



Link Attraction Factors

A study of the factors that influence the number of links a URL published to Digg's homepage accumulates.

By Dan Zarrella

<http://danzarrella.com>

2008

Introduction & Dataset

One of the most valuable aspects of being listed on Digg's homepage is that the story is seen by a large number of highly savvy social media users (including a large number of bloggers) who are likely to link to pages they find interesting on other sites that they participate in or own.

Studies have been done on the various characteristics of Digg stories and how they correlate to a story's popularity. We know that stories submitted by popular users are much more likely to go popular, and we know which times are the best for submitting stories.

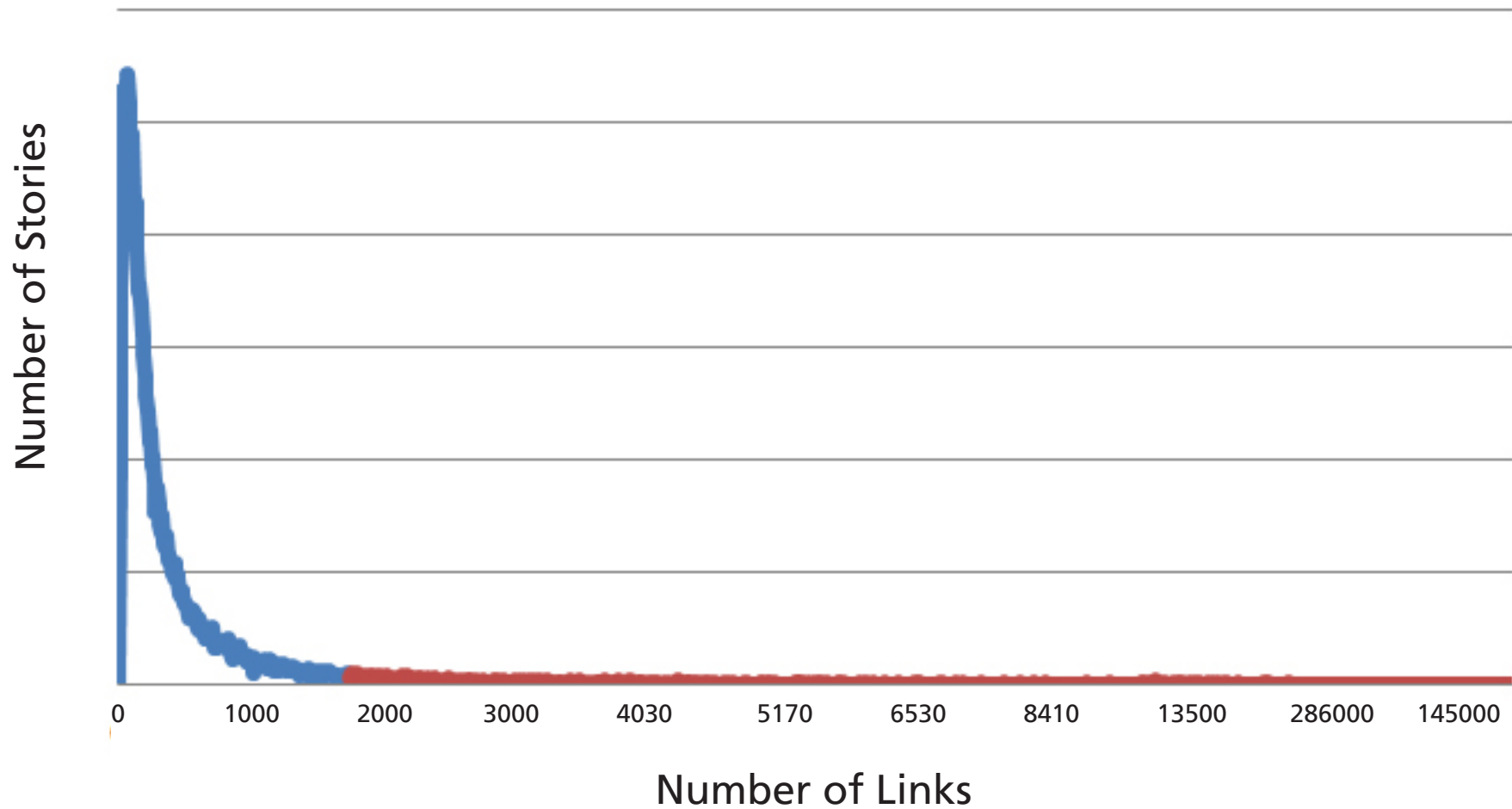
What hasn't been studied, however, is what happens to a story once it *makes* Digg's homepage; beyond the number of votes the story got, how "viral" did it go? By measuring the number of incoming links pointing to popular stories on Digg, I've discovered which characteristics increase or decrease the number of links a story gets.

Using Digg's API, I constructed a database of information on 33,322 of the 39,000 stories that became popular and were listed on Digg's homepage. I also indexed the textual content of the page for each story using scripts, CSS and HTML removal functions, as well as the number of incoming links the URL listed in the Digg story has. I used Yahoo's Site Explorer API to gather this data over the course of several weeks. The addition of the page content took the size of the database from about 16mb to over 500mb, so I also created an optimized table without the text content to use for calculations that did not require it.

The distribution graph of the number of incoming links the stories accumulate, has a very long tail to the right, at least partially due to the fact that some of the URLs seen on Digg's homepage were popular root domains (like apple.com). I calculated the "outer fence" to identify statistically extreme outliers with the standard $3 * IQR + Q3$ formula, resulting in an upper boundary of 1717 links. In my calculations I only use pages that have less than 1717 links, leaving 30,676 eligible stories, 92.06% of the total database.

