2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing

Next Generation of Impersonator Bots: Mimicking Human Browsing on Previously Unvisited Sites

Y. Yang, N. Vlajic, U. T. Nguyen
Dept. of Electrical Engineering & Computer Science
York University
Toronto, Canada
yangcs@cse.yorku.ca, vlajic@cse.yorku.ca, utn@cse.yroku.ca

***Abstract*—The development of Web bots capable of exhibiting human-like browsing behavior has long been the goal of practitioners on both side of security spectrum – malicious hackers as well as security defenders. For malicious hackers such bots are an effective vehicle for bypassing various layers of system/network protection or for obstructing the operation of Intrusion Detection Systems (IDSs). For security defenders, the use of human-like behaving bots is shown to be of great importance in the process of system/network provisioning and testing.**

In the past, there have been many attempts at developing accurate models of human-like browsing behavior. However, most of these attempts/models suffer from one of following drawbacks: they either require that some previous history of actual human browsing on the target web-site be available (which often is not the case); or, they assume that 'think times' and 'page popularities' follow the well-known Poisson and Zipf distribution (an old hypothesis that does not hold well in the modern-day WWW).

To our knowledge, our work is the first attempt at developing a model of human-like browsing behavior that requires no prior knowledge or assumption about human behavior on the target site. The model is founded on a more general theory that defines human behavior as an 'interest-driven' process. The preliminary simulation results are very encouraging - web bots built using our model are capable of mimicking real human browsing behavior 1000-fold better compared to bots that deploy random crawling strategy.

Keywords — bot modeling, interest-driven human browsing

I. INTRODUCTION

Current day Internet abounds in the number and types of Web bots (software programs used to automate the process of retrieval and collection of Web resources). Starting from 2012, traffic generated by bots has accounted for the majority (nearly 60%) of all Web traffic, pushing human-generated traffic to sidelines [1]. Some of the Web-roaming bots perform useful jobs and are referred to as 'benign', while others tend to misuse/obstruct legitimate Internet resources and are referred to as 'malicious'. Example of benign bots are *spider bots* (used by search engines) and *media bots* (provide updates on weather conditions, news, sports). Examples of malicious bots are *spam bots* (harvest email addresses for the purpose of email spamming) and *click bots* (automate clicks on online advertisements to fraudulently generate revenue). According to [1], roughly 50% of all bot-generated traffic in the Internet is generated by good and 50% by bad bots.

Programs designed to mimic/emulate the way humans browse the Internet are a particularly important category of Web bots. Namely, from the perspective of security defenders, these bots are an invaluable tool used in the process of capacity planning and load testing. On the other hand, malicious hackers have an equally keen interest in these types of bots as (e.g.) they can be used to launch powerful hard-to-defend-against DoS/DDoS attacks. (In the literature, Web bots designed to mimic the behavior of legitimate Web clients are generally referred to as *impersonator bots* [1].) Clearly, for both security defenders and malicious hackers, it is critical that the behavior of their human-like acting bots be as close as possible to the behavior of real human visitors to the target system/site. Otherwise, only sub-optimal results can be expected.

To date there have been many attempts to develop models that accurately emulate the way humans browse the Web. These models can generally be grouped in two main categories: 1) models that assume the existence of previous browsing history (i.e., Web logs) from which it is possible to extract sufficient information/knowledge on how humans browse the target site; 2) models built on the classical assumptions concerning *web-page think times* (believed to follow Poisson distribution) and *web-page popularities* (believed to follow Zipf distribution). Unfortunately, both of these models/assumptions are becoming increasingly problematic. For example, many of today's Web-sites are very dynamic in terms of the content they provide and the user populations they attract. Hence, for these sites, log history has very little if any value in modeling/predicting the current or future visitor behavior. On the other hand, the well-known assumption about human browsing behavior being shaped by Poisson and Zipf law is slowly losing ground, as many recent studies contradict this classical hypothesis.

The goal of our work is to propose a radically new approach to the development of human-like behaving bots. In particular, our approach makes no assumptions about specific probability laws governing human Web browsing, nor it relies on any prior log history on which it would be possible to base the modeling of current/future user behavior. The only information that our model relies on is the actual (textual and hyperlink) content of the target site. Using this readily available information, and applying the rules of the so-called 'theory of interest-driven human behavior', our model aims to emulate human-like browsing behavior on new previously unvisited sites (or sites that renew their content and/or structure frequently).

