

DIGITAL MARKETING

Eric Enge Interviews Google's Matt Cutts

by **Eric Enge** on October 8th, 2007 | ~ 28 minute read



The Digital Essentials, Part 3

Developing a robust digital strategy is both a challenge and an opportunity. Part 3 of the Digital Essentials guide series explores five of the essential technology-driven experiences customers expect, which you may be missing or not fully utilizing.

[Get the Guide](#)

 [Picture of Matt Cutts](#)

Matt Cutts

Matt Cutts joined Google as a Software Engineer in January 2000. Before Google, he was working on his Ph.D. in computer graphics at the University of North Carolina at Chapel Hill. He has an M.S. from UNC-Chapel Hill and B.S. degrees in both mathematics and computer science from the University of Kentucky.

Matt wrote SafeSearch, which is Google's family filter. In addition to his experience at Google, Matt held a top-secret clearance while working for the Department of Defense, and he's also worked at a game engine company. He claims that Google is the most fun by far.

Matt currently heads up the Webspam team for Google. Matt talks about webmaster-related issues on his [blog](#).

Interview Transcript

Eric Enge: Let's talk about different kinds of link encoding that people do, such as links that go through JavaScript or some sort of redirect to link to someone, yet the link actually does represent an endorsement. Can you say anything about the scenarios in which the link is actually still recognized as a link?

Matt Cutts: A direct link is always the simplest, so if you can manage to do a direct link that's always very helpful. There was an interesting proposal recently by somebody who works on FireFox or for Mozilla I think, which was the idea of a ping attribute, where the link can still be direct, but the ping could be used for tracking purposes. So, something like that could certainly be promising, because it lets you keep the direct nature of a link while still sending a signal to someone. In general, Google does a relatively good job of following the 301s, and 302s, and even Meta Refreshes and JavaScript. Typically what we don't do would be to follow a chain of redirects that goes through a robots.txt that is itself forbidden.

Eric Enge: Right.

Matt Cutts: I think in many cases we can calculate the proper or appropriate amount of PageRank, or Link Juice, or whatever you want to call it, that should flow through such links.

Eric Enge: Right. So, you do try to track that and provide credit.

Matt Cutts: Yes.



Eric Enge

Eric Enge is part of the Digital Marketing practice at Perficient. He designs studies and produces industry-related research to help prove, debunk, or evolve assumptions about digital marketing practices and their value. Eric is a writer, blogger, researcher, teacher, and keynote speaker and panelist at major industry conferences. Partnering with several other experts, Eric served as the lead author of The Art of SEO.

[More from this Author](#)

Subscribe to the Weekly Blog Digest:

[Sign Up](#)

Categories

[Digital Marketing](#)
[SEO](#)

Follow Us



Eric Enge: Right. Let's talk a bit about the various uses of NoIndex, NoFollow, and Robots.txt. They all have their own little differences to them. Let's review these with respect to 3 things: (1) whether it stops the passing of link juice; (2) whether or not the page it still crawled; and: (3) whether or not it keeps the affected page out of the index.

Matt Cutts: I will start with robots.txt, because that's the fundamental method of putting up an electronic no trespassing sign that people have used since 1996. Robots.txt is interesting, because you can easily tell any search engine to not crawl a particular directory or even a page, and many search engines support variants such as wildcards, so you can say don't crawl *.gif, and we won't crawl any GIFs for our image crawl.

We even have additional standards such as Sitemap Support, so you can say here's a link to where my Sitemap is can be found. I believe the only robots.txt extension in common use that Google doesn't support is the crawl-delay. And, the reason that Google doesn't support crawl-delay is because way too many people accidentally mess it up. For example, they set crawl-delay to a hundred thousand, and, that means you get to crawl one page every other day or something like that.

We have even seen people who set a crawl-delay such that we'd only be allowed to crawl one page per month. What we have done instead is provide throttling ability within Webmaster Central, but crawl-delay is the inverse; its saying crawl me once every "n" seconds. In fact what you really want is host-load, which lets you define how many Googlebots are allowed to crawl your site at once. So, a host-load of two would mean, 2 Googlebots are allowed to be crawling the site at once.

Now, robots.txt says you are not allowed to crawl a page, and Google, therefore, does not crawl pages that are forbidden in robots.txt. However, they can accrue PageRank, and they can be returned in our search results.

