CHICAGO TRIBUNE

CONTENT VELOCITY ANALYSIS

KALEV LEETARU

OVERVIEW

This report presents the findings of a small pilot study examining content velocity on the Chicago Tribune's website, http://www.chicagotribune.com/. Conversations with the Tribune indicated they were unable to provide detailed metrics on content additions to their website, so a custom crawling system was deployed to archive the 105 "gateway" pages (the article listing pages of each section) every half hour for 34 days, from 9/15/2010 through 10/19/2010. The resulting 136,605 snapshots were used to characterize the paper's linking velocity over this period.

Primary findings are that 83% of the Tribune's links were to the DoubleClick.net advertising network, with just 11% of links pointing to Tribune pages with an average link lifespan of 56 hours and a range from 18 hours to 7 days. Only 5% of pages were linked to from more than one gateway page, suggesting the Tribune requires users to employ extensive subnavigation to locate all stories, and crawling-based archival systems must utilize more extensive crawling strategies to locate all content. No master "latest content" portal pages or RSS feeds are available, meaning that a page-by-page crawl of all top- and sub-level gateway pages is required to compile a list of new pages, greatly increasing the resource requirements for archiving the Chicago Tribune.

Roughly 39% of Tribune URLs are linked for a day or less and an average of 735 new links are posted each day to Tribune content. The highest number of new links are added on Thursdays, while Sundays have the fewest updates. Content sections exhibit strong stratification in number of links, percentage of links directed to other Tribune pages, and link lifespan, suggesting that content characterizations must be topically-oriented on large news sites, rather than site-wide.

ANALYSIS

During the 34 day monitoring period, there were a total of 366,290 links in the monitored gateway pages. Of these, 304,684 (83%) were links to the DoubleClick.net advertising network, while 41,220 (11%) were to Chicago Tribune pages, and the remaining 20,386 (6%) were to other external sites, including Tribune partners. That more than 83% of the links across the Chicago Tribune's website during this period pointed to paid advertising demonstrates the importance advertising plays even in mainstream online news websites, and the density of paid commercial messages against the journalistic news content developed by the papers themselves. Note that this study did not attempt to measure the screen real estate in pixels or percentages occupied by advertising versus journalistic content, but such a high link density would suggest a significant amount of link space is dedicated to advertising content.

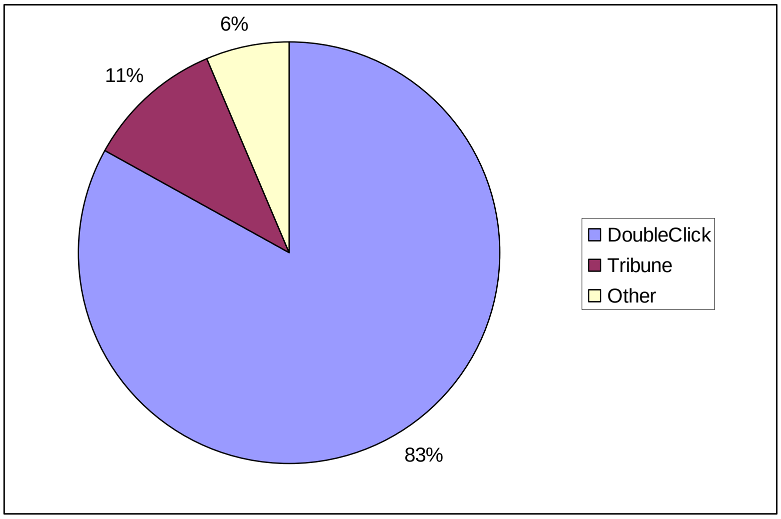

Figure 1 - Breakdown of Chicago Tribune links by target

Appendix A, Domains by Links, contains a complete list of all domains linked to by the Chicago Tribune and the number of distinct URLs linked to at that domain. Of particular note is that several of the Tribune's partner sites are among the most-linked-to domains, including chicagobreakingsports.com and chicagonow.com. Two other papers in the Tribune stable, the LA Times and the Orlando Sentinel, are frequently linked from the Tribune. In the world of print, a newspaper very rarely runs a story from a primary competitor, yet in a strange twist, in a one month period, the Tribune forwarded its visitors to more than 515 articles on the Sun Time's site, making it only the second most-linked to external domain, behind Twitter. In all, the Tribune linked to URLs hosted by 458 different domains and subdomains.

