Evaluating Adversarial Interfaces

An Automated Approach

Ryan Tate, Gregory Conti, Alexander Farmer, and Edward Sobiesk

Contrary to the idealistic notion that human-computer interfaces are designed to actively help users accomplish their goals, millions of people who use the World Wide Web every day encounter a wide variety of interfaces that aggressively divert users toward the goals of the designer and away from those of the user. We label such strategies "adversarial interface design." Techniques of adversarial design include coercion, confusion, distraction, exploiting errors, forced work, interruption, manipulating navigation, obfuscation, restricting functionality, shock, and trickery (10). See Figure 1 for several illustrations of these techniques. A key distinction of adversarial design from bad design is the deliberate attempt by the designer to manipulate the user or subvert user intentions.

Rather than functioning as allies of the user, many webpage designers are adversaries with tremendous power. We acknowledge the frequent necessity of and associated business models employing advertisements on the Web, and that some ads are considered beneficial by some users. However, ads affect Web user experiences almost every time users go online – sometimes to very significant levels. The user community fights back through efforts such as Adblock Plus (ABP), an open-source ad blocker available at http://adblockplus.org. Millions of Adblock Plus users actively remove unwanted ads (noise) from webpages. This illustrates the severity of the user-designer struggle over ad placement within web interfaces.

Digital Object Identifier 10.1109/MTS.2016.2518249
Date of publication: 9 March 2016

Noise (i.e., unwanted ads) in a webpage frequently consists of ads from Content Distribution Networks (CDNs) that interfere with user-desired content. We define noise using Adblock Plus as a classifier and seek to measure the extent to which ads have invaded webpages to divert user goals toward those of the designer. A modified version of Shannon's Model of a general communications system (24) provides a conceptual framework of the situation (see Figure 2).

The ability to consistently compare and chart the impact of adversarial interface techniques as a scientific measurement will help raise awareness of the effects that adversarial content has on popularity and projected user experience for interface designers, advertisers, regulators, search engine companies, and users. A means of measuring such content would allow designers and content providers to increase awareness of the overall impact of ads on the user experience as well as compare strategies against their competitors. If the level of interference engendered by the adversarial interface techniques employed on a given page becomes high enough, users may seek alternate webpages to conduct their activities. It was shown that this threshold is different for each user and will vary based on the user experience and user-desired content of the site (10).

From a societal implications perspective, adversarial interface techniques degrade the value and trust of technology. Adversarial interface design is not limited to the Web and is common in desktop, mobile, and physical interfaces. In this Digital Age, there will always be friction between usability and commerce. However, adversarial interface techniques are an exceptionally egregious example of an environment that leads to a collective loss of time and productivity. Adversarial techniques, whether

FIGURE 1. Examples of adversarial interface techniques from the World Wide Web. (a) TheFreeDictionary.com result page is so cluttered with noise that finding the desired word definition (A) is difficult. (b) The Fox News website is interrupted by a Fidelity Investments advertisement placing the news story at an unreadable angle. (c) A survey participation request obscures a Network World article. (d) Wired Magazine's scrolling ad for magazine subscriptions (A) and a looping animated GIF (B). (e) A vodka bottle (A) flies over Weather.com while a man raises his glass in a toast (B). (f) An expanding ad from the National Geographic Channel covers a New Scientist article.

explicitly or inadvertently, also appear to significantly impact more vulnerable user populations such the young, the elderly, the physically challenged, etc., who may be especially susceptible to their deceptive and confusing techniques. Our cataloging of adversarial content in popular websites provides insight into the extent to which advertising and other adversarial interface techniques exceed user tolerance, and provides a comparative ranking of website behaviors. Such rankings may be further divided by content category to provide increased clarity, such as news, weather, sports, porn, and gaming sites, as well as between sites.

In this article, we automate original metrics and provide a study measuring adversarial content in over

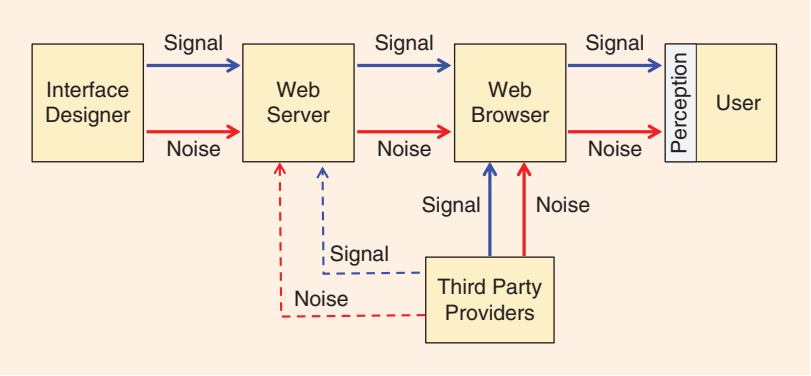

FIGURE 2. Modified Shannon Model. Users perceive the message, a webpage including signal (desired content) and noise (ads), through a web browser which obtains the complete message via multiple channels including third party channels (e.g., CDNs) as determined by the interface designer.

200 top Alexa global websites. Our metrics are not an attempt to measure good website design, provide new usability metrics, or judge the use of advertising to support free online products and services. Our approach is unique in providing metrics for measuring potentially adversarial content, and strictly calculates the level of content embedded in a page from advertisements, certain adversarial interface techniques, and other noise.

In the remainder of the article, we summarize adversarial interface design techniques, explain our metrics and supporting equations, describe how our metrics calculate webpage adverseness ratings, discuss the design of our automated tool which is implemented as a browser extension, and explain our website survey methodology. We then present and analyze the results of our survey, suggest promising directions for future work, place our research in the field of related work, and conclude.

Adversarial Interface Design

As discussed at the beginning of this article, the categories of adversarial interface techniques are many and diverse. However, our desire to automate analysis and the necessity to appropriately scope the problem has resulted in this work focusing on three specific adversarial categories: distraction, forced work, and interruption by ads. These three adversarial design techniques are illustrated in Figure 1. The designer of a website can potentially create an adversarial environment or implicitly influence the user experience through choice of web advertising companies and related third-party content at varying levels of aggressiveness. As a result, the designer possesses great power in crafting the user experience.

In our work, we assume a user's primary task is not to find advertising, and that users therefore do not want advertising, whether untargeted or targeted. This assumption is supported by recent advertising click-through rates commonly reported at around 0.09%.

Metrics

Our primary research objective is to create metrics to measure the visual adverseness of a webpage. We do not address non-visual attributes and, while we evaluate webpages, the metrics are also applicable to desktop and mobile software. Our implementation provides a single value useful to compare websites and analyze for various trends. The single value method is based on the principles of Jaquith (16):

> "The goal of metrics is to quantify data to facilitate insight. A good metric should be consistently measured, cheap to gather, expressed as a cardinal number or percentage, expressed using at least one unit of measure, and ideally, contextually specific."

We analyze our success against these criteria later in the paper.

Static Analysis

A webpage displayed in a browser consists of pixels, each of which we categorize into either content (C) or whitespace (W). The union of C and W equates to the set of all pixels on the page. We subcategorize C, defining user-desired content as signal (S) and undesired content (i.e., ads) as noise (N). On some webpages $N >> S$ (the number of noise pixels is far greater than the number of signal pixels), leading to user annoyance and frustrating task accomplishment. A logical means of measuring C, S, N, and W is by the number of pixels dedicated to each. In practice, the S and N pixels are clustered into regions based on HTML code. Therefore, a displayed webpage is the combination of whitespace and various signal and noise regions; see Figure 3.