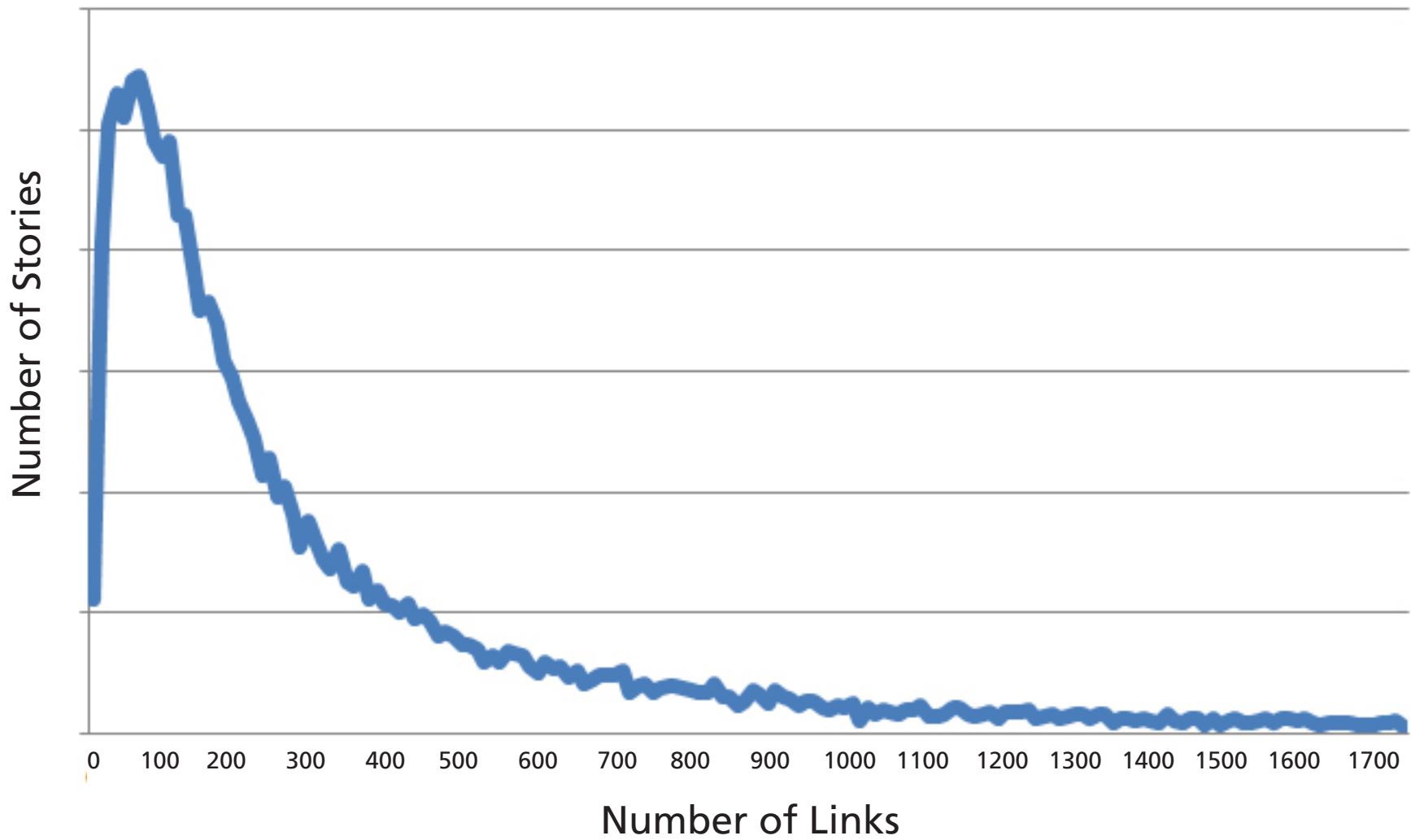
Stories removed are shown in red on this distribution graph:

With-Outliers Link Popularity Distribution



After removing the outliers the clean dataset's distribution looks like this:

Non-Outlier Link Popularity Distribution



Results

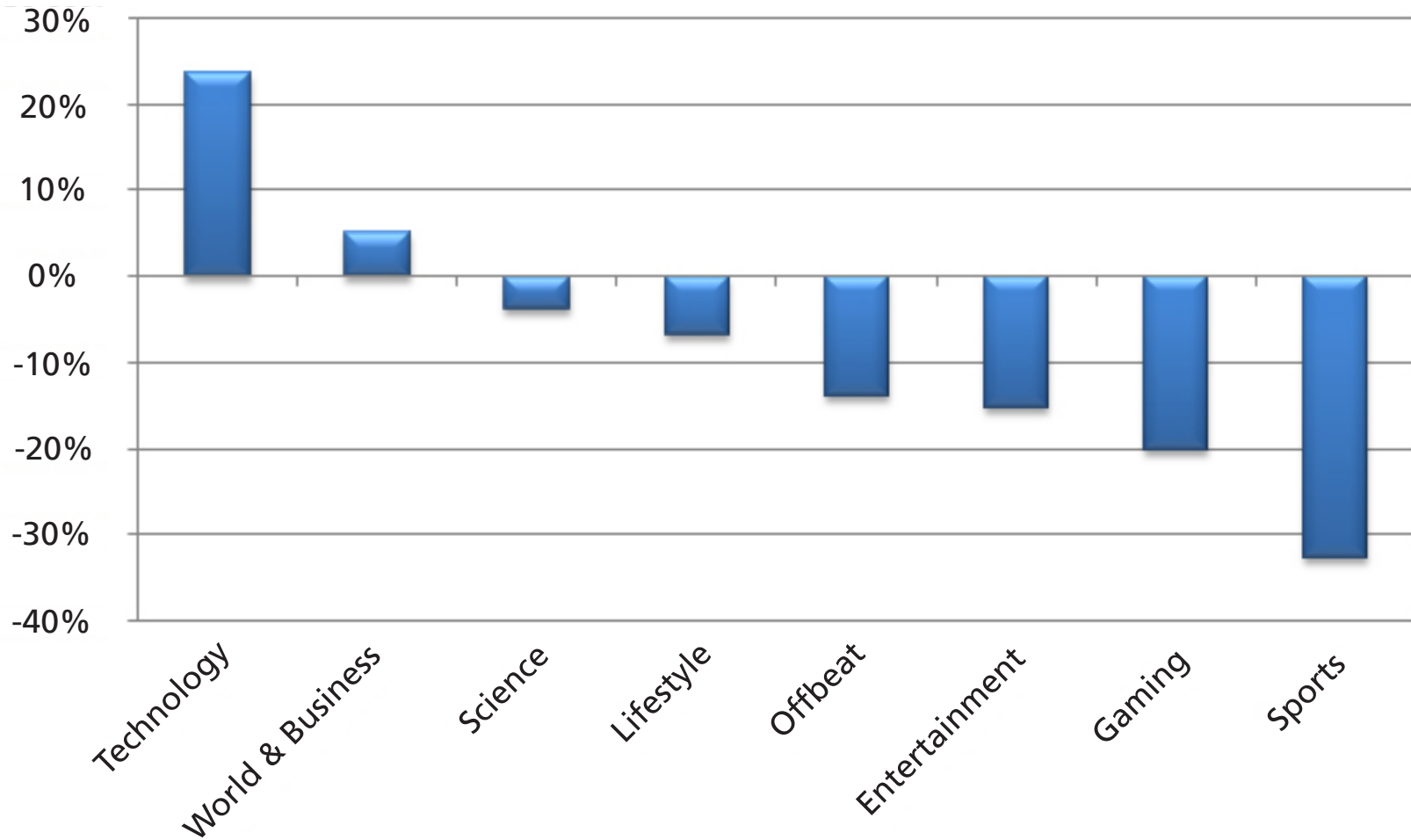
With this data set I first set out to test the quantitative values I had for each story by calculating their correlation with the number of incoming links each story has. There are 5 numeric values I could test this way: the number of diggs a story has; the number of comments; the popularity (measured by the number of popular stories submitted by the user) of the submitter; and the lengths of the title and description. In all five cases correlation was too low to be statistically significant (values above 0.5 are generally considered significant correlations). While the number of votes or comments or the popularity of the submitter may influence a story's chances of going popular, they have no measurable effect on the number of links a story gets once it is listed on Digg's homepage.