Just 1,950 of the 41,220 Tribune links (5%) were linked to from more than one gateway page, and those that were tended to be links to other main pages like links from sub-level pages such as /sports/cubs/ back to the main /sports/ page. This is a critical finding from the standpoint of archiving the Tribune, as it suggests that to fully collect the entirety of the Tribune's content, crawling only root-level pages like /sports/ will not be enough: it will be necessary to crawl every single sub-level page like /sports/cubs/, /sports/football/, etc. This significantly increases the resources necessary to archive the paper.

Appendix B, Tribune URLs by Lifespan, shows all 41,220 Tribune URLs linked from the monitored gateway pages during the 34 day monitoring period, ordered by the number of minutes the URL was linked to. Since snapshots were taken every 30 minutes, lifespan is measured in multiples of a half hour, with 30 minutes indicating the link was found in only a single snapshot (likely a high-velocity content section), while those with long spans measuring in days indicate URLs that were kept linked from the site for long periods of time. Note that the lifespan of a URL measures only the length of time a link was kept to it from one of the 105 monitored gateway pages, not the length of time the article was kept on the Tribune site. In most cases, pages remain on the site indefinitely, and access transitions from gateway links to requiring keyword search engine use for discovery.

There were 2,029 URLs (5%) linked for 30 minutes or less, while an additional 928 were linked for 60 minutes or less. Nearly 39% of URLs (16,249) were linked for a day or less. The figure below shows the total number of URLs at each lifespan duration. There appears to be fairly even distribution of lifespans beyond the one day mark, with the only outlier being a strong clustering at a duration of 6 days, suggesting a weekly rotation cycle for some Tribune content sections. That 61%

of the Tribune's content is nearly evenly distributed among linking lifespans (as opposed to primarily being clustered in the 1-3 day range) suggests that the Tribune uses a management strategy emphasizing using a large number of gateway pages to keep links online, as opposed to a small number of high-velocity linking pages.

Eliminating URLs linked for two or more weeks (these are likely gateway page links or links to standing features) leaves 30,226 URLs (73%), with a mean link lifespan of 56 hours.

