEncoder.encode ("Fake Name") + "&fromEmail=" +
URLEncoder.encode ("fakeSender@email.com");

printout.writeBytes (content);
printout.flush ();
printout.close ();

// Gets response data.
DataInputStream input = new DataInputStream (urlConn.getInputStream
());
String str;
while (null != ((str = input.readLine()))) { }
```

Distinguishing between humans and computers is a well studied topic, and is related to the famous Turing Test problem[2, 3]. The application of Turing Test to this problem involves introducing an intermediate "Turing" test in the online user's action, with the assumption that it would be difficult for the software to respond accurately to the Turing test. The rationale is that such a test does not inconvenience casual users, who are likely to use that feature once, but is a significant deterrent for a user who attempts to perform the same action multiple times.

Note: In the classical Turing test, the entity presenting the test is a human user. However, in the current test, the entity presenting the test is the software system powering the website. For this reason, it is sometimes referred to as Computerized Turing Test. One particular system within this category is the CAPTCHA [4], which is a registered trademark.

There is a variety of Turing tests available, and they work with varying degrees of success against state of the art

computer programs. Next, we present an outline of various mechanisms and discuss their advantages and disadvantages.

4.1 Visual/Image Verification (Turing Numbers)

In this mechanism, an image is displayed that contains a text that is obscured by using image distortion techniques, such as highlights, background clutter, shadows and random line segments. This mechanism is widely deployed by websites that require registration. While specialized Optical Character Recognition (OCR) implementations can defeat this test, it does make the spamming software significantly harder. A recent implementation of image verification by Dutta et al [5] has been shown to protect against latest OCR algorithms.

One disadvantage of this approach is that it seriously impacts the web accessibility for the users who may be visually impaired.

An example of image verification is shown in Figure 1.

Figure 1: Example of image verification (Turing Numbers)

4.2 Audio Verification

In this mechanism, an audio is played by the website, and the human user is expected to answer a short question based on the audio. Again, the motivation is that this makes the spamming softare significantly more complex, as it must include a module to read an audio file, decipher the words, and then calculate the answer. Recent works using this technique include [6] and [7].

4.3 Mathematical Expressions

In this mechanism, the website presents a mathematical expression, and the human user is expected to calculate the answer to the mathematical expression. For example, the website's challenge question may be: "What is seventeen plus two plus thirteen minus twenty?"

This is considered a relatively weak form of Turing Test, even though in practice it may be safe due to "security by obscurity".

The primary disadvantage of this method is that a custom calculator can be built quite easily that accepts mathematical expressions and some elements of natural language processing.

Another significant disadvantage of this method is that it impacts the accessibility for users who have cognitive disabilities.

4.4 Logic Riddles or Puzzles

This mechanism is similar to the mathematical expressions, except that in place of a mathematical expression, a logic puzzle is presented to the website user. For example, a challenge question may be: "What is the name of the fruit that is its own color?" (Orange), or "What is the name of a yellow curved fruit?" (Banana).

This form of Turing test is very vulnerable to a dictionary attack, as the logic puzzles may be limited in number.

4.5 Video Verification

This mechanism is similar to audio expression, except that a video is shown to the website user. Following the video, a challenge question may be: "What did the man wearing yellow shirt talk about". The response can either be free form, or optionally 4 choices are presented to the user.

This form of Turing test suffers from a problem that some web browsers and operating systems may not show the video clip.

Clearly enough, image, audio and video verification methods pose problems for users using textual browsers.

4.6 Ascii Image Verification

This Turing test is similar to image verification, with the difference that it presents the image using Ascii characters, thereby continuing to be accessible to users using textual browsers.
An example of Ascii image verification is shown in Figure 2.
Security of the Ascii image verification technique has not been considered so far in the literature, though again, like in practice it may be safe due to "security by obscurity".

Figure 2: Example of Ascii image verification

4.7 Accessibility of Suggested Methods

The methods presented in Sections 4.1- 4.6 vary in the degree of success that they provide distinguishing humans from computers. They also vary in the degree of accessibility. We compare them on the following counts:

- Cultural/Intellectual Compatibility: The methods that use certain intellectual or cultural

information may not be accessible for all users. For example, consider the two challenge questions presented in Section 4.4. A person who is not a native English speaker may have difficulties answering the first question. Similarly, a person who is colorblind may have difficulties answering the second question. Audio verification may pose challenges to a person who is hearing impaired or is accustomed to a different accent.

- Browser Compatibility: The methods using non-textual mechanisms, that is, image verification, audio verification and video verification all present problems in terms of browser compatibility. Users with older browsers may not be able to use these features.
- Convenience: To compare the various mechanisms in terms of convenience, we apply the 10 second versus non 10 second rule. Essentially, a mechanism that takes more than 10 seconds for a human user to verify inconveniences the user significantly so as to affect the usability of the feature. As per this rule, both video verification and mathematical expressions can be considered to be unnecessarily tedious.

We present a summary of the accessibility of suggested methods in Table 3.

Table 3: Comparison of various methods for distinguishing human users and computers

Method	Cultural Comp.	Browser Comp.	Convenience
Image Verification	✓		✓
Audio Verification			
Mathematical Expression	✓	✓	✓
Logic Riddles or Puzzles		✓	✓
Video Verification			
Ascii Image Verification	✓	✓	✓

Due to the accessibility issues, many combination methods have been proposed as well. The idea is that by providing multiple methods, the accessibility of the system can be improved.