A browser determines the positioning and dimensions of content regions as it interprets HTML code, which is written as a tree of element tags. Browsers render these elements (e.g., <img> for images) on the screen based on default or specified markup and/or CSS rules prescribing element presentation (such as height and width). We categorize the pixels in the region rendering an element on the screen as either signal (S) or noise (N), defining advertising elements, whether targeted or not, as N. Due to advertising display standards, many N regions will occur in standardized sizes such as 160x600 pixels (Wide Skyscraper) and 300x250 (Pop-Up) (15). However, N regions could occur in any shape,

such as the flying vodka bottle illustrated in Figure 1(e), and may overlap, a situation addressed below.

Intuition Behind the Metric

Starting with basic intuition, we derive a metric counting HTML element pixels, while weighting both adversarial content and whitespace. The intuition tallies the number of visible pixels associated with noise N and signal S to derive an adverseness rating (A) for a given page in the range from 0 to 1. A basic adverseness equation is formally:

$$A = \frac{N}{C} = \frac{N}{N+S}. \tag{1.0}$$

Absent trickery, N and S regions are easy for humans to detect. Moderate to high values of A indicate the application of adversarial interface design. However, the classification of N is subjective, which we address in the section "Noise Classification." As our goal is automated metric calculation, determining N and S must be precise and consistent.

To more formally define N and S, consider that each region of a webpage consists of a group of pixels associated with an HTML element. We define N_i as the set of pixels associated with a noise HTML element i on a webpage (such as a Flash advertisement). We similarly define S_j as the set of pixels associated with a signal HTML element j (such as an image). We therefore derive a webpage's N and S from the cardinality of the union of all N_i and S_j. To illustrate this, if all content is signal, then A is 0, but A is 0.5 with equal signal and noise. Eq. (1.1) evolves eq. (1.0) to calculate an adverseness rating as the size of the set of all noise pixels divided by the size of the set of all content pixels:

$$A = \frac{\left|\bigcup_i N_i\right|}{\left|\bigcup_i N_i\right| + \left|\bigcup_j S_j\right|}. \tag{1.1}$$

This simple equation provides a basis for more realistic approaches. It is too simple, for one, because it assumes that all regions of a page have an equal impact on human perception. For example, simple black text on a white background may occupy much more space than a lurid advertisement, but draw far less attention. Many advertisements seek to exploit pre-attentive processing, such as size, shape, color, intensity, or orientation to attract user attention. A better approach would be to weight regions based on perceptual impact.

Element Weighting

We describe strategies for deriving element weighting factors in conjunction with eq. (1.1) to provide a framework for more accurate measurements of adversarial content. Many are best informed and validated by future human subject testing and combinations of weighting factors may yield the best results. Generally, each N_i and S_j is weighted based on desired element-wise weighting, w_i and w_j.

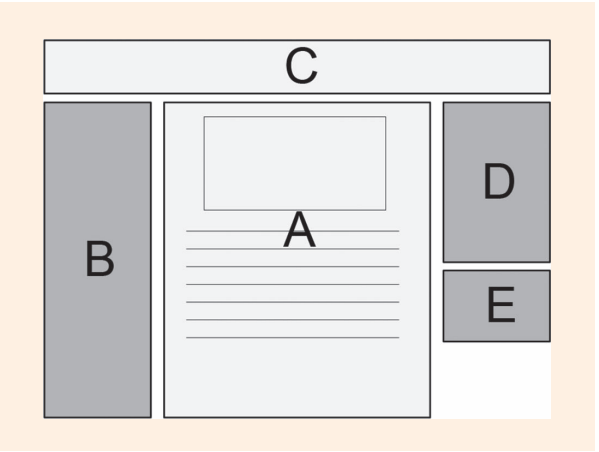

FIGURE 3. A generic webpage with content regions A-E and whitespace composed of the remaining pixels. Content can be subcategorized as signal (light shaded regions A, C) or noise (dark shaded regions B, D, E).

Stack Order Weighting

To account for element overlap, we order elements according to increasing stack order where higher indexed elements appear on top. Equation (1.2) derives eq. (1.1) to account for the weighting of individual HTML elements and for overlap by counting only pixels not occluded.

$$A = \frac{\sum_i w_i \left| N_i - \bigcup_{k>i} N_k \right|}{\sum_i w_i \left| N_i - \bigcup_{k>i} N_k \right| + \sum_j w_j \left| S_j - \bigcup_{l>j} S_l \right|} \tag{1.2}$$

for all $i, j, k, l \in C$. Equation (1.2) provides the ability to weight elements, for example by perceptual impact, and is amenable to automated calculation with the simplest implementation setting each w_i and w_j to 1.

Whitespace Weighting

We assume the more whitespace on a page, the more adversarial noise is diluted – that whitespace makes noise more obvious (though one could reverse this assumption with algebraic manipulation). Based on our assumption, a single noise image on a largely blank webpage would result in a lower rating than the same image on a busy webpage. Defining whitespace, W, as the number of all pixels not in N or S on a webpage, eq. (2.0) produces a whitespace weighted adverseness rating (A_W):

$$A_W = A\left(1 - \frac{W}{W+N+S}\right). \tag{2.0}$$

Above, A_W ranges from 0 to 1 and increases with N. As whitespace increases, A_W decreases, illustrating the relative dilution of A by increasing whitespace.

Bucketization

This manual approach assigns w_i and w_j based on user perception and categorization when applying a well-defined adverseness rubric. Weighting may occur from 0 (no perceptual impact) to 1 (severe perceptual impact). This captures nuances easily perceived by humans that might be missed by an automated metric, such as an attractive person's picture. A carefully designed rubric with many user evaluations may mitigate subjectivity.

Luminance and Contrast Ratio

Hue, brightness, and contrast play a role in the impact of noise. We suggest study of the W3C recommended formulas for measuring brightness and color difference [29]. Additionally, the Web Content Accessibility Guidelines v1.0 and v2.0 deal with issues such as relative luminance and contrast ratio formulas and take into account human perception [28], [29]. Relative luminance and contrast ratio may bear the greatest promise.

Bytes per Region

Lending well to automation, this approach measures bytes allocated to N, S, and W regions under the significant assumption that more N bytes indicates more aggressive content. For example, an MPEG movie would be larger and more perceptually aggressive than a similarly dimensioned image. The bytes assumption could be hindered by issues such as bad design or small frame size, but would work well for Flash or other animated noise.

Proximity

The distance between S, N, and W regions may be significant. The close proximity of noise to important S elements, such as search results, may increase adverseness. A proximity-based weighting algorithm might involve pair-wise analysis between each S and N element pair, creating a two dimensional matrix of values used to derive appropriate weights.

Centrality

Numerous web-based eye tracking studies illustrate regions where users are more likely to focus their attention. Advertisers seek to place their advertisements in these regions. A centrality-based weighting factor would more heavily weight regions, whether S or N, that fall in these central locations.