The contribution of our work is multifold. We believe that the use of bots built using our model will find a whole

978-1-4673-9300-3/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCloud.2015.93

host of applications in the areas of Web-site system provisioning, load testing and resource sharing, which are particularly important in the context of Cloud computing. Our work will also open a new chapter in the design of IDS systems capable of dealing with sophisticated next-generation application-layer DDoS attacks.

The content of this paper is organized as follows. In Section II, we provide a brief survey of related works on the subjects of modeling/mimicking of human browsing and theory of interest-driven human behavior. In Section III, we present the conceptual and mathematical details of our model of human browsing behavior using the theory of interest-driven behavior. In Section IV, we present some of our preliminary experimental results. In Section V we outline potential directions for future research.

II. Related Work

A. Modeling of Human Browsing Behavior

In earlier research works, the following two approaches to modeling of human browsing behavior have commonly been used:

a) *Generalized statistical modeling*. The key presumption of this approach is that the actual log history of the target site is not accessible or available. Thus, the model is built on what are believed to be common characteristics of human browsing behavior across all/most Web sites. For example, according to some early papers on the characterization of Web traffic ([2], [3]), probabilities of Web page requests are believed to follow the well-known Zipf distribution, while times spent on each individual page (so-called think times) are believed to follow Poisson or Pareto distribution. However, this hypothesis - that human browsing behavior is homogeneous (i.e., site independent) and can be modelled with the above mentioned distributions - is increasingly being questioned ([4], [5], [6], [7]). Namely, many argue that with the ever increasing size and complexity of the Web and its applications, the way humans look for and retrieve information 'on-line' has also undergone considerable change and diversification. Consequently, the old 'one size fits all' approach to modeling of human browsing behavior is no longer appropriate.

b) *Predictive modeling based on log history*. The works deploying this approach do not make any assumptions about the nature of human browsing. Instead, they (only) assume that the log history of the target site is available for analysis. By studying (i.e., learning from) the past behavior of human visitors to the target site, these works aim to make predictions about future visitor behavior. The modeling/prediction techniques commonly used in these works are Markov Models and Bayesian Networks (e.g., [8], [9]). Now, the main obvious drawback of this entire group of works is that they are clearly inapplicable in cases when past server logs are not accessible or available (e.g., in the case of a test Web site or a Web site hosted by a third-party). Furthermore, as indicated earlier, many of the current day Web sites are known to experience dynamic and often unpredictable change – not only in their content and organization, but also in the number and type of visitors requesting their information/services. (E.g., the content of a news agency web-site may change in the matter of minutes, and depending on the type of 'current news' different demographic groups may be visiting the site.). Consequently, in such sites, using past log information to make prediction about the users' future browsing behavior is likely to produce suboptimal results, at best.

B. Theory of Interest-Driven Human Behavior

The theory of interest-driven human behavior was first introduced in [5] and [6]. According to these two break-through works, many real-world human activities are mainly driven by personal interests, and could not be treated simply as tasks needing execution (i.e., do not fit the paradigm of Poisson-based queueing theory). Specifically, human behavior seems to be driven by an interplay between 'personal interests' and 'frequency of events/actions'. As stated in [5], "frequency of events/actions is determined by the interest, while the interest is simultaneously affected by the occurrence of events/actions." In other words, a new event/action may initially spark interest in the same/similar type of activity and intensify its occurrence. However, as the overall number of repetitions (i.e., frequency) of this activity increases, the respective interest is likely to subside. This, in turn, will gradually decrease the frequency of activity back down to (near) zero. The interdependence between the frequency of an event/activity and the interest in the given event/activity is illustrated in Figure 1.

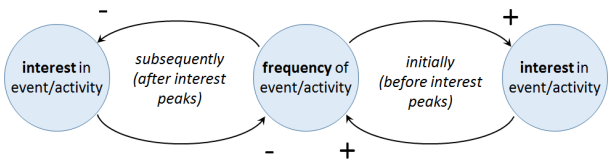

Figure 1 Causality between frequency of an event activity and interest in the given event/activity in human behavior

III. Model of Human Browsing Behavior Using Interest-Driven Theory (HBB-IDT)

A. Key Assumptions and Definitions

In this section, we lay the foundation for further discussion by introducing the key assumptions and definition as pertaining to our model of **H**uman **B**rowsing **B**ehavior using **I**nterest-**D**riven **T**heory (HBB-IDT).