A few other methods for human verification have also been proposed, for example, [8], [9] and [10].

5. Protecting against Statistics Hacking using Clustering

In Section 4, we presented various methods that can counter statistics hacking by distinguishing humans and computers. In this section, we explore an entire different strategy. In this strategy, the software system does not try stop the malicious usage. Rather, it alters its counting methodology to ignore the duplicate actions from the malicious users.

Following methods of protection are available against Statistics Hacking, that we will explore individually.

5.1 Ignoring Multiples in Statistics based on IP Address and Time

This is a relatively easy method, in which the usage statistics module simply ignores multiple actions taken by the "same" user, where the same user is interpreted to be the same IP address and time unit. The time unit may be selected to be 1 hr, 1 day, or something else depending upon the exact system.

All similar actions performed from the same IP and within the same time unit (hour/day) are counted as one action. Again, to further diminish repeat counts, the last byte of the IP address may be ignored in this process.

5.2 Ignoring Multiples in Statistics based on Similarity in Action

In this method, a pattern is established on the underlying action, and its impact is assessed, irrespective of the time or the source of the action. For example, in case of BBC's "Email this to a Friend" feature, a pattern may be established on the sender's and/or receiver's email address. The business interpretation can be that each receiver's email address only counts once.

This mechanism can be easily defeated by using a random email address as sender and a random email address as receiver. Even if the system is able to ignore bounced or undelivered emails, the statistics hacker can use a domain catch all address (that is, all emails sent to a particular domain are delivered to one address). So, while this mechanism is not perfect, it is in principle different from Time and/or IP based grouping.

5.3 Ignoring Misusing IPs or Domains

Once an IP or a domain is found engage in Statistics Hacking, website administrator may ignore usage from that IP or domain when calculating statistics.

§

Notes:
(i) **Goal of Clustering**: None of the methods are foolproof, and many may ignore multiple valid uses. Still, as the statistical counts are usually not mission critical, it may be acceptable to under report the usage counts.
(ii) Blocking is not an option: Once an IP or a domain is found engage in Statistics Hacking, website administrator may block that IP or domain from further using the service. This method is often impractical as it allows malicious users to compromise the usability of shared public computers.

6. Conclusions and Future Work

In this paper, we highlighted a form of vulnerability which can lead to "Statistics Hacking". This form of hacking is of particular importance to news and other public websites that also want to allow users to interact and give feedback to the news. We also presented the other forms of hacking, similar or related to Statistics Hacking and highlighted the differences and similarities.

We reviewed leading news websites and presented a summary of the scope of the existing problem. We documented a method for exploiting this vulnerability in a known leading website that can lead to modification of the system statistics by malicious users (statistics hacking).

We outline two broad categories of methods that can counter statistics hacking. The first category consists of many methods already available to distinguish humans from computers. We compare the different methods within this category on basis of security and accessibility. The second category uses well known clustering techniques applied to the web statistics. We compare the different methods within this category on basis of security.

This work can be extended in two distinct ways. We outlined various methods for conducting Turing test, that is, distinguishing computers and humans in Section 4. That analysis can be extended to jointly consider security and accessibility features. In Section 5, we presented counting principles which can be used to limit the effects of Statistics Hacking, click fraud and link spamming. This work can be extended to present a quantitative analysis, which would be of interest in publishing and advertising algorithms.

References

[1] "BBC NewsWebsite," http://news.bbc.co.uk/, Retrieved between October 1st, 2006 and December 15th, 2006.

[2] A. Turing, "Computing machinery and intelligence," Mind, vol. LIX, no. 236, pp. 433–460, October 1950.

[3] S. G. Sterrett, "Nested algorithms and the 'original imitation game test'," Minds and Machines, 2002.

[4] "CAPTCHA Project," http://www.captcha.net, Retrieved on December 8th, 2006.

[5] R. Datta, J. Li, and J. Wang, "Imagination: A robust image-based captcha generation system," in Proceedings of the ACM Multimedia Conference, November 2005.

[6] M. T. Goodrich, M. Sirivianos, J. Solis, G. Tsudik, and E. Uzun, "Loud and clear: Human-verifiable authentication based on audio," in ICDCS 2006: 26th IEEE International Conference on Distributed Computing Systems, 2006.

[7] G. Kochanski, D. Lopresti, and C. Shih, "A reverse turing test using speech," in Proceedings of the International Conferences on Spoken Language Processing, 2002.

[8] J. McCune, A. Perrig, and M. K. Reiter, "Seeing is believing: Using camera phones for human verifiable authentication," in IEEE Symposium on Security and Privacy, 2005, pp. 110–124.

[9] Y. Y. Gu, Y. Zhang, and Y. T. Zhang, "A novel biometric approach in human verification by photoplethysmographic signals," in 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003, pp. 13–14.

[10] C. Park, J. Paik, T. Choi, S. Kim, Y. Kim, and J. Namkung, "Multi-modal human verification using face and speech," in IEEE International Conference on Computer Vision Systems, 2006.

Amrinder Arora received the B. Tech. degree in Computer Science and Engineering from the Indian Institute of Technology, Delhi, in 1998. He received MS and DSc degrees from the George Washington University, Washington DC in 2001 and 2006 respectively.

His research interest includes online algorithms, theory of computation and software engineering.