Dynamic Weighting

Many webpages are dynamic and interactive. While dynamic analysis significantly complicates automated adverseness measurement, any metric that assumes webpage content is static is over-simplified. One could weight more heavily dynamic elements based on the magnitude of their movement, or combine a series of adverseness ratings (for example, eq. (1.2)) over time, perhaps every millisecond.

Noise Classification

One of the great challenges of measuring the adverseness of a webpage is that adversarial (noise) content is by nature subjective. What one user considers to be distracting, another may consider helpful or even desirable. Furthermore, noise is culturally dependent. To reasonably classify noise requires the opinion of an entire community of users. For that, we exploit the community-driven work of *Adblock Plus*. *Adblock Plus* classifies content on popular webpages as undesirable and removes HTML elements entirely through a continually updated element selector list. Noise classification by *Adblock Plus* exploits the user community's efforts to define and identify unwanted ads.

Browser Extension Design and Implementation

We automated a tool that measures webpage adverseness based on eqs. (1.2) - stack order weighting - and (2.0) - whitespace weighting. We chose these equations because they 1) are non-subjective and do not require user testing and 2) have a straightforward implementation that current technologies support. The tool is an extension of Mozilla Firefox version 18.0.2. Our objective was to automate calculation to the greatest extent while maintaining the option of classifying S or N elements manually after initial automated processing.

Automated Classification of N, S, and W

Our tool uses the browser-based W3C Document Object Model API to accurately measure content pixels (noise and signal). Figure 4 depicts the appearance of a generic webpage where various HTML elements are positioned according to CSS style rules. The tool defines an element's pixels as content if they fall within a bounding rectangle that separates an element from its padding, border, and margin pixels (based on the CSS box model), which count as whitespace. Towards the center of Figure 5, the image's padding, border, and margin are depicted along with the image itself.

The sum of the area (i.e., pixels displayed in a browser) of all noise and signal regions, as defined above, after a page loads, provides the measurement of webpage content. Figure 6 illustrates how the tool classifies content from the same page depicted in Figures 4 and 5. Figure 7 depicts the final geometric regions for Figure 4's generic webpage. The area of the red rectangles is noise, and the area of the blue rectangles is signal. All other pixels are whitespace.

In defining content, we included the following HTML elements: image, video, audio (control), object, and embed. Each is easily captured with a minimum bounding rectangle. CSS background images are not treated as content as it would be difficult to account for the numerous overlap situations. All measurement techniques discussed are recursively applied to iframe documents and are included in the respective signal, noise, and whitespace totals. Significant additional implementation details are described in the original work we extended from [26].

Automated Noise Classification

To perform automated noise classification, we modified the code-base of *Adblock Plus*, first described in the "Noise Classification" section above. Millions of ABP users subscribe to filter lists developed by the user community to identify ad content that is blocked through a set of CSS selectors. ABP "blocks" elements of a webpage by either preventing the browser from requesting an external resource (e.g., flash) via HTTP, or by rendering an HTML element within a document not visible. We adjusted ABP 2.2.1 (modifying Policy.processNode and querying elemhide.css) to annotate rather than block ads by inserting an arbitrary class attribute to the tag of element and its children within the live document. The arbitrary class attribute facilitates extension to extension communication (modified ABP to our tool).

During our experiments, we manually verified that our personal, subjective identifications of noise matched almost perfectly the automated identification produced by ABP. Using the ABP code-base also increases the maintainability of our tool as community standards change. Our tool additionally provides the ability to manually modify classifications if desired.

Calculating Results

Once classification is complete, final calculations based on eq. (1.2) and (2.0) essentially sum pixels in the set of rectangular areas as shown in Figure 7. Our tool draws rectangles to assist in manually validating identification results: red rectangles surround noise elements and blue rectangles surround signal elements. Figure 8 shows a screenshot of our tool (a Firefox browser extension) after completing an automated run on a webpage.

Evaluation Description

In this section, we describe the data, settings, and conditions of our experiment. The goal in designing the experiment was to calculate our metric results in a consistent manner that allows comparison between websites and reproduction of our work.

We used a web browser window dimension of 1024x768 pixels, resulting in a constant webpage display region of 1010x627 – approximating the display size of a 15-inch laptop. This captures what a user sees without scrolling, but the parameters of the tool easily adjust to capture an entire page including portions not initially visible. We used Firefox (version 18.0.2) running on a PC with Windows 7.

We ran our experiment on the first 660 of Alexa Internet's top 1000 global websites (as of September 2013). To achieve greater fidelity, we more closely examined 25 top U.S. sites, looking for sites designed for use by general web visitors, not a closed or targeted group as this might skew our evaluation. Toward this end, we excluded sites requiring user accounts, such as Bank of America and Windows Live, and directory sites like Blog Spot. For most sites we examined the default homepage as well as search page results for "cars" in order to evaluate a typical use of some websites. These decisions are reflected in our "Results" section.

In early research, we anticipated that manual classification of signal and content would be necessary. We envisioned a desktop application facilitating manual annotations on screenshots that would then be processed. This concept evolved into a browser extension because we could leverage the Firefox document API to perform automated calculations of elements on the page.

A key breakthrough came with the integration of our modified version of ABP as the classifier, which

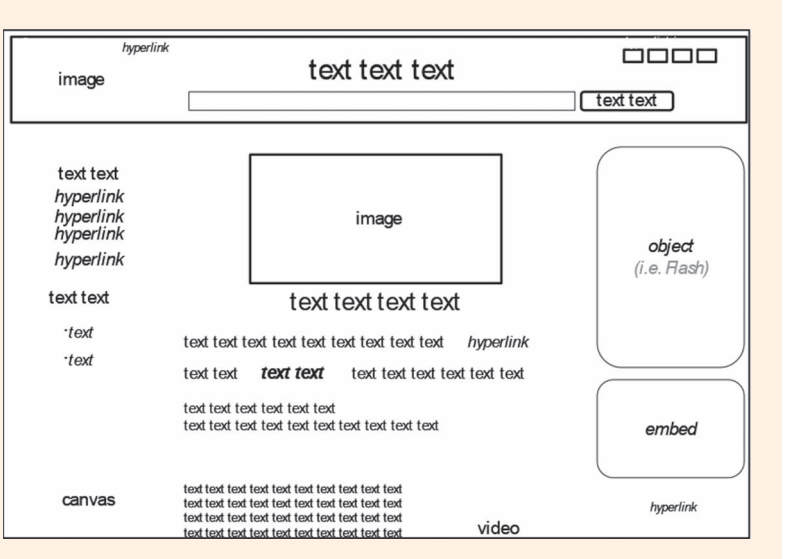

FIGURE 4. The appearance of a generic webpage displayed in a browser. The various HTML elements are positioned according to CSS, style rules, or other methods.