Criteria	Correlation
Number of Diggs	0.22649
Number of Comments	0.313194
Popularity of Submitter	0.02973
Length of Title	-0.00248
Length of Description	-0.00035

To test the effect of other, non-numerical factors, I first calculated the average number of incoming links the URLs in my database have: 299. I then found the average number of links a story matching certain criteria has and compared that number to the overall average.

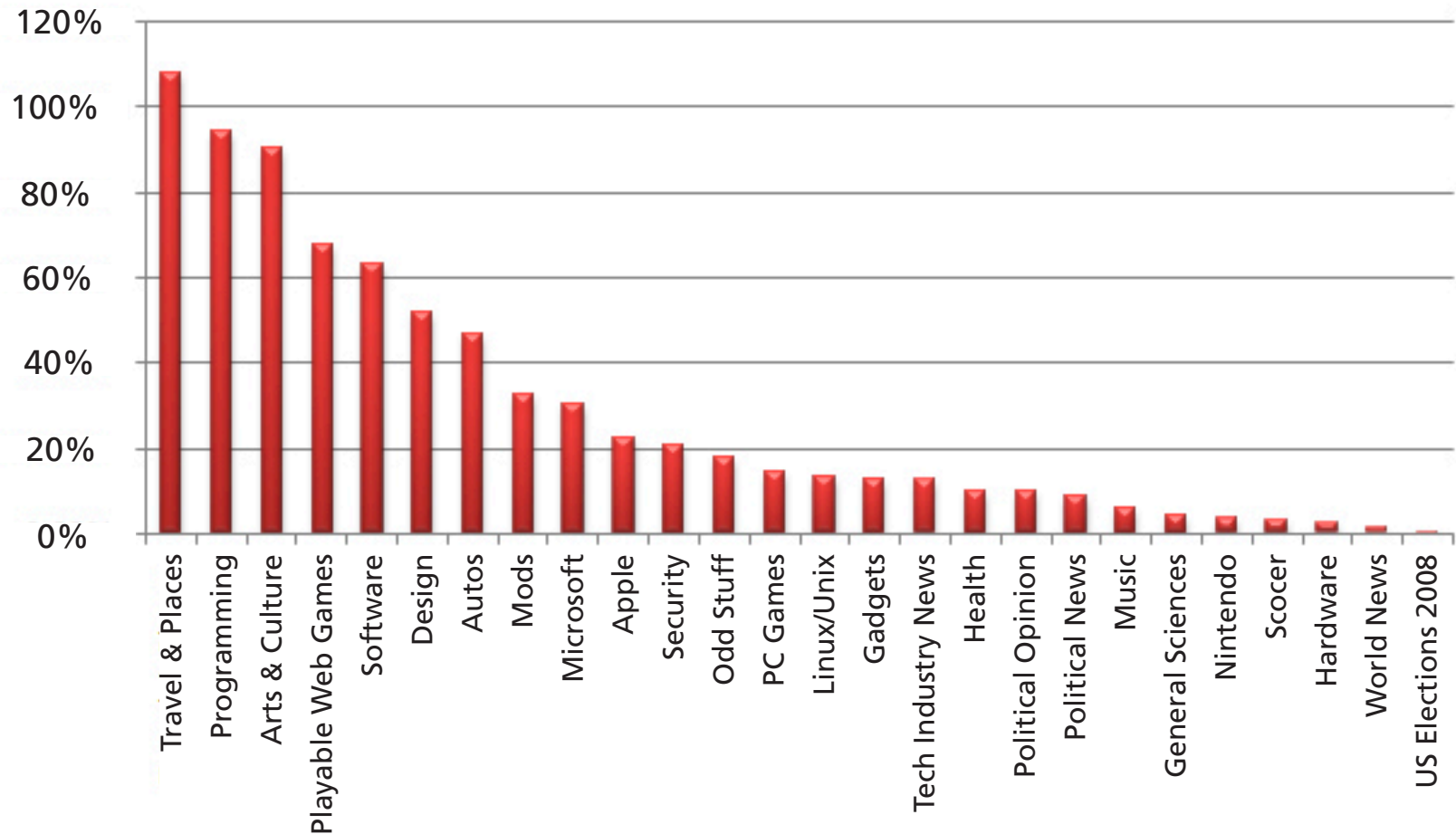
To visualize the effect these factors have on a URL's link accumulation I calculated and graphed the difference from the average for each criteria as a percentage. For instance, if a certain test shows that a type of story gets 598 links on average that is a 100% difference from the overall average. Stories that match this one get 100% more links on average than a normal story. If a factor causes stories to only have 150 links on average, that criteria has a -50% difference from the norm.