a) Browsing Process

In general, there are two sets of features that characterize any browsing process, whether generated by a human or bot: a) the sequence of web pages visited/requested, and b) the stay time on each of the visited pages. In our work, we employ the following definitions of the key terms describing the browsing process (*web page*, *successor page*, *stay time*):

- ***Web Page***. A web page is an HTML document with a unique URL and meaningful human-readable text content. A web page may contain several hyper-links (URLs) to other HTML documents, multi-media resources, MIME files, CSS or JAVASCRIPT files. Note that in our work files and embedded objects that do not contain human-readable text (CSS, JSON data, JAVASCRIPT codes, etc.) are not considered a web page, even though they may also be referenced by a unique URL and their requests/retrievals may also be recorded in server logs. Generally, requests to these non-human-readable objects are caused by a request to a web page that contains or points to them.
- ***Successor Page***. During a browsing process, web page B requested immediately after the currently visited web page A is called page A's successor.
- ***Stay Time***. The stay time on a web page is defined as the time interval between the user requesting the given page and the user requesting this page's successor. In the literature, this time interval is also commonly referred to as *Think Time*. Generally, the length of stay time on a web page is determined by the user's interest in the page's theme and content.

b) Web Page Theme and Content

In our work/model, we assume that the user's interest in each visited page is influenced by three factors: the user's interest in the page's *theme*, the user's interest in the page's *content*, and the *content quality* of the given page. We distinguish and define these three concept as follows:

- ***Theme***. Theme is the set of main general topic(s), subject(s) or idea(s) conveyed in a web page. For example, the possible themes of a news page about Apple's stock are Business, Technology and Finances.
- ***Content***. The content of a web page is defined as the substantial information provided/found in its text. Web pages belonging to the same theme may (i.e., likely will) provide different contents. For example, a news page about Microsoft's stock would belong to the same theme as a page on Apple's stock, but their actual contents would obviously be different.
- ***Content Quality***. We define *content quality* as the actual attractiveness of a webpage's content to the user. It can generally be expected that a webpage with more information and sufficient but not overly high similarity to the content of previously visited webpage is more attractive to the user, and thus has higher content quality.

Now, during the browsing process, the choice of a particular web page and the stay time on it are impacted by a fine interplay among the three above mentioned parameters. Namely, once the user decides to open/retrieve a web page based on his/her interest in that page's general *theme*, the stay time on the given page will depend on his/her interest in the page's actual *content* as well as the page's *content quality*. In some cases, even though the user may be very interested in the page's theme, he/she may quickly jump/move to another page if the page's content is not attractive enough or its quality is not satisfactory.

c) Web Page Closeness

In order to measure the 'contextual proximity' between two web pages, which we refer to as *web page closeness*, we establish the following three metrics:

(1) ***Theme Closeness***, depicts the degree at which the themes of two web pages are similar. To decide how close the themes of two web pages actually are, we examine the text-similarity between their 'digests' comprising hyperlink tip-texts, page titles and keywords. In information retrieval and text mining, to allow for comparison of different pieces of texts, each word/term is notionally assigned a different dimension. Subsequently, an entire piece of text is characterized by a vector, where the value of each dimension corresponds to the number of times that the respective term appears in the given text. Under this model, one of the most practical method to compute similarity between two pieces of text is cosine similarity [4], which corresponds to the inner product or their respective vector representations. Following this approach, we use expression (1) to calculate the theme closeness between web pages i and k.

$$S(i,k) = \frac{\sum_{j=1}^{|H|} R(h_j,i) \cdot R(h_j,k)}{\sqrt{\sum_{j=1}^{|H|} R^2(h_j,i)} \cdot \sqrt{\sum_{j=1}^{|H|} R^2(h_j,k)}} \qquad (1)$$

In (1), $S(i,k) \in (0,1)$ annotates the theme closeness between web pages i and k (i.e., between their respective digests), set $H=\{h_j\}$ is the theme-domain list, and $R(h_j,i)$ keeps the count of the number of times that the terms associated with theme h_j appear in page i (i.e., its digest). In our work, we use WordNet Domains Hierarchy (WDH, version 3.2) [5] theme-domain list to generate H. This theme list classifies 115,424 English words into 161 domains, allowing that some words be mapped into several different domains.