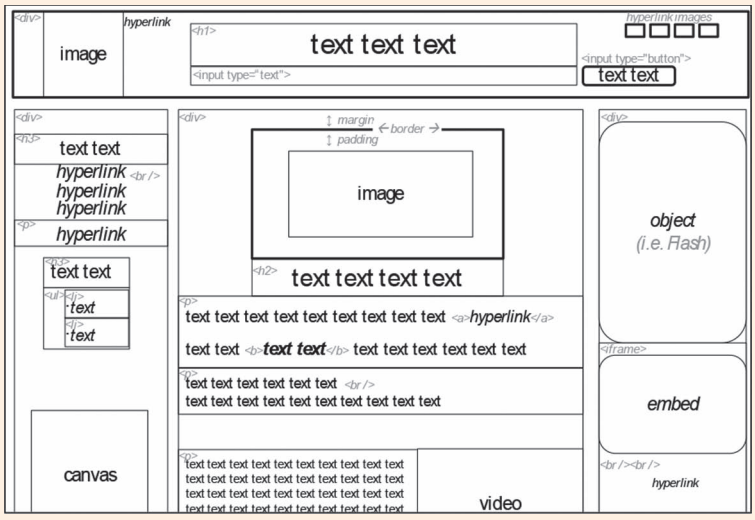

FIGURE 5. The generic webpage of Fig. 4 with select markup of HTML element tag names and a border drawn around many elements to indicate boundaries. The padding, border, and margin of the center image element are annotated.

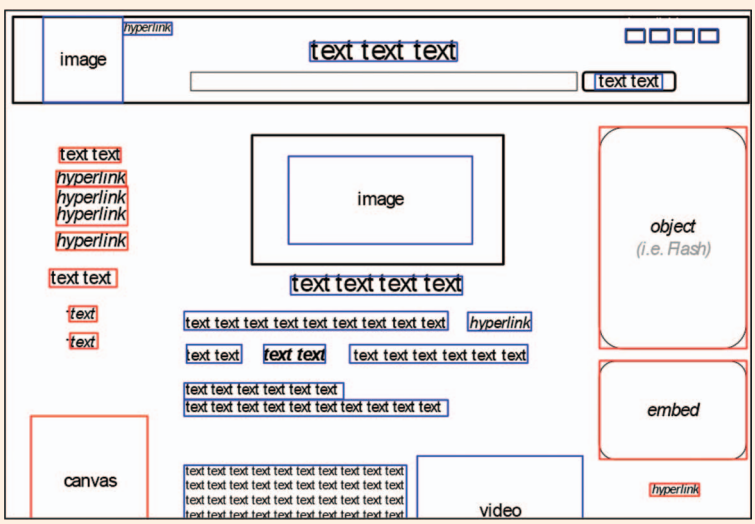

FIGURE 6. This depicts noise (red) and signal (blue) regions from the generic webpage of Fig. 4 and 5. Signal and noise regions are based on bounding rectangles excluding padding, border, and margin.

facilitated fully automated noise classification. We used the default installation "EasyList" ABP filter list (17) because it is both well-regarded and community developed (https://easylist.adblockplus.org/en/), though the tool accepts any ABP filter list. We discuss the strong performance of ABP as a classifier in the Analysis section.

For each website, our tool classified and annotated all noise and signal regions. It then saved 1) a screenshot of the visible portion of the annotated webpage, 2) a text file with all calculations, and 3) a copy of the complete webpage in the Mozilla Archive Format (MAFF). As we conducted the webpage evaluations, we manually examined each of the 25 U.S. sites after automated classification to confirm accuracy of the red (noise) and blue (signal) annotations generated by our tool. The tool allows for manually adjusting an incorrect classification, but this proved unnecessary.

We visited each site once. Our experiment did not consider pop-up or pop-under windows in our calculations, although these frequently contain pure noise. We did not employ any additional plug-ins beyond our tool and those required for correct rendering of the display, e.g., Flash.

Evaluation Results

In this section, we present the results of the experiment described above in the "Evaluation Description" section. Also as discussed earlier, these results were completely automated but manually verified for the top 25 U.S. sites by one of the authors.

We tested the first 660 of Alexa's top 1000 global websites and our automated tool successfully measured adverseness in 598 cases. The error cases demonstrate that the automated tool requires future work to account for the diversity of the web. Of the 598, ABP identified noise for 212 websites. Many of the foreign websites had no applicable ABP filter rules and, generally, ABP rules were applicable for the most popular sites. Figure 9 shows the Cumulative Distribution of adverseness ratings based on eq. (1.2) and (2.0) - whitespace weighting - for those 212 global websites where ABP had labeled noise (ads).

We evaluated the following 25 U.S. websites in greater detail: About.com, Amazon, AOL, Ask, Bing, CNN, Comcast, Craigslist, eBay, ESPN, Facebook, Fox News, Google, Huffington Post, IMDB, Microsoft, NY Times, Pinterest, Tumblr, Twitter, Walmart, Weather Channel, Wikipedia, Yahoo, and You Tube. Appendix 1 shows a detailed summary of our tool's results for those 25 websites. The websites in Figures 10, 11, and 12 are sorted from left to right in decreasing order of the adverseness

rating determined by eq. (1.2) using the "cars" search result pages.

Figure 10 shows the adverseness rating determined by our tool using eq. (1.2) and (2.0) –whitespace - on both the homepage and on a search for "cars" as explained in Section 5. Eight sites not listed in Figure 10 (Craigslist, eBay, Facebook, Microsoft, Pinterest, Tumblr, Twitter, Walmart, Wikipedia) had adversarial ratings of zero. We address zero ratings in the next section.

We depict the count of signal (blue), noise (red), and whitespace (gray) pixels for each website in Figure 11 ("cars" search) and Figure 12 (homepages). As explained in the Evaluation Description section, the display dimension of 1010x627 pixels resulted in a maximum of 633 270 pixels.

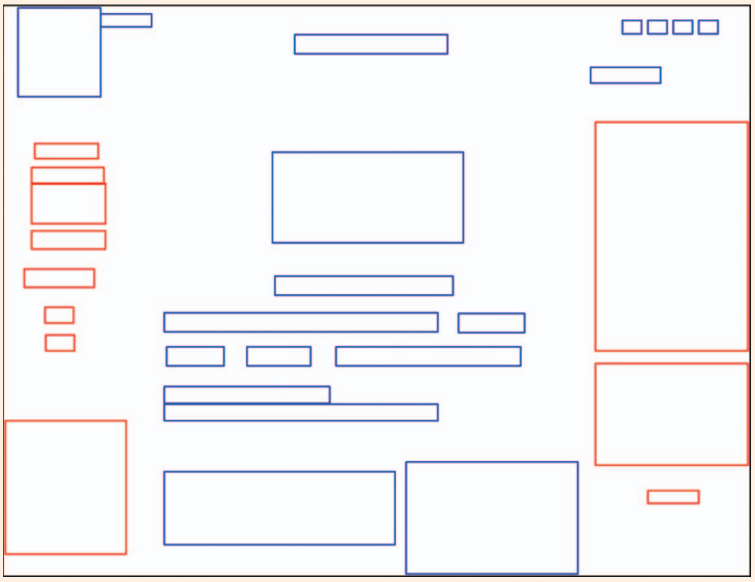

FIGURE 7. Depicts the final geometric situation for Fig. 4's generic webpage. The area of the red and blue rectangles are noise and signal, respectively. All other area is whitespace.

Results Analysis

The major contributions of this research are:

1) demonstrating a successful and efficient method for automating computation of an adverseness rating for a webpage,
2) creating effective metrics that allow for comparing the visual adverseness of websites,
3) providing insights into the relative prevalence of adversarial interfaces in popular websites,
4) identifying strengths and weaknesses of our techniques, and
5) proposing promising areas for future work.