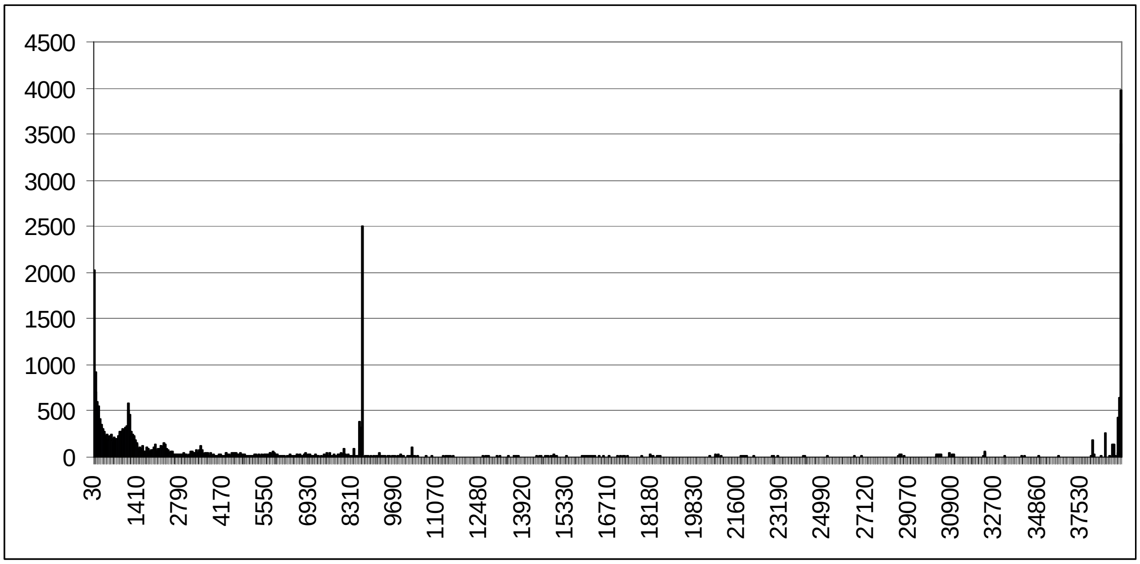

Figure 2 - Histogram of number URLs updated by time interval

Looking only at Tribune links, the figure below shows the total number of new links added to the site each day after the start day on 9/15/2010, exhibiting strong weekly stratification. There is an average of roughly 735 new links per day. While the Drudge Report had the most updates on Tuesdays and the fewest on Saturdays, [1] the Tribune has the highest number of new pages added on Thursdays, with the fewest new pages on Sundays. A linear trendline has been plotted through the center of the timeline indicating that the Tribune appears to have reduced its update rate in nearly linear fashion throughout the duration of this analysis. It is unclear whether this is a longer-term reduction in the Tribune's update velocity, or whether it simply reflects a specific news cycle profiled during this period.

[1] http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2500/2235

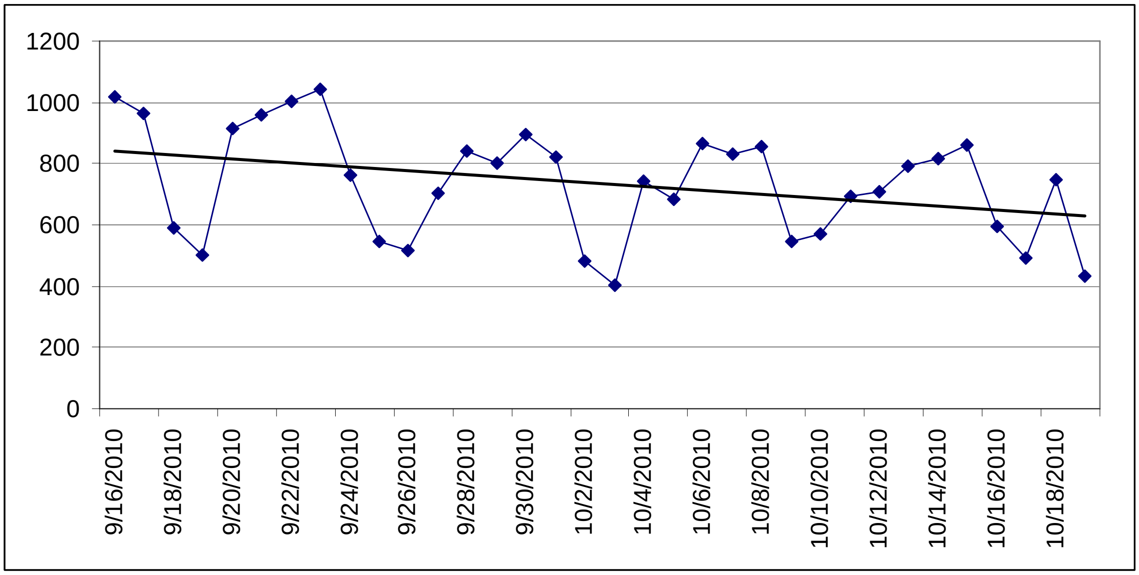

Figure 3 - Number new Tribune links seen by day

Appendix D, Stats by Section, contains a breakdown of each of the 105 gateway pages and the total number of links to all sites, number of links to other chicagotribune.com pages, number of Double Click advertising links, and average lifespan in hours of chicagotribune.com links, for each section. The blogs and special elections section of the Tribune have very little advertising, high link lifespans, and low content volumes. The only three sections to have sub-24-hour lifespans are the main Business, Sports, and Celebrity sections. Even the front page has a link lifespan of 25 hours, suggesting relatively slow turnover on featured stories, despite that section having the highest total number of links. Among individual content sections, the Sports section has the highest number of links at 9,940 (3,384 to Tribune pages), nearly one thousand more than the business section. Ranking by total number of links to Tribune pages, as opposed to all links, Nation & World news comes in at third, followed by Entertainment and then Local News.