Average Links by Container



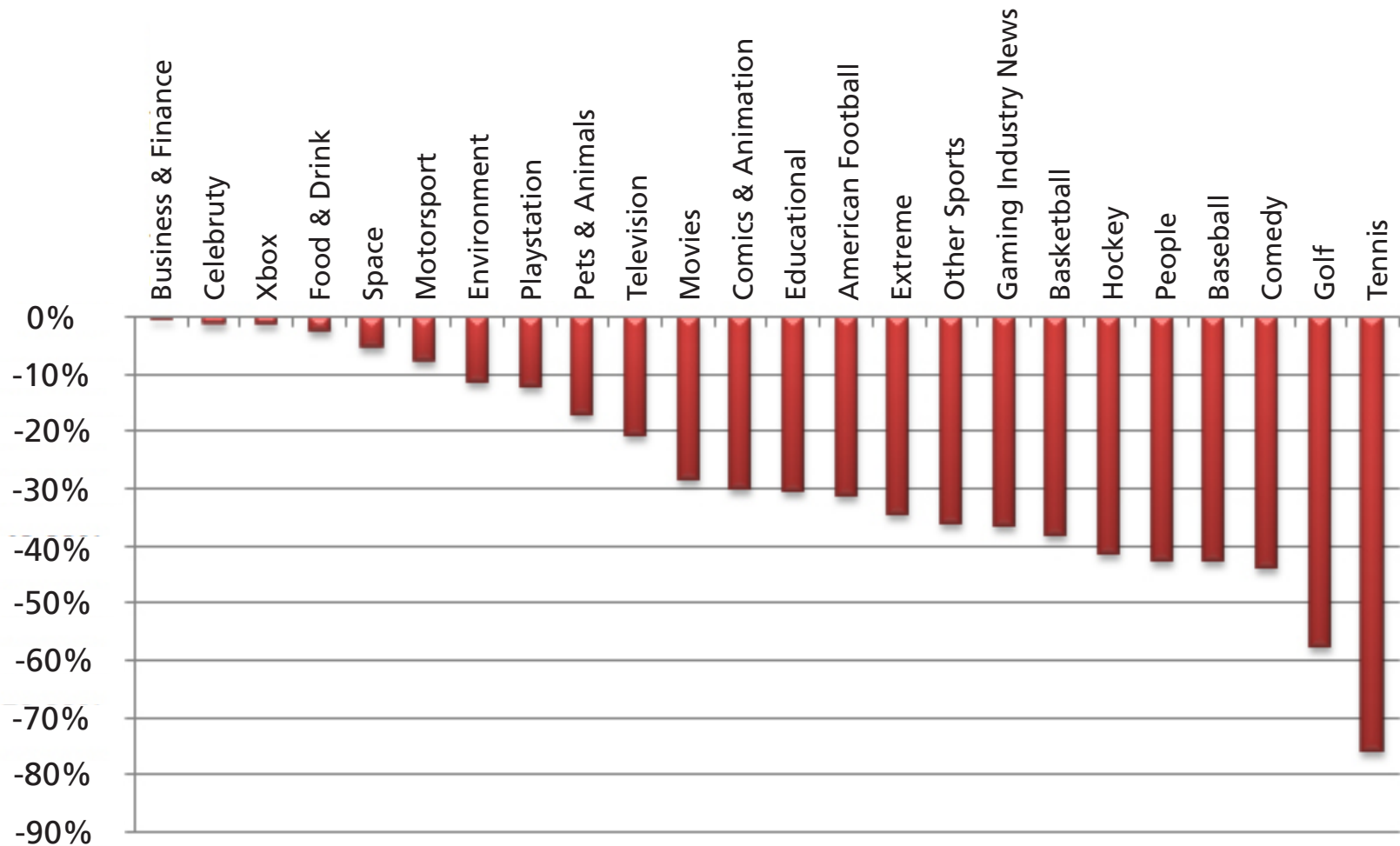
The first criteria I tested was the container the story was listed in on Digg. The above graph shows reasonable expectable results, with stories in the Technology container accumulating 23.66% more links than the average and stories listed under Sports receiving 32.89% less. One surprise here is that URLs listed in the Gaming container (typically thought of as very "Digg-like") got 20.39% less links than normal pages.

Average Links by Topic



Next I tested the topic the stories were listed in, since there are so many of them I had to break the graph down into those that seemed to have a positive influence on the number of incoming links a URL received and those than seemed to have a negative influence. Similarly unsurprising results here, with technical and “geek”-oriented topics (technology, engineering) dominating the high-link-popularity side of the graph. Travel & Places took first place however, and stories listed in that category received 107.75% more links than average.