In the early days, lots of very popular websites didn't want to be crawled at all. For example, eBay and the New York Times did not allow any search engine, or at least not Google to crawl any pages from it. The Library of Congress had various sections that said you are not allowed to crawl with a search engine. And so, when someone came to Google and they typed in eBay, and we haven't crawled eBay, and we couldn't return eBay, we looked kind of suboptimal. So, the compromise that we decided to come up with was, we wouldn't crawl you from robots.txt, but we could return that URL reference that we saw.

Eric Enge: Based on the links from other sites to those pages.

Matt Cutts: Exactly. So, we would return the un-crawled reference to eBay.

Eric Enge: The classic way that shows it you just list the URL, no description, and that would be the entry that you see in the index, right?

Matt Cutts: Exactly. The funny thing is that we could sometimes rely on the ODP description (Editor: also known as [DMOZ](#)). And so, even without crawling, we could return a reference that looked so good that people thought we crawled it, and so that caused a little bit of earlier confusion. So, robots.txt was one of the most long standing standards. Whereas for Google, NoIndex means we won't even show it in our search results.

So, with robots.txt for good reasons we've shown the reference even if we can't crawl it, whereas if we crawl a page and find a Meta tag that says NoIndex, we won't even return that page. For better or for worse that's the decision that we've made. I believe Yahoo and Microsoft might handle NoIndex slightly differently which is a little unfortunate, but everybody gets to choose how they want to handle different tags.

Eric Enge: Can a NoIndex page accumulate PageRank?



Matt Cutts: A NoIndex page can accumulate PageRank because the links are still followed outwards from a NoIndex page.

Eric Enge: So, it can accumulate and pass PageRank.

Matt Cutts: Right, and it will still accumulate PageRank, but it won't be showing in our Index. So, I wouldn't make a NoIndex page that itself is a dead end. You can make a NoIndex page that has links to lots of other pages.

For example, you might want to have a master Sitemap page and for whatever reason NoIndex that, but then have links to all your sub Sitemaps.

Eric Enge: Another example is if you have pages on a site with content that from a user point of view you recognize that it's valuable to have the page, but you feel that is too duplicative of content on another page on the site

That page might still get links, but you don't want it in the Index and you want the crawler to follow the paths into the rest of the site.

Matt Cutts: That's right. Another good example is, maybe you have a login page, and everybody ends up linking to that login page. That provides very little content value, so you could NoIndex that page, but then the outgoing links would still have PageRank.

Now, if you want to you can also add a NoFollow metatag, and that will say don't show this page at all in Google's Index, and don't follow any outgoing links, and no PageRank flows from that page. We really think of these things as trying to provide as many opportunities as possible to sculpt where you want your PageRank to flow, or where you want Googlebot to spend more time and attention.

Eric Enge: Does the NoFollow metatag imply a NoIndex on a page?

Matt Cutts: No. The NoIndex and NoFollow metatags are independent. The NoIndex metatag, for Google at least, means don't show this page in Google's index. The NoFollow metatag means don't follow the outgoing links on this entire page.

Eric Enge: How about page A links to page B and page A has a NoFollow metatag, or the link to page B has a NoFollow on the link. Will page B still be crawled?

Matt Cutts: It won't be crawled because of the links found on page A. But if some other page on the web links to page B, then we might discover page B via those other links.

Eric Enge: Right. So there are two levels of NoFollow. There is the attribute on a link, and then there is the metatag, right.

Matt Cutts: Exactly.

Eric Enge: What we've been doing is working with clients and telling them to take pages like their about us page, and their contact us page, and link to them from the home page normally, without a NoFollow attribute, and then link to them using NoFollow from every other page. It's just a way of lowering the amount of link juice they get. These types of pages are usually the highest PageRank pages on the site, and they are not doing anything for you in terms of search traffic.

Matt Cutts: Absolutely. So, we really conceive of NoFollow as a pretty general mechanism. The name, NoFollow, is meant to mirror the fact that it's also a metatag. As a metatag NoFollow means don't crawl any links from this entire page.

NoFollow as an individual link attribute means don't follow this particular link, and so it really just extends that granularity down to the link level.

We did an interview with Rand Fishkin over at [SEOMoz](#) where we talked about the fact that NoFollow was a perfectly acceptable tool to use in addition to robots.txt. NoIndex and NoFollow as a metatag can change how Googlebot crawls your site. It's important to realize that typically these things are more of a second order effect. What matters the most is to have a great site and to make sure that people know about it, but, once you have a certain amount of PageRank, these tools let you choose how to develop PageRank amongst your pages.