(2) ***Content Closeness***, depicts the degree at which the contents of two webpages are similar. As in the case of theme closeness, the content closeness between pages i and k can be calculated using the following formula

$$S'(i,k) = \frac{\sum_{j=1}^{|G|} R(g_j,i) \cdot R(g_j,k)}{\sqrt{\sum_{j=1}^{|G|} R^2(g_j,i)} \cdot \sqrt{\sum_{j=1}^{|G|} R^2(g_j,k)}} \qquad (2)$$

In (2), $S'(i,k) \in (0,1)$ annotates the content closeness between web pages i and k, set $G=\{g_j\}$ is the list of all meaningful English words possibly found in a text/page (115,424 words according to WDH), and $R(g_j,i)$ keeps the count of the number of times that word g_j has appeared in page i.

(3) ***Visibility Closeness***, between the current page and one of its links (i.e., linked pages) depicts the likelihood that the user visually spots/finds the given hyperlink inside the current page. In many web sites, links that are visually more pronounced and/or better positioned are more likely to contain contextually related information. Furthermore, they are also more likely to be chosen for viewing/retrieval by the user. In our work we use formula shown in (3) to calculate the visibility closeness between web pages i and j.

$$V(i,j) = \begin{cases} 0 & , \quad \text{if there is no links from } i \text{ to } j \\ 0.5 - 0.5\sin\left(\left(\frac{Loc(i,j)}{L(i)} - 0.5\right)\pi\right), & \text{if there is a link from } i \text{ to } j \end{cases} \quad (3)$$

In (3), $V(i,j) \in (0,1)$ annotates the visibility closeness between pages i and j, $L(i)$ is the overall number of characters appearing in web page i, and $Loc(i,j)$ is the number of characters appearing in web page i before the hyper link to j. Figure 2 depicts the change in $V(i,j)$ as a function of $Loc(i,j)$. Clearly, the value of $V(i,j)$ is high when the link to page j appears at the top or in the upper half of page i. and then this value gradually decreases.

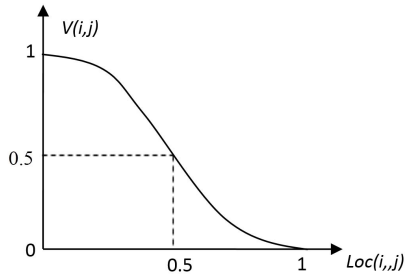

Figure 2 $V(i,j)$ as a function of $Loc(i,j)$

B. HBB-IDT Model

The actual outline of our model of **H**uman **B**rowsing **B**ehavior using **I**nterest-**D**riven **T**heory (HBB-IDT) is provided in Figure 3. The figure captures not only the key states/parameters, but also their interdependence and dynamicity. In the figure, "+" implies positive and "–" implies negative correlation. P_n is the current visited webpage, P_{n-1} is the page visited before P_n, and P_m is one of the candidate pages to be visited next.

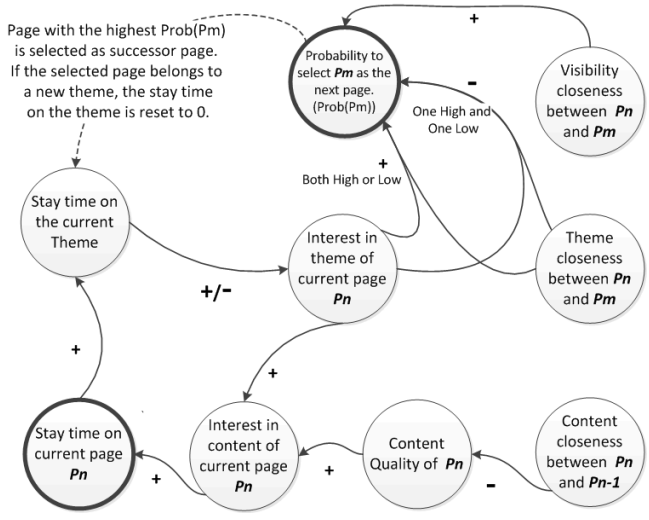

Figure 3 HBB-IDT Model

From Figure 3, the following properties of our model can be observed:

1) ***Rules governing change in theme***. In reality, it is quite common that a user looks at several different topics/themes (in succession) during a single browsing session. The transition from the current theme to another is generally correlated with the (cumulative) ***stay time on current theme***. Namely, when the user first opens a web page on a new theme, it is reasonable to assume that his/her interest in this theme is high. Following this, the user is also likely to open other pages on the same/similar theme. However, as the stay time on the same theme increases, the user will gradually become less interested (i.e., bored) with this theme, and he/she will be more likely to open a webpage on a different theme. We capture this phenomenon in our model by indicating that the interest in the current theme (i.e., theme of currently visited page) is initially positively but then negatively impacted by the cumulative stay time on the given theme (see Figure 3).