Average Links by Topic



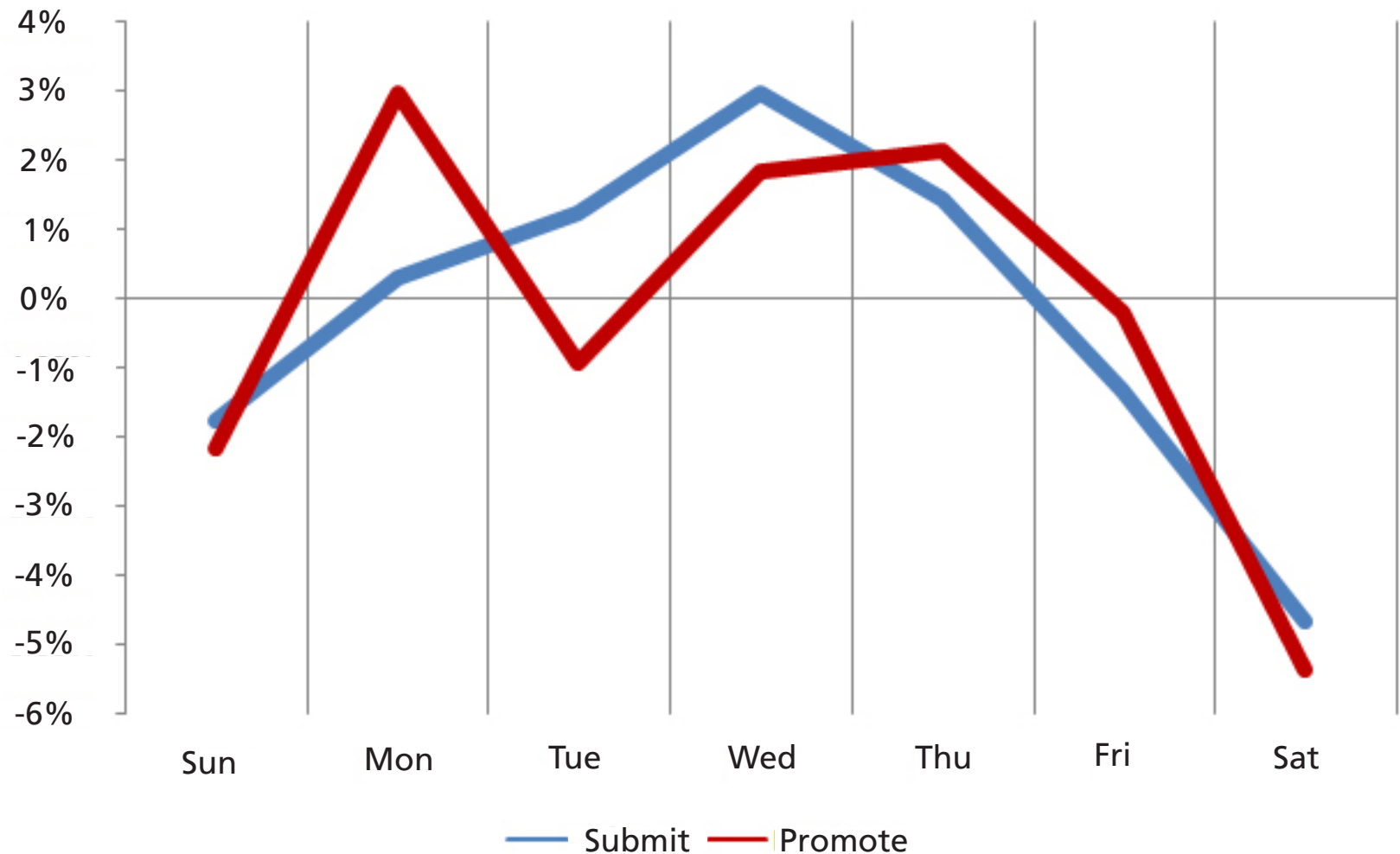
On the low-link-popularity side of the graph, stories listed in sports and entertainment topics are seen to garner fewer links than the rest of Digg's topics. Tennis performs the worst, with stories listed there getting 76.25% less links, followed closely by Golf at 57.63% less.

Average Links by Topic

Topic	Average Links
Travel & Places	621.18
Programming	581.82
Art & Culture	568.82
Playable Web Games	501.58
Software	487.47
Design	453.40
Autos	439.58
Mods	396.02
Microsoft	390.19
Apple	365.88
Security	361.64
Odd Stuff	352.04
PC Games	338.39
Linux/Unix	336.64
Tech Industry News	336.33
Health	329.36
Political Opinion	328.97
Political News	325.62
Music	317.25
General Sciences	312.42
Nintendo	309.30
Soccer	308.43
Hardware	306.05
World News	302.52