In the subsequent subsections, we elaborate on each of these topics.

Automated Computation of Adverseness

We were pleased to find that our techniques and metrics in conjunction with the modified code base of ABP provided an extremely effective, automated method of identifying, labeling, and rating adverseness on webpages. Of the 25 closely evaluated websites, we did not find a single missed ABP classification. This strong performance of our tool suggests that techniques such as these could be employed on a larger scale. Ideally, we envision a future that includes fully automated analysis of a large swath of the Web, with adverseness ratings impacting both search engine results and user click-through choices.

Metrics Facilitating Webpage Comparison

One of the most valuable aspects of our results is the ability to potentially compare websites against each other in terms of adverseness. We demonstrate this in Figures 9-12, in which clearly some websites are more adversarial than others when using eq. (1.2).

Our metric is effective in appropriately rating webpages that use aggressive adversarial techniques. The adverseness rankings of websites in our experiment were strongly influenced by larger ads. On search pages, this often correlated with large advertised link sections preceding the actual search engine results. This is well illustrated in Figure 8 where the desired content (the actual search results) barely makes the viewable page area. On many search pages, the advertised links were difficult to distinguish from the actual search results. Our restriction, of evaluating only the top 627 pixel rows displayed in the browser, heavily rated webpages such as search results that exhibit advertisements before desired results. This restriction, the deliberate omission of ads (e.g., Wikipedia), and/or the withholding of ads until login resulted in eight sites with zero noise for our closely inspected dataset. In doing so, we assess that those eight sites achieve a more favorable initial impression in users. We can contrast that impression with the common technique of highly adversarial pages, which place large Flash objects or images across the top of a page or alongside primary content, thus dominating a user's initial impression.

Prevalence of Web-Based Adverseness

We expected that most users experienced adversarial content as prevalent on the Web. Several examples in Figure 1

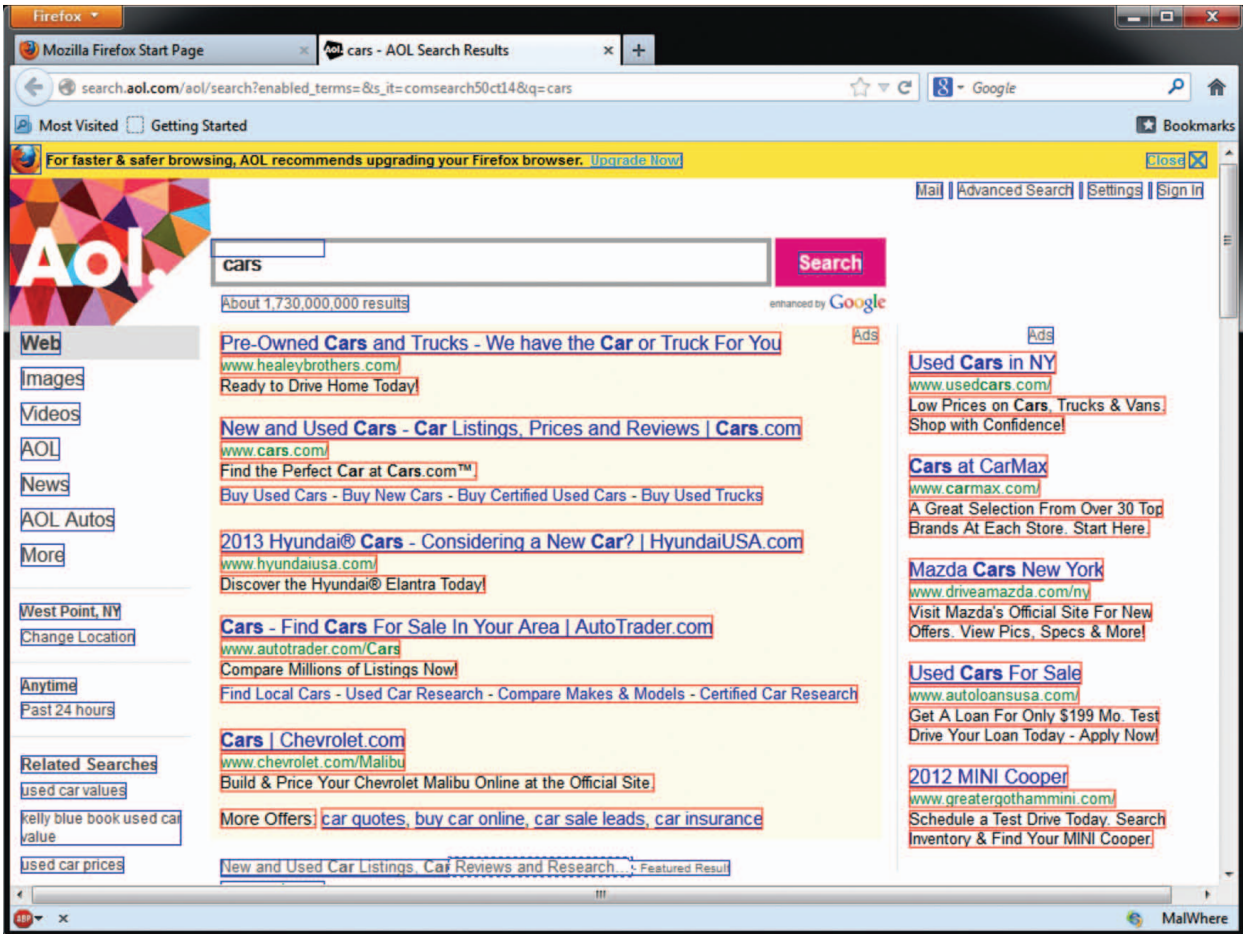

FIGURE 8. A screenshot of our tool (implemented as a Firefox browser extension) after completing an automated run on an AOL webpage. The tool identifies noise regions with red rectangles and signal regions with blue rectangles. The calculated value for eq. (1.2) - adverseness without regard for whitespace - is 0.76 and the calculated value for eq. (2.0) - adverseness with whitespace weighting included - is 0.22.

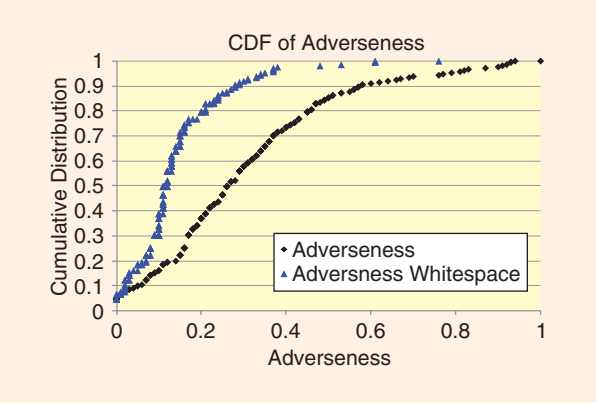

FIGURE 9. This graph depicts the Cumulative Distribution of adverseness rating determined by our tool using eq. (1.2) - black diamonds (bottom line) - and (2.0) - blue triangles (top line).

likely resonated with our readers. Our results, though, go beyond intuition and provide a repeatable, quantifiable, and feasible method of calculating webpage adverseness.