Ranking sections by the percentage of their links that point to Tribune pages versus external properties, the special elections.chicagotribune.com section was highest at 95%, followed by the Tribune's major blogs. The highest non-blog section was its Sports page at 34%, its main page at 29%, and its Nation & World news at 28%. The pages with the lowest density of Tribune links are several of its travel sections with less than 4%. That no major content section has more than a third of its links pointing to Tribune pages shows the tremendous role external resources and advertising play on the Chicago Tribune's website.

Table 1 - Gateway pages ordered by average link lifespan

Section	Total	Tribune	%Tribune	DoubleClick	Lifespan
/business/	9003	2180	24.21	5134	18.97
/sports/	9940	3384	34.04	5192	20.21
/entertainment/celebrity/	5403	1090	20.17	3900	22.08
/features/horoscopes/	5503	244	4.43	5200	25.08
/	13030	3855	29.59	5200	25.48
/technology/deals/	3345	689	20.60	2602	25.61
/news/nationworld/	5704	1628	28.54	3900	29.38
/news/local/chicago/	4927	426	8.65	3900	29.77
/sports/football/bears/	5310	1083	20.40	3903	35.15

/sports/college/	5362	1118	20.85	3897	35.19
/health/	5394	789	14.63	4096	36.19
/news/education	4477	464	10.36	3894	36.62
/news/opinion/blogs/	8041	735	9.14	3892	37.48
/news/opinion/share/	5573	345	6.19	5168	38.01
/entertainment/	5762	1458	25.30	3848	38.10
/news/politics/	4710	619	13.14	3903	38.84
/news/local/	6309	1057	16.75	3879	43.35
/news/opinion/	5037	831	16.50	3900	45.79
/news/columnists/all/	4867	908	18.66	3898	47.37
/news/columnists/all/	4867	908	18.66	3898	47.37
/news/corrections/	4301	337	7.84	3903	50.16
/sports/baseball/whitesox/	5028	850	16.91	3900	50.73
/sports/highschool/	4847	838	17.29	3798	52.01
/sports/hockey/blackhawks/	4808	696	14.48	3903	53.07
/features/columnists/	5627	359	6.38	5204	53.35
/sports/baseball/cubs/	4966	808	16.27	3903	55.50
/travel/	5827	582	9.99	5175	59.53
/sports/basketball/bulls/	4636	501	10.81	3900	61.74
/features/food/	4492	414	9.22	3903	63.51
/news/opinion/commentary/	4676	501	10.71	3894	64.61
/news/opinion/editorials/	4397	434	9.87	3903	67.04
/features/gardening/	4290	318	7.41	3900	71.08
/business/yourmoney/	5880	255	4.34	5204	71.90
/entertainment/tv/	4963	407	8.20	3901	73.22
/business/columnists/	5566	307	5.52	5200	73.65
/features/style/	5617	347	6.18	5204	79.39
/features/tribu/	4380	432	9.86	3887	80.27
/business/investors/	2847	197	6.92	2600	80.98
http://newsblogs.chicagotribune.com/clout_st/	429	390	90.91	2	83.75
/travel/escapes/	4263	277	6.50	3900	84.22
/entertainment/dining/	4483	404	9.01	3903	84.73
/sports/golf/	4339	376	8.67	3903	88.00
/news/local/suburbs/	1539	291	18.91	858	90.48
/sports/tennis/	4167	213	5.11	3903	93.86
/sports/smack/	6045	365	6.04	3900	95.24
http://newsblogs.chicagotribune.com/tribnation/	486	400	82.30	2	95.80
/entertainment/events/	4407	376	8.53	3903	96.14
/news/opinion/letters/	4552	589	12.94	3903	97.72
/business/smallbusiness/	5493	234	4.26	5200	99.85
http://elections.chicagotribune.com/	148	142	95.95	2	100.53
/features/family/	5619	244	4.34	5200	101.67
/news/local/southsouthwest	4381	230	5.25	3897	102.07
http://leisureblogs.chicagotribune.com/the_theater_loop/	535	455	85.05	5	108.16
/news/watchdog/	4205	239	5.68	3903	110.04