In the actual implementation of our model, we use the following formula to calculate the overall stay time on (current) theme:

$$d_i = \begin{cases} 0 & , \quad \text{if } i = 1 \text{ or } S(i,i-1) \leq \theta \\ d_{i-1} + t_i & , \quad \text{if } S(i,i-1) > \theta \end{cases} \quad (4)$$

In (4), t_i is the stay time on the current page i, $S(i,i-1)$ is the theme closeness between the current and previously visited page, and θ is a pre-set threshold that determines whether the current and previous page belong to the same theme. Building on (4), we employ formula (5) to capture the above described change in the user's interest in the current theme.

$$c_i = Max\left(0, \quad \frac{c' - 1}{d_m^2}(d_i - d_m)^2 + 1\right) \quad (5)$$

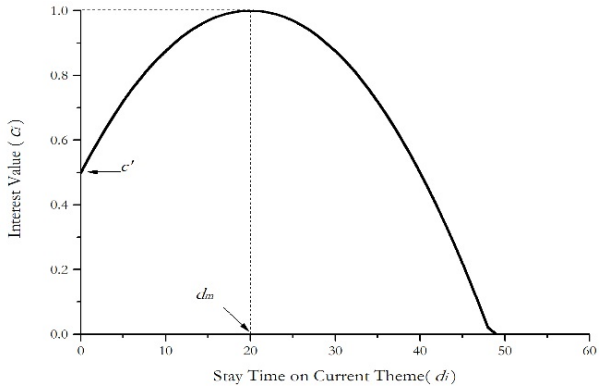

Figure 4 c_i as a function of d_i

In (5), $c_i \in (0,1)$ annotates the user's interest in the current theme, c' is a pre-set constant or a random number $\in (0,1)$ representing the user's initial interest value in a new theme, d_m is also a pre-set constant representing the time interval at which the user's interest in any given theme is expected to

reach its highest value (and then start declining after that), and d_i is the actual stay time on the given/current theme (see (4)). The character of the change in the user's interest in the current theme (c_i) relative to the stay time on the given theme (d_i) is depicted in Figure 4.

2) ***Rules governing change in content and length of stay time***. Change in the user's interest in the current content effectively implies that the user has decided to open another/new webpage. When a user opens a new webpage, his/her interest in this new content will mainly depend on (i.e., be positively correlated with) his/her interest in the general *theme* of the web page as well as the page's *content quality*, as shown in Figure 3. Higher interest in the content of the current page will ultimately imply longer stay time on the given page, as also indicated in Figure 3.

In the actual implementation of the HBB-IDT model, we use formula (6) to calculate the content quality of the current page ($q_i \in (0,1)$)

$$q_i = \begin{cases} \frac{L(i)}{L_{max}}, & i = 1 \\ \frac{L(i)}{L_{max}}\left(1 - S'(i, i-1)\right), & i > 0 \end{cases} \quad (6)$$

In (6), *L(i)* is content length (i.e., word count) of page *i* and L_{max} is the maximum content length found in all pages of the given site. Clearly, from (6), pages that contain more text/information and provide substantially different/novel content relative to the previous page(s) will be associated with better content quality.

Given c_i (see (5)) and q_i (see (6)), we are able to estimate/predict the user's stay on page *i* as:

$$t_i = \frac{L_{max}}{Z} \cdot c_i \cdot q_i \quad (7)$$

Note in (7) Z represents the average human reading speed and, consequently, L_{max}/Z represents the maximum possible stay time the average user is expected to experience/exhibit on the given site. It should be clear that (7) complies with our previous discussion as a human user is expected to spend longer time viewing page *i* (time close to L_{max}/Z) if both his interest in the page's theme (c_i) and the page's content quality (q_i) are high, and shorter time otherwise.