Our results show that adversarial design is prevalent to a moderate degree on the web. While only 35% of tested websites had content ABP blocked, the average adverseness of the 212 sites with ABP noise was 0.31. This suggests as much as one third of the content of popular websites monitored by the ABP community is advertising. This also explains why ABP is popular! Figure 9 indicates that 60% of evaluated websites have dedicated at least one quarter of their content to ads. Further, it indicates that approximately 40% of websites employing ads exceed the adverseness ratings of Google and Yahoo (the two highest Alexa websites). The effect of whitespace weighting greatly mitigates this effect (making ads more obvious).

Of the 25 U.S. sites we examined, about half of the search pages and many of the homepages possessed significant amounts of adversarial content (above 25% using eq. (1.2)). This supports the hypothesis that the amount of noise tolerated by users is a function of the quality of content and the availability of less

adversarial sites as an alternative (10). On the other hand, the Pearson correlation coefficient of Alexa ranking to adverseness rating for the 212 websites was 0.17, slightly correlating low Alexa rank to low adverseness rating.

To compare like-purposed sites, we analyze Figure 10 more closely. The adverseness results from the "cars" search indicate that top rated search engines have significantly less adverseness than competitors. Alexa's top four search engines (Google, Yahoo, Bing, Ask) had significantly lower adverseness ratings, between 0.29 and 0.43, than its over-50 rated search engines (AOL, About, and Comcast) with adverseness between 0.51 and 0.79. We leave a larger study of this potential insight to future work.

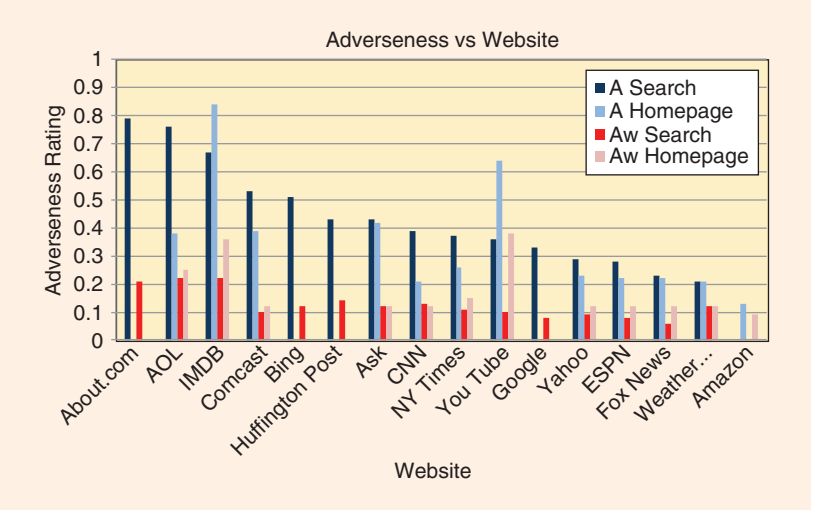

FIGURE 10. This graph depicts the adverseness (A) rating determined by our tool using eqs. (1.2) - A - and (2.0) - Aw. For each site, each equation was calculated on both the homepage and on a search for "cars" as explained in the "Evaluation Description" section. Eight sites, not shown, had no adversarial content.

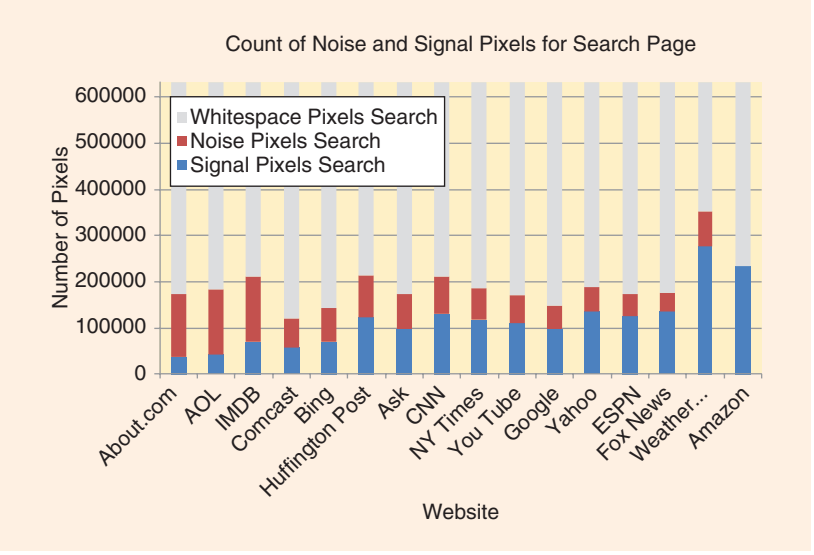

FIGURE 11. This graph depicts the raw signal (blue/bottom), noise (red/center), and whitespace (gray/top) pixel counts for each website from Figure 10 based on the search for "cars."

Strengths and Weaknesses

We believe our work has accomplished our primary goals of both defining reasonable metrics for rating adverseness and for demonstrating that such metrics can be automated. One of the greatest challenges is the customization of the Web and its impact on the ability to automate and consistently rate adverseness. Wikipedia, Craigslist, Pinterest, Microsoft, Twitter, eBay, Facebook, and Tumblr were rated at zero adverseness. As we review these results, we agree with some and question others. In the case of Wikipedia, we believe the site genuinely has no adversarial content. It is a strength that our tool accurately reflects this. The design of our experiment failed, however, in the case of Facebook advertising to logged-on users because we decided to not log on and only evaluated the homepage. Outside of our experiment, we tested the tool while logged on to Facebook, and it successfully classified sponsored ads as noise, but items like recommended pages as signal. This type of website challenges a fully automated and consistent method for rating adverseness because 1) generic web crawlers do not log on to websites and 2) the nature of targeted advertising might result in different levels of adverseness for different individual users. As the Web becomes more tailored to profile, geography, platform, and other characteristics, this challenge may increase.

A second significant challenge is that, in practice, websites can be a complicated mess (e.g., pictures nested in iframes inside a table paired with animation). Although our tool accounts for most static aspects of this messiness, we envision advertising techniques our tool would miss. For example, extensive use of canvas elements or background images such as the large AOL image in the upper left corner of the webpage in Figure 8. Additionally, our tool only currently supports rectangular regions, which restricts whitespace calculation. We believe this is

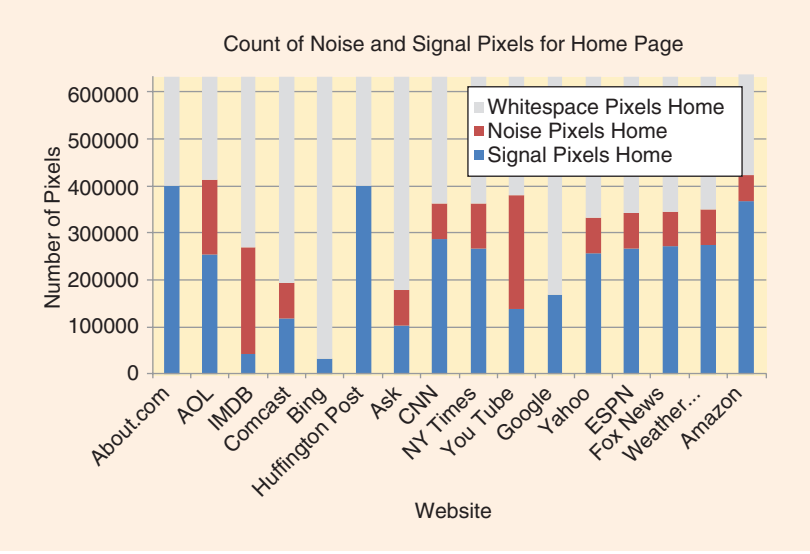

FIGURE 12. This graph depicts the raw signal (blue/bottom), noise (red/center), and whitespace (gray/top) pixel counts for the homepage of websites in Figure 10.

reasonable to facilitate automated analysis, but higher accuracy requires more robust techniques. Further, we might reconsider some portion of the area around ads, rather than classifying all of it as whitespace; strategies to highlight ads should impact adverseness ratings.