/features/games/	5451	199	3.65	5196	110.67
/health/agentorange/	2825	177	6.27	2598	114.83
/news/strange/	4566	601	13.16	3903	115.10
/entertainment/movies	4458	354	7.94	3900	115.47
/news/local/northnorthwest	4380	235	5.37	3900	124.02
/business/problemsolver	5608	298	5.31	5200	125.72
/travel/midwest/	5416	166	3.06	5194	127.27
/features/lottery/	5416	155	2.86	5204	127.47
/travel/family/	4156	202	4.86	3897	128.26
/sports/soccer/	4214	253	6.00	3903	128.29
/news/local/suburbs/evanston/	1517	176	11.60	866	128.56
/news/local/suburbs/joliet/	1512	172	11.38	866	129.76
/news/local/suburbs/arlington_heights	1457	170	11.67	863	131.12
/news/local/suburbs/tinley_park/	1404	170	12.11	866	131.49
/news/local/suburbs/wheaton/	1532	170	11.10	866	131.51
/news/local/suburbs/orland_park	1396	171	12.25	866	131.62
/news/local/suburbs/wilmette-kenilworth/	1379	173	12.55	863	131.82
/news/local/suburbs/northbrook	1470	171	11.63	863	131.95
/news/local/suburbs/libertyville	1514	171	11.29	866	131.96
/news/local/suburbs/schaumburg/	1448	168	11.39	866	132.19
/news/local/suburbs/crystal_lake/	1475	168	11.31	866	132.19
/news/local/suburbs/des_plaines/	1485	168	11.19	866	132.19
/news/local/suburbs/glen_ellyn/	1501	168	11.60	866	132.19
/news/local/suburbs/plainfield/	1497	169	11.29	866	132.26
/news/local/suburbs/deerfield	1393	172	12.35	866	132.26
/news/local/suburbs/elgin/	1481	169	11.40	866	132.27
/news/local/suburbs/bolingbrook/	1483	169	11.41	866	132.27
/news/local/suburbs/grayslake/	1391	170	10.81	866	132.34
/news/local/suburbs/gurnee/	1436	170	11.00	866	132.34
/news/local/suburbs/elmhurst	1545	170	12.22	866	132.34
/news/local/suburbs/downers_grove/	1573	170	11.84	866	132.34
/news/local/suburbs/hinsdale	1449	170	11.73	866	132.34
/news/local/suburbs/naperville	1619	170	10.50	866	132.35
/travel/deals/	5414	164	3.03	5200	132.45
/news/local/suburbs/glenview	1453	172	11.84	866	132.48
/news/local/suburbs/oak_park-river_forest	1650	172	10.42	866	132.48
/news/local/suburbs/winnetka-northfield/	1382	172	12.45	866	132.48
/news/local/suburbs/highland_park-highwood/	1452	173	11.91	866	132.55
/news/watchdog/nursinghomes/	4148	192	4.63	3900	133.18
/entertainment/theater/	4306	258	5.99	3900	133.33
/travel/other/	5459	206	3.77	5196	135.91
/travel/unitedstates/	5465	208	3.81	5200	136.04
http://featuresblogs.chicagotribune.com/printers-row/	123	79	64.23	2	140.56
http://leisureblogs.chicagotribune.com/turn_it_up/	443	406	91.65	2	142.45

/travel/chicago/	5453	196	3.59	5197	146.86
/news/local/west	4429	217	4.90	3900	146.86
/news/local/suburbs/barrington/	1553	169	10.88	1030	154.80
http://featuresblogs.chicagotribune.com/features_julieshealthclub/	679	378	55.67	2	162.62
http://newsblogs.chicagotribune.com/towerticker/	378	270	71.43	2	163.84
/features/askamy/	4071	238	5.85	3768	168.75
http://leisureblogs.chicagotribune.com/taking-off/	165	90	54.55	2	172.70