3) ***Rules governing selection of successor page***. In our model, the first web page to be visited (web page i=1) can be randomly chosen among all web pages in the site. After that, the model decides whether to choose a particular page (page *j*) by examining the following three parameters: the user's interest in the current theme (c_i), the theme closeness between the current page and page *j* (*S(i,j)*), and visibility closeness between the two pages (*V(i,j)*). Now, out of the three parameters, the value of *V(i,j)* is an independent factor that positively feeds into the probability of choosing page *j* as the successor page. (Clearly, links/URLs that are more visually pronounced and 'catchy' have a higher chance of being selected/requested, and vice versa.) On the other hand, the impact of c_i and *S(i,j)* on the selection of page *j* is more complicated as it requires that the values of these parameters relative to each other be examined. In particular, if the user's interest in the current theme is high (c_i is high), and page *i's* theme is very close to page *j's* theme (*S(i,j)* is high), then the probability of visiting page *j* next is high. Similarly, if the user's interest in the current theme is low, and page *i*'s theme is very different from page *j*'s theme, then the probability of visiting page *j* in the next step is also high. In all other cases, page *j* chances to be chosen as the successor page are low(er).

In our implementation of the HBB-IDT model, we use formula (8) to estimate the relative chances of page *j* becoming the successor of page *i* (*E(i,j)*). Ultimately, the page with the highest value of *E(i,j)* will become the actual successor of page *i*.

$$E(i,j) = \varphi V(i,j) + (1-\varphi)\frac{2 \cdot c_i \cdot S(i,j)}{c_i^2 + S^2(i,j)} \quad (8)$$

In (8), φ is a pre-set parameter that depict the importance/weight of visibility closeness *V(i,j)* relative to the other two parameters. (E.g., for a high φ, we expect the user to be more influenced by the visual organization of a web-page then the elements related to its content/theme, and vice versa.) The combined effect of *S(i,j)* and c_i on *E(i,j)* is shown in Figure 5. It is obvious from Figure 5 that when the two factors are both high or both low, *E(i,j)* will be significantly higher than when either of these factors is low.

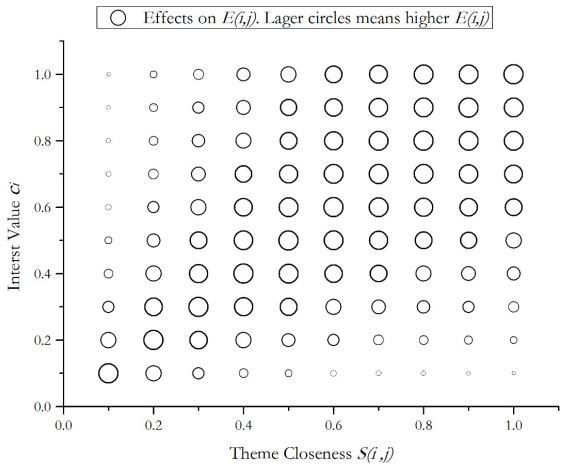

Figure 5 *E(i,j)* as a function of c_i and *S(i,j)*

IV. Experimental Results

A. Emulation and Evaluation Software

Based on the model outlined in the previous section, we have developed a software framework for emulation and evaluation of HBB-IDT model. The framework is built in Java and comprises the following components:

(1) **Content Gatherer,** which crawls and downloads all web pages from the target website. Note, even though our model does not require any knowledge of the system logs or

prior human behavior on the target site, the system does assume that the site's content (i.e., the content of its individual webpages) is readily known and available.
(2) **Data Analyzer** is responsible for analyzing the content of individual webpages and determining their respective themes and linkage maps. Data Analyzer's results are stored in an SQL Server database to support the next function.
(3) **HBB-IDT Crawler** is the core of the system. By relying on the data provided by (1) and (2), and by implementing the functions outlined in Section 3, HBB-IDT crawler browses the target site by implementing HBB-IDT rules.
(4) **HBB-IDT Evaluator** receives the browsing sequences generated by HBB-IDT Crawler and calculates how well they match against real human-generated sequences as well as random browsing sequences.

B. *Experimental Setup*

To evaluate our model in real-world conditions, we chose the most popular news website in Canada, www.cbc.com as our study case. The reason for this is: (1) the website covers a wide range of themes and contents and thus is likely to satisfy a large number of different interests; (2) textual components are the primary source of information in most of its web pages, which meets our research assumptions; (3) it is mostly a one-way interaction website (i.e., most users requesting and not posting information), which also agrees with the assumptions of our research.