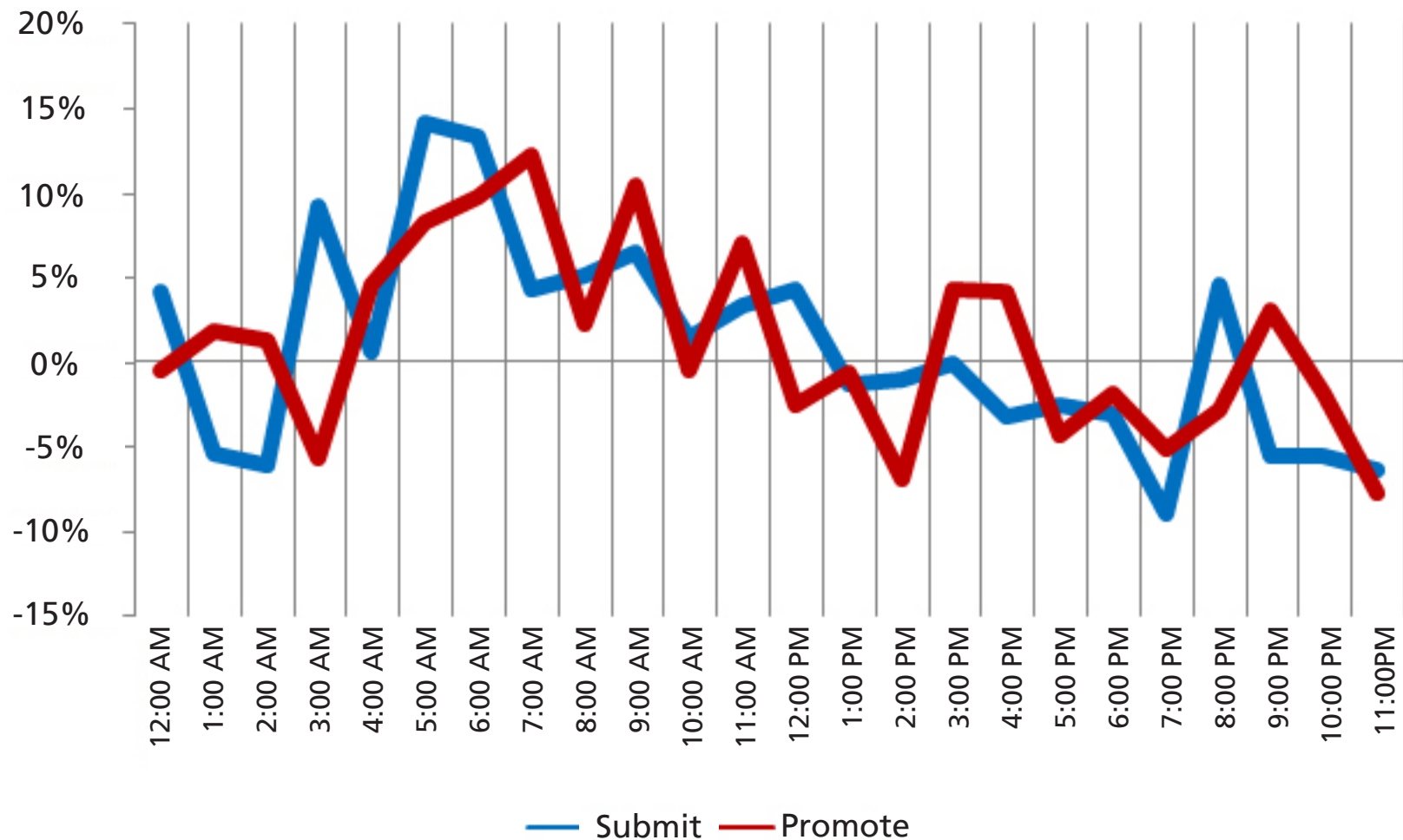
Topic	Average Links
US Elections	299.61
Business & Finances	297.15
Celebrity	295.44
Xbox	294.61
Foot & Drink	291.00
Space	282.97
Motorsport	276.16
Environment	264.50
Playstation	262.01
Pets & Animals	247.82
Television	236.99
Movies	213.82
Comics & Animation	208.35
Educational	207.19
American Football	205.75
Extreme	195.20
Other Sports	191.03
Gaming Industry News	189.76
Basketball	184.74
Hockey	174.71
People	171.32
Baseball	171.25
Comedy	167.82
Golf	126.69
Tennis	71.00

Average Links by Day of Week



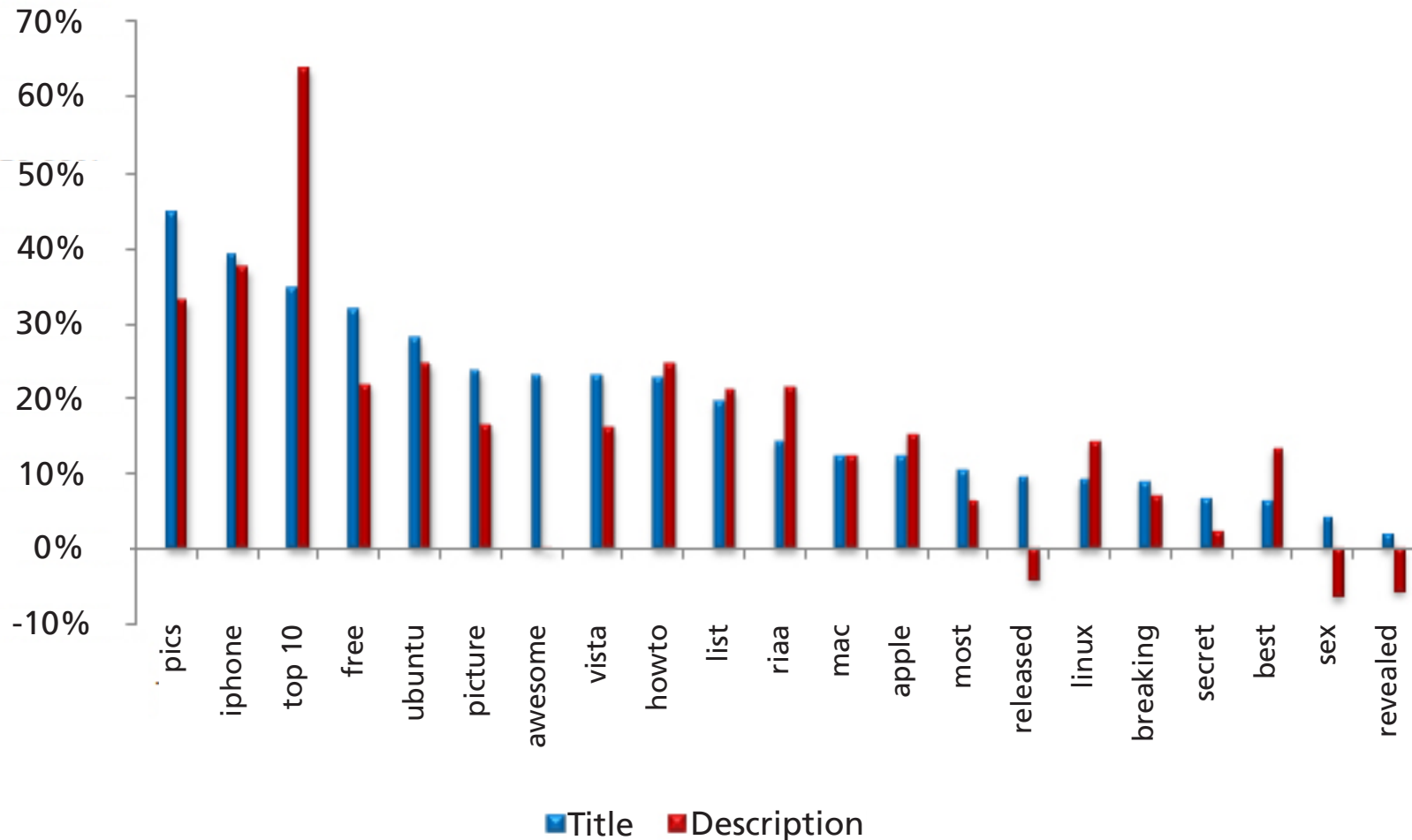
I then tested the day of the week the stories were submitted to Digg and promoted to its homepage on. Here we have a clear pattern showing that stories submitted and promoted during the business week (and especially in the beginning of it) tend to get more links than those submitted or promoted on weekends. The differences from the average based on day of week are small compared to criteria like topic or container, but the case for weekday submission and promotion is made.