A final significant challenge is that our techniques measure the adverseness of only the static visual aspects of a webpage. Our tool does not address the effects of user interaction, dynamic (scripted) webpage elements, animation, or sound. We examined instantaneous metrics at page load, but change over time is particularly important to analyze with the impact of blinking objects, animations, and video, which may change S, N, and W and associated weighting factors dramatically. Several examples in Figure 1 illustrate dynamic effects for which we believe most users would argue for higher adverseness ratings than our tool produces. Interactivity is also key. Some adversarial interface techniques are only triggered through user interaction, such as "mouseover" expanding ads. Other adversarial techniques occur only once or for a specific duration, such as at the beginning of a video. We might also measure workload forcing users to obtain desired content by, for example, closing pop-up ads or through navigation over several pages (a technique to increase ad exposure). We could measure pop-up and pop-under advertisements using our metrics, but other metrics covering workload, time, or cognitive load may be necessary supplements.

Promising Areas for Future Work

We see efforts to combat adversarial techniques as a promising and emerging area of future research. Although the opportunities are many and broad, we suggest that the next logical steps to our efforts are to incorporate change over time into our metrics and to conduct a formal user study. We envision work that calculates metrics throughout an entire site visit to better determine the impact of adversarial interface techniques on the user experience. A user study could explore whether we can measure "unfair," "deceptive," or "misleading" adversarial techniques. The user study could also establish reasonable weights for the more subjective metrics described in the "Metrics" section, and investigate whether there are categories of ads that would not be considered noise. Looking at thousands of websites and evaluating them numerous times and under various conditions and geographical areas will bring this area of research closer to application in domains such as search engines. Web designers may use this information to reduce the adverseness of their sites and improve user experiences. Finally, experiments should be conducted with logged-on users.

Related Work

There is important related work surrounding the topic of interface metrics for adversarial interface design. This paper extends prior work on malicious interfaces [8], [10], [11]. This work defines methods used to subtly or aggressively influence users to take actions desired by the web designer, especially if contrary to user intent. We note, however, that prior work on interfaces did not include a means of measuring the impact that adversarial techniques have on users, which we extend in this article based on the original technique in [26].

Significant work has been done studying attention given to website advertisements. Using eye-tracking software, researchers found that users typically spend very little time glancing at advertisements, usually less than a second [4], [18], [19]. The effect of users becoming accustomed to seeing advertisements on websites is termed banner-blindness. Advertising and the struggle for user attention is covered in many documents from the advertising community. We suggest the work of Ogilvy as a starting point and *Advertising Age* magazine for more modern treatment [21]. However, we have found that many emerging adversarial advertising techniques are considered closely guarded information by advertising agencies due to perceived competitive advantage.

Adversarial interface design also threatens to violate usability best practices, regulations, and laws including the World Wide Web Consortium's (W3C) Web Content Accessibility Guidelines (WCAG) and Section 508 Amendment of the Rehabilitation Act of 1973 [23], [30]. More recently, legal scholars have explored the contractual implications of web design, consumer experience with a product or service as notice, and consumer opinion on behavioral targeting by marketers [6], [14], [27].

Incentives driving interface designers have yielded an additional important body of work. Goldfarb and Tucker's research studying online display advertising and the effectiveness of "obtrusive" and "highly visible" advertising found that obtrusive advertising and advertisements that match website content are more effective independently but less effective in combination. They also estimate that a move from targeted advertising could cause a 65% drop in advertising effectiveness and that advertisers may move to "more visually arresting" (read adversarial, or aggressive) advertising in order to compensate for their inability to target users [12]. Financial incentives are a key driving factor behind adversarial interface design. Particularly important is Acquisti *et al.*'s work on behavioral economics that suggest aggressive advertising online, particularly advertisements that interrupt the user's task flow, will decrease financial benefits to advertisers [1].

We do not seek to replicate traditional usability and user experience evaluation techniques, which employ user evaluation to test the effectiveness of interfaces designed to assist users. We instead study situations where the designer is an adversary. Our ultimate objective is automated and scalable measurement solutions. To this end, we found inspiration in Harty's work developing automated tests to find tab order usability flaws in websites [13]. Various readability tests including Flesch-Kincaid, Gunning fog index, and Dale-Chall, demonstrate that complexity of textual data can be measured effectively using automation. We also commend Norman's design rules based on his analyses of human error as providing useful insight, but this work assumes the designer is an ally not an adversary [20]. In addition, Fogg's "captology" and related work on persuasive technology demonstrates the extensive capabilities of technology to influence human behavior. Finally, recent work by Brignull defines and characterizes "dark patterns" – design patterns that are not mistakes, but instead trick people based on an understanding of human psychology. In the words of Brignull, dark patterns "do not have the user's interests in mind" [5].

The discipline of media literacy also intersects with the concept of adversarial interface design as it studies the ability to analyze and evaluate digital media. For an excellent starting point we suggest the work of Renee Hobbs and the work of Gus Andrews' *The Media Show*, which included an episode on adversarial interface design [3].

Chen and Janicke demonstrated that information theory can provide quality metrics for information visualization systems [7]. Suo *et al.* measure computer security visualization system design complexity in terms of "separable dimensions, complexity of visual attribute interpretation, and the visual search efficiency" [25]. Alexander and Smith provided a taxonomy of disinformation as a step toward "disinformation theory," and importantly do not assume cooperation between the sender of information and the receiver in their modified Shannon model [2]. Finally, Conti studied techniques for attacking information visualization system usability and demonstrated that an adversary with access to the data being visualized could, perhaps significantly, influence the display of information [9]. However, he did not assume an adversary with extensive control of the interface itself.

Security researchers have studied the use of metrics to measure or identify attacks, determine compliance with security processes, and judge the effectiveness of security devices and techniques. Jaquith [16] provides an excellent overview. Security and usability researchers have studied the usefulness of countermeasures and provided insights into user behavior when dealing with Internet attacks [22]. In addition, researchers developed countermeasures for protecting users from intrusive advertising, behavioral targeting, and adversarial webpages. Examples include, *Adblock Plus*, the TOR anonymity network, NoScript, Privoxy, Greasemonkey, and Ghostery.

While our work studies adversarial interface design on the World Wide Web, significant related techniques are employed on the desktop as well as in the physical world. Among numerous other strategies, we encounter applications that interrupt the user to encourage paid updates to free software, complex privacy policies and user agreements displayed in tiny scrolling text boxes, and defaults set to install unwanted software. Distraction and aggressive advertising occur frequently in physical world interfaces as well, typically due to financial incentives. For example, forcing captive movie audiences to view advertisements, the design of hotel televisions to encourage pay-per-view purchases, and billboards dotting the highways.