At this stage of our research, accessing original CBC Web logs has not been possible. Thus, in order to obtain true human browsing sequences on this site (which are ultimately needed to evaluate our HBB-IDT model), we have set up a mirror-site of www.cbc.com/news directory. (Subsequently we have asked a number of volunteers to browse through the mirror site the same way they would do on the original site.) The major steps in the set-up of the mirror site have included:
(1) Downloading web pages of www.cbc.ca/news using HTTrack Website Copier v3.48 - a popular software that facilitates the creation of web site mirrors.
(2) Uploading the mirror website to a Microsoft's cloud virtual server, and configuring it in IIS 9.0. The home URL of mirror site has been set to http://cse.cloudapp.net, and the site is made accessible from anywhere on the Internet. Note, to catch necessary logs and identify different human users (even those coming from the same/shared IP address), an ASP file (assigned to be the default index page) has been put in the root directory of the site. Then, the IIS' logging module has been set to catch not only standard information but also the field CS(COOKIE). Therefore, any user visiting the mirror site would be assigned a unique ASPSESSIONID after entering http://cse.cloudapp.net through their browser.

C. *Experiment Execution*

A mirror copy of the original CBC site was created on May 19, 2015. The mirror site consisted of 22,740 files occupying a total of 1.6 GB. Immediately following the creation of the mirror site, a group of volunteers was asked to visit the site over the following 24h period. At the end of the given period, we were able to group all logged requests into 31 sessions (using IP address and ASPSESSIONID data). Some of the sessions were obviously generated by our human volunteers - sessions numbered 1, 6, 8, 11, 25 (see Table 1). Remaining sessions were generated by well-known bots that happened to crawl through the mirror site.

D. *Experimental Results*

Recorded sessions generated by human visitors were compared against sessions generated by our HBB-IDT crawler, as well as against sessions generated by a bot deploying random browsing strategy. The comparison has revealed that our model was able to emulate real human browsing several 100- to several 1000- fold better than the randomly browsing bot. (The actually obtained results for a number of specific human sessions are provided in Table 1.)

TABLE I. PERFORMANCE OF OUR MODEL VS. RANDOM CRAWL MODEL

Session Number	Probability that out HBB-IDT crawler has generated this session vs. probability that random crawler has generated this session
1	220.80
6	5385.00
8	262.00
11	93.84
25	2821164.00

V. FUTURE WORK

Although the research presented in this paper is still ongoing, we hope our HBB-IDT model and our initial experimental results will be a catalyst for a broader discussion, and will ultimately mark a new era in the design and utilization of human mimicking bots. Our future research effors will involve a broader scale evaluation of the HBB-IDT model – involving a larger number of volunteers as well as working with original server logs.

REFERENCES

[1] Incapsula, "*2014 Global Bot Traffic Report*", 2014.
[2] P. Barford, M. Crovella, "*Generating Representative Web Workloads for Network and Server Performance Evaluation*", ACM International Conference on Measurement and Modeling of Computer Systems SIGMETRICS, Madison, Winsconsin, USA, Jul 1998.
[3] Z. Liu, N. Niclausse, C. Jalpa-Villanueva, "*Traffic model and performance evaluaiton of Web Servers*", Elsevier Journal of Performance Evaluation, Vol 446, Issue 2-3, pp. 77-100, Oct 2001.
[4] R. Morris, D. Lin, "*Variance of Aggregate Web Traffic*", IEEE INFOCOM Conference, Tel Aviv, Isreal, Mar 2000.
[5] Barabasi, Albert-Laszlo. "*The origin of bursts and heavy tails in human dynamics*", Nature 435, pp. 207-211, May 2005.
[6] T. Zhou, H. Xiao-Pu, W. Bing-Hong, "*Towards the understanding of human dynamics*", Book Chapter in Science matters: humanities as complex systems, pp. 207-233, 2008.
[7] R. D. Smith, "*The Dynamics of Internet Traffic: Self-Similarity, Self-Organization, and Complex Phenomena*", Journal of Advances in Complex Systems, Vol 14, Issue 06, Dec 2011.
[8] M. Deshpande, G. Karypis, "*Selective Markov models for predicting Web page accesses*", ACM Transactions on Internet Technology (TOIT), Vol 4, Issue 2, pp. 163-184, 2004.
[9] M. A. Awad, I. Khalil, "*Prediction of user's web-browsing behavior: Application of markov model*", IEEE Transactions on Systems, Man and Cybernetics, Vol 42, Issue 4, pp. 1131-1142, 2012.