Average Links by Hour (PST)



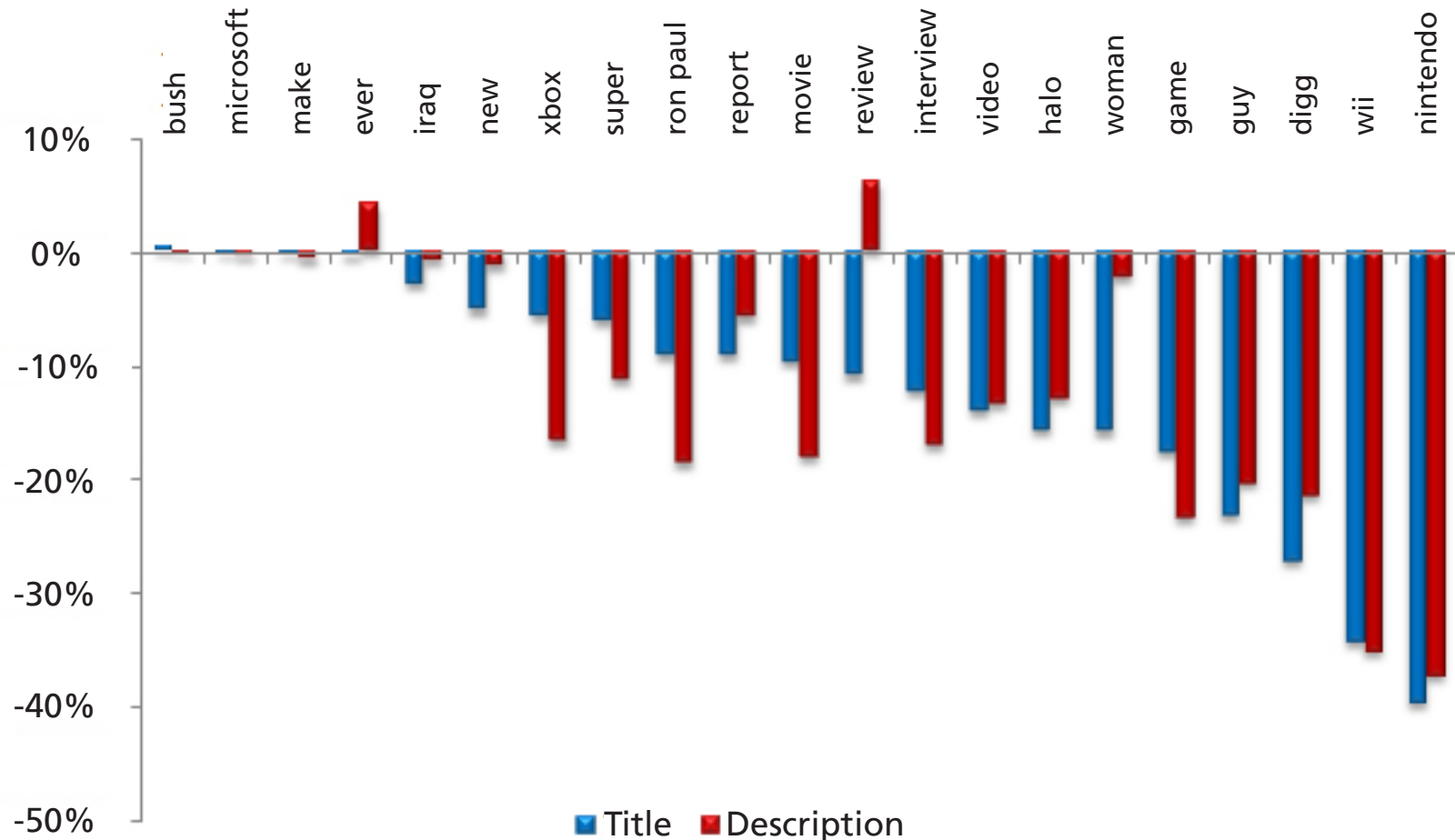
After looking at days of the week I looked at hour of the day the stories were submitted and promoted. Again, we see a clear pattern of business-hour (especially early in the day) submissions and promotions getting more links than evening and night. However, the difference from the average here is larger than I found in day of the week. Stories that are submitted or promoted between the hours of 4am and 9am PST (9am and 1pm EST) get more links than those that are submitted or promoted outside of that period.