Pervasive Adversarial Interfaces

Adversarial interfaces are pervasive on the Internet, including both desktop and mobile, but are employed to varying degrees. Our work found that popular websites often aggressively attempt to divert visitors away from user desired tasks toward other goals supportive of the website's business objectives. While we acknowledge that many online services are advertiser supported and users do not pay directly for their use, users do pay a heavy

cost in frustration, lost time, and failed task accomplishment. Automated metrics that effectively characterize the use of adversarial interface techniques on various sites offer the potential to empower end users, inform regulators, and allow more appropriate design decisions by advertisers, publishers, and designers.

At the beginning of this article, we stated that one of our goals was to create metrics based on the principles of Jaquith [16]. We believe that our proposed and implemented metrics (eqs. (1.2) and (2.0)) have achieved this goal. In addition, we have demonstrated the ability to fully automate our metrics, and we have produced meaningful results on 200 popular websites. Our work harnesses the power of the community-driven *Adblock Plus* effort to define, classify, and update adversarial content. We also propose promising, feasible areas for future work that we hope will significantly improve the user experience on the Web.

Appendix

Results from our experiments, including a full listing of URLs, a CSV list of metric results, and our corpus of screenshots and archived website data can be found at http://www.gregconti.com.

Author Information

Ryan Tate is with the United States Army Cyber School at Fort Gordon, GA.

Gregory Conti and ***Edward Sobiesk*** are with the Army Cyber Institute at West Point, NY.

Alexander Farmer is with United States Army Intelligence and Security Command at Fort Meade, MD.

Acknowledgment

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Military Academy, the Army Cyber Institute, the Department of the Army, the Department of Defense, or the United States Government.

The authors would like to thank Ryan Calo, Woodrow Hartzog, Chris Hoofnagle, Tim Jones, Lee Tien, the Electronic Frontier Foundation, and the Hackers on Planet Earth (HOPE) community for their support and encouragement of our adversarial interface work. We would also like to thank Wladimir Palant, lead developer of *Adblock Plus*.

References

[1] A. Acquisti and S. Spiekermann, "Do interruptions pay off? Effects of interruptive ads on consumers' willingness to pay," *J. Interactive Marketing*, vol. 25, no. 4, pp. 226-240, 2011.
[2] J. Alexander and J. Smith, "Disinformation: A taxonomy," *IEEE Security and Privacy*, vol. 9, no. 1, pp. 58-63, Jan./Feb. 2011.
[3] G. Andrews, "Evil interfaces," *The Media Show*, 2009; https://www.youtube.com/watch?v=AnV7nSNksho.
[4] M. Bayles, "Just how 'blind' are we to advertising banners on the web?" *Usability News*, vol. 2, Jul. 2000.
[5] H. Brignull, Dark Patterns, 2012; http://wiki.darkpatterns.org, accessed Oct. 9, 2012.
[6] R. Calo, "Against notice skepticism in privacy (and elsewhere)," *Notre Dame Law Rev.*, vol. 87, pp. 1027, 2012.
[7] M. Chen and H. Janicke, "An information-theoretic framework for visualization," *IEEE Trans. Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1206-1215, Nov./Dec. 2010.
[8] G. Conti, "Evil interfaces: Violating the user," presented at Hackers on Planet Earth (HOPE) Conf,, Jul. 2008.
[9] G. Conti, M. Ahamad, and J. Stasko, "Attacking information visualization system usability: Overloading and deceiving the human," presented at Symp. Usable Privacy and Security (SOUPS), 2005.
[10] G. Conti and E. Sobiesk, "Malicious interface design: Exploiting the user," presented at Int. World Wide Web Conf., 2010.
[11] G. Conti and E. Sobiesk, "Malicious interfaces and personalization's uninviting future," *IEEE Security and Privacy*, vol. 7, no. 3, pp. 64-67, May/Jun. 2009.
[12] A. Goldfarb and C. Tucker, "Online display advertising: Targeting and obtrusiveness," *Marketing Science J.*, vol. 30, no. 3, pp. 389-404, May/Jun. 2011.
[13] J. Harty, "Finding usability bugs with automated tests," *Commun. ACM*, vol. 54, no. 2, pp. 44-49, 2011.
[14] W. Hartzog, "Website design as contract," *American Univ. Law Rev.*, vol. 60, iss. 6, art. 2, 2011.
[15] "IAB display advertising guidelines: The new 2012 portfolio," Interactive Advertising Bureau, Feb. 26, 2012; http://www.iab.net/guidelines/508676/508767/displayguidelines
[16] A. Jaquith, *Security Metrics: Replacing Fear Uncertainty, and Doubt*. Addison Wesley, 2007.
[17] MonztA, Famlam, and Khrin, "EasyList filters designed for Adblock Plus," 2013; https://easylist.adblockplus.org/en/, accessed on Feb. 21, 2013.
[18] M. Pagendarm and H. Schaumburg, "Why are users banner-blind? The impact of navigation style on the perception of web banners," *J. Digital Information*, vol. 2, no. 1, 2001.
[19] J. Nielson, "Banner blindness: Old and new findings," *Alertbox*, Aug. 20, 2007; http://www.useit.com/alertbox/banner-blindness.html, accessed Oct. 7, 2012.
[20] D. Norman, "Design rules based on analyses of human error," *Commun. ACM*, vol. 26, no. 4, pp. 254-258, 1983.
[21] D. Ogilvy, "Ogilvy on advertising," *Vintage*, 1985.
[22] K. Onarlioglu, U. O. Yilmaz, E. Kirda, and D. Balzarotti, "Insights into user behavior in dealing with Internet attacks," presented at Network and Distributed System Security Symp. (NDSS), 2012.
[23] "Section508.gov," U.S. General Services Administration; https://www.section508.gov/, accessed Oct. 8, 2012.
[24] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27. no. 3, pp 379-423, 1948.
[25] X. Suo, Y. Zhu, and S. Owen, "Measuring the complexity of computer security visualization designs," presented at Workshop on Visualization for Computer Security (VizSec), 2007.
[26] R. Tate, G. Conti, and E. Sobiesk, "Automated webpage evaluation," in *Proc. 2nd Ann. Conf. Research in Information Technology (RIIT '13)*. New York, NY: ACM, 2013, pp. 77-82.
[27] J. Turow, J. King, C. Hoofnagle, A. Bleakley, and M. Hennessy, "Americans reject tailored advertising and three activities that enable it," *Berkeley Law*, Sept. 29, 2009; https://www.law.berkeley.edu/centers/bclt/research/privacy-at-bclt/berkeley-consumer-privacy-survey/.
[28] "Techniques for accessibility evaluation and repair," *W3C Working Draft, World Wide Web Consortium*; http://www.w3.org/TR/AERT#color-contrast, Apr. 26, 2000.
[29] "1.0: W3C recommendation," *Web Content Accessibility Guidelines (WCAG)*; http://www.w3.org/TR/WCAG10/, May 5, 1999.
[30] "2.0: W3C recommendation," *Web Content Accessibility Guidelines (WCAG)*; http://www.w3.org/TR/WCAG20/#contrast-ratiodef, Dec. 11, 2008.