INTEGRATED EXTERNAL LINKS AND JAVA SCRIPT

Some sections, such as Julie's Health Club [2] have a standard set of external links provided for each posting. For the Health Club features blog, each post has a list of keywords at the bottom with links to a Technorati search for each, along with a link to a Technorati search of all external blog posts linking to that blog post. This results in a high density of external linking in these content areas. Since these links are part of the overall "experience" of those content sections, and to help distinguish content sections that rely more or less heavily on these links, they are considered along with all other links for a given section. Further investigation of the site suggests that these "integrated" external links are present only on the Tribune's small number of hosted blogs and are not used in its primary content sections.

In addition, Health Club uses an embedded Java Script widget from Outbrain.com to display a list of external sponsored advertising links beneath each posting based on its content (known as "contextual advertising"). Downloading, executing, and evaluating the output of embedded Java Script blocks like this was beyond the scope of this study, but suggests that dynamic content generated on-the-fly when a user visits a page is on the rise among newspaper websites, especially for contextual sponsored advertising content.

PROCESS AND METHODOLOGY

At the start of the project, Bill Adee, former Digital Editor of the Chicago Tribune and current Vice President of Digital Development and Operations for the Chicago Tribune Media Group was contacted to see what details the Tribune itself could offer on velocity and churn rates. Email correspondence with Mr. Adee [3] suggested that 100% of the Chicago Tribune's print content is reproduced on its website, http://www.chicagotribune.com/, and that 15% of its web site's contents is only available online (such as blogs). When asked "How often are web-based articles updated, to add new information, correct errors, etc? What percentage of its online content is later edited to reflect updated information, as opposed to a new story being issued with the updated details", Mr. Adee responded that online stories were "updated often," especially on their partner Chicago Breaking News Center site (http://www.chicagobreakingnews.com/). In further correspondence, [4] Mr. Adee noted that there was no central RSS feed for the entire Chicago Tribune website because multiple content management systems are used to populate the site. He was unable to provide more detailed indicators on specific rates of change within each content section.

[2] http://featuresblogs.chicagotribune.com/features_julieshealthclub/

[3] Personal correspondence with Bill Adee 9/7/2010.

[4] Personal correspondence with Bill Adee 9/7/2010.

Given that the Chicago Tribune was unable to furnish specific detail on the velocity of change on its site, it was necessary to set up a targeted web crawling system to externally monitor the rate of new content posted to the site. The size of the Chicago Tribune website and total volume of content made it infeasible to attempt to download the full article contents of the entire site to monitor for individual article change to examine how often articles are updated after posting. A random selection of roughly 50 URLs from across sections, however, showed no content updates to those articles during a 24 hour period, so it is unclear how widespread post-updates to article content may be, and it is likely that this may be concentrated more heavily on their partner "breaking news" properties as opposed to the flagship http://www.chicagotribune.com/ domain. More thoroughly measuring this phenomenon would require mass-downloading of large portions of the Chicago Tribune's site on an hourly basis, which would entail significant resource consumption and thus was outside of the scope of this study.

Initially, it was hoped to use the site's RSS feeds as a quick mechanism to identify all new content across the site. Many news websites offer site-wide master or section-wide master RSS feeds that contain a complete list of all new content of the current day arranged in chronological order. Unfortunately, the Tribune's RSS feeds [5] do not function in this manner and appear to be ordered based on popularity, rather than publication/posting date. For example, the "Latest News" feed, which would presumably contain the most recent stories posted to the site, included only 10 stories at 9:45AM on November 8, 2010, with four of the stories dated the previous afternoon and the other six ranging from 4AM to 9:40AM November 8th. Other feeds examined showed similar ordering unrelated to post date and did not reflect the majority of content contained on the table of contents pages of each section. Thus, it was determined that the RSS feeds would not suffice to measure content velocity and a full-fledged external web crawl was necessary.