Average Links by Keyword



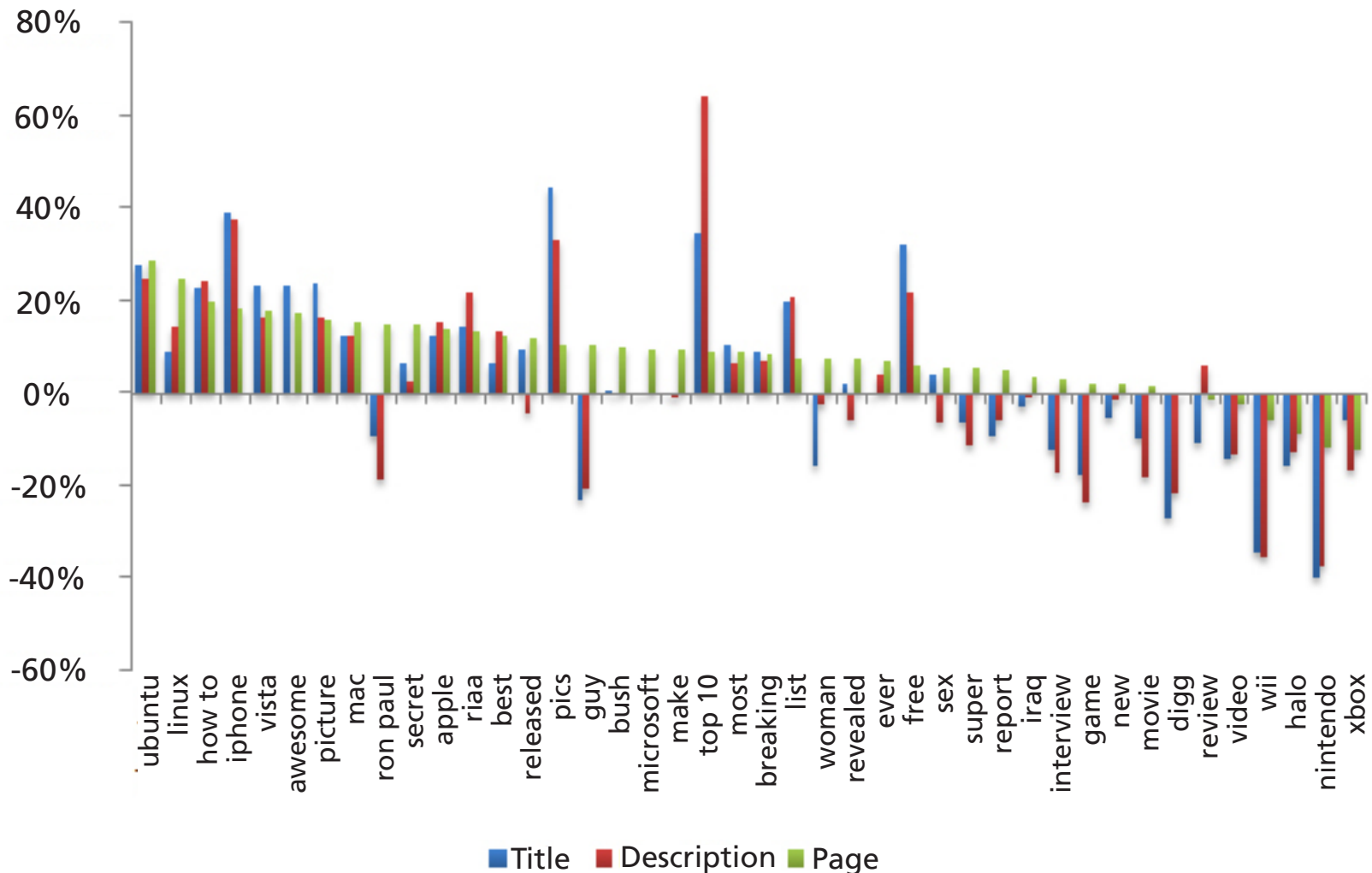
I then began testing if the occurrence of certain popular keywords in the title or description effected the average number of incoming links the stories received. Again, due to the number of words I tested I had to split the graph into positive-effect keywords and negative-effect keywords. In the top half we see keywords like "how to," "top 10," "list," "free" and "pics," as well as superlatives commonly used with "top 10" and list-type stories. We also see that certain brand names can have a positive effect on the number of incoming links a story received.

Average Links by Keyword



On the negative-effect side of the graph, we see one major surprise: stories that mention the word "digg" in their title received 27.30% less links and stories that contained it in the description field got 21.56% less links than the average story. Self-reference may help a story become popular and reach Digg's homepage, but it reduces the number of links it will get once it's there. Here we also see the words "report," "interview" and "review" all have a negative impact on link acquisition. Stories containing these words contrast with positive-effect keywords like "how to," "top 10," "list," "free" and "pics" in that the negative-effect words indicate longer, heavier and more textual content, while the positive ones indicate quick, well-chunked reads.

Average Links by Keyword



If you compare the data for keyword occurrence in the title and/or description to occurrence in the textual content on the story's target URL, we see that while title and description tend to have similar influence, the effect had by on-page occurrence of the words varies a good deal. Above is a combined keyword occurrence graph, sorted by the average links accumulated by URLs with the keyword in the content of the page. I've also created two tools which use this keyword occurrence data, one displays the effect a keyword has on stories using it in their title and description, and the other analyzes entire potential titles.

Average Links by Keyword

Average links when keyword occurs in...

Keyword	Desc	Title	Page
ubuntu	372.51	382.31	384.86
linux	341.13	326.08	372.37
how to	372.03	366.85	358.51
iphone	410.59	415.45	354.06
vista	347.32	367.90	352.61
awesome	298.25	368.19	350.39
picture	347.84	369.60	345.54
mac	335.99	335.93	344.14
ron paul	243.08	271.50	344.03
secret	306.34	318.67	342.69
apple	344.50	335.63	340.82
riaa	363.28	341.75	339.62
best	338.83	317.11	335.38
released	286.18	327.65	334.56
pics	397.57	432.21	330.54
guy	237.88	229.27	330.26
bush	298.96	299.89	328.66
microsoft	297.69	298.30	327.65
make	296.73	298.15	327.55
top 10	490.02	402.85	326.36

Average links when keyword occurs in...

Keyword	Desc	Title	Page
most	318.24	330.08	325.03
breaking	319.54	325.38	323.75
list	361.74	357.38	320.99
woman	292.11	251.40	320.99
revealed	281.98	304.90	320.84
ever	311.44	297.61	319.37
free	364.38	394.24	317.45
sex	279.49	311.34	315.29
super	265.14	280.60	314.72
report	281.42	271.45	314.33
iraq	296.52	290.24	309.37
interview	247.86	262.11	307.90
game	228.66	246.01	305.06
new	295.07	283.32	304.62
movie	244.69	269.82	303.49
digg	234.55	217.37	299.29
review	317.49	266.09	294.87
video	258.76	256.70	291.47
wii	193.65	196.20	281.07
halo	260.07	251.85	271.98
nintendo	187.16	179.95	264.18
xbox	248.90	281.77	262.96