Given the size of the Chicago Tribune's site and the need for tight-interval regular snapshots of the site's structure, it was decided to use a "gateway page crawl," which downloads only the primary "table of contents" ("gateway") pages of each content section (such as the front page of the sports section and its links to all of the sports stories of the moment) and measures update velocity based on the average lifespan of a secondary page linked from that gateway page. In other words, the rate of change of the sports section is determined by downloading the main sports front page (/sports/) every 30 minutes and examining the list of links on that page. In conversations with CRL, it was decided to monitor only the Chicago Tribune's flagship property http://www.chicagotribune.com/ as opposed to its myriad partner properties in order to make results attributable solely to the Chicago Tribune's editorial controls, eliminating the role that partner organizations would play in the update cycles of affiliated websites.

Given the complex site layout of a typical Chicago Tribune web page, and the need for high accuracy on measuring rate of change, it was decided that a manual review of the site would be used to identify the gateway pages, rather than an automated content characterizer. Automated content characterizer crawlers use algorithms that balance the size of the page's textual content against the density of internal links and the overall site's link structure to determine if a page is likely a news page or a table of contents page. Such algorithms must crawl the entire site in order to generate baseline data and to find all possible gateway pages, which was not feasible for this study given the size of the Tribune's site.

Instead, a manual review of the Tribune site was used to identify all 105 gateway pages, such as http://www.chicagotribune.com/news/local/. A complete list of all pages is included in Appendix C. In many cases, both a root-level page such as http://www.chicagotribune.com/sports/ and a sub-level page like http://www.chicagotribune.com/sports/baseball/cubs/ were monitored

[5] http://www.chicagotribune.com/services/rss/

independently. A manual examination of root and sub-level pages showed that the Tribune displays only a small number of the top stories on its top-level pages, leaving the majority of its content in each section to be linked only from the sub-level pages. This required archiving top-level and sub-level gateway pages individually.

A customized crawling script was written to download the complete contents of all 105 gateway pages every 30 minutes and this was run for approximately one month (34 days), from 9/15/2010 through 10/19/2010, with a 3 second delay between each page fetch to reduce load on the Tribune's servers. A TCP timeout of 20 seconds was set for the crawling agent, but never triggered, indicating the Tribune's servers were responsive during all crawling runs. To ensure returned results matched those seen by a human user, robots.txt adherence was disabled in the crawlers and an HTTP User Agent field was transmitted mimicking a Macintosh user with the Google Chrome browser. Aggregate network latency resulted in an average of 43 snapshots per day, as opposed to the theoretical 48, for a total of 1,301 snapshots of each of the 105 gateway pages, or 136,605 total snapshots covering the entirety of the http://www.chicagotribune.com/ site over a 34 day period.

CONCLUSIONS AND FUTURE WORK

This study has examined a 34 day period of the Chicago Tribune, examining the rate of change and linking patterns of its flagship online property http://www.chicagotribune.com/. The lack of a master chronological list of new and updated content via RSS, portal page, or other mechanism is a significant hindrance not only to scholarly archival and analysis of the paper, but, more critically, for the Tribune itself. If Tribune readers must wade through multiple sub navigation pages to view all articles under a given content section, rather than subscribing to an RSS feed or other push mechanisms to receive a list of all new content each day, that makes it difficult for its readership to fully consume and utilize its content. This is by no means limited to the Tribune and reflects a larger problem with many news media websites.

Resource limitations prevented this study from examining the possibility of individual articles changing over time through post-updates as opposed to new articles being released, and no attempt was made to characterize how article content or positioning influenced linking lifespan, though these could be examined in future studies.

Overall, the picture of a modern mainstream news website reflects one heavily dependent on advertising revenue, integrating paid sponsored links throughout its web properties in excess of 83% of all links, with the average content link remaining on the site from 18 hours to 7 days. Content sections exhibit significant stratification in linking behavior, suggesting site-wide characterizations are less useful for understanding large news media websites and a more intricate content-specific characterization approach must be utilized. In conjunction with a previous longitudinal study of the Drudge Report, [6] this study demonstrates the considerable insights that can be derived from a media property through external monitoring and the utility of customized crawling activities in informing both media research and archival.

[6] http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2500/2235