Eric Enge: Right. Another example scenario might be if you have a site and discover



that you have a massive duplicate content problem. A lot of people discover that because something bad happened. They want to act very promptly, so they might NoIndex those pages, because that will get it out of the index removing the duplicate content. Then, after it's out of the index, you can either just leave in the NoIndex, or you can go back to robots.txt to prevent the pages from being crawled. Does that make sense in terms of thinking about it?

Matt Cutts: That's at the level where I'd encourage people to try experiments and see what works best for them because we do provide a lot of ways to remove content.

Matt Cutts: There's robots.txt.

Eric Enge: Sure. You can also use the URL removal tool too.

Matt Cutts: The URL removal tool is another way to do it. Typically, what I would probably recommend most people do, instead of going the NoIndex route, is to make sure that all their links point to the version of the page that they think is the most important. So, if they have got two copies, you can look at the backlinks within our Webmaster Central, or use Yahoo, or any other tools to explore it, and say what are the backlinks to this particular page, why would this page be showing up as a duplicate of this other page? All the backlinks that are on your own page are very easy to switch over to the preferred page. So, that's a very short term thing that you can do, and that only usually takes a few days to go into effect. Of course, if it's some really deep URL, they could certainly try the experiment with NoIndex. I would probably lean toward using optimum routing of links as the first line of defense, and then if that doesn't solve it, look at or consider using NoIndex.

Eric Enge: Let's talk about non-link based algorithms. What are some of the things that you can use as signals that aren't links to help with relevance and search quality? Also, can you give any indication about such signals that you are already using?

Matt Cutts: I would certainly say that the links are the primary way that we look at things now in terms of reputation. The trouble with something like other ways of measuring reputation is that the data might be sparse. Imagine for example that you decided to look at all the people that are in various yellow page directories, across the web, for the list of their address, or stuff like that. The problem is, even a relatively savvy business with multiple locations might not think to list all their business addresses.

A lot of these signals that we look at to determine quality or to help to determine reputation can be noisy. I would convey Google's basic position as that we are open to any signals that could potentially improve quality. If someone walked up to me and said, the phase of the moon correlates very well with the site being high quality, I wouldn't rule it out, I wouldn't take it off the table, I would do the analysis and look at it.

Eric Enge: And, there would be SEOs out there trying to steer the course of the moon.

Matt Cutts: It's funny, because if you remember [Webmaster World](#) used to track updates on the Google Dance, and they had a chart because it was roughly on a 30-day schedule. When a full moon came around people started to look for the Google Dance to happen.

In any event, the trouble is any potential signal could be sparse, or could be noisy, and so you have to be very careful about considering signal quality.

Eric Enge: Right. So, an example of a noisy signal might be the number of Gadgets installed from a particular site onto people's iGoogle homepage.

Matt Cutts: I could certainly imagine someone trying to spam that signal, creating a bunch of accounts, and then installing a bunch of their own Gadgets or something like that. I am sad to say you do have to step into that adversarial analysis phase where you say okay, how would someone abuse this anytime you are thinking about some new network signal.

Eric Enge: Or bounce rate is another thing that you could look at. For example, someone does a search and went to a site, and then they are almost immediately back at the Google search results page clicking on a different link or doing a very similar search. You could use that as a signal potentially.



Matt Cutts: In theory. I don't think we typically don't confirm or deny whether we'd use any given particular signal. It is a tough problem because something that works really well in one language might not work as well in another language.

Eric Enge: Right. One of the problems with bounce rate is that with the web moving so much more towards just give them answer now. For example, if you have a Gadget, you want the answer in the Gadget. If you use subscribed links, you want the answer in the subscribed links. When you get someone to your site, there is something to be said for giving them the answer they are looking for immediately, and they might see it and immediately leave (and you get the branding/relationship benefit of that.)

In this case, it's actually a positive quality signal rather than a negative quality signal.

Matt Cutts: Right. You could take it even further and help people get the answer directly from a snippet on the search engine results page, and so they didn't click on the link at all. There are also a lot of weird corner cases, you have to consider anytime you are thinking about a new way to try to measure quality.

Eric Enge: Right, indeed. What about the toolbar data, and Google analytics data?

Matt Cutts: Well, I have made a promise that my Webspam team wouldn't go to the Google Analytics group and get their data and use it. Search quality or other parts of Google might use it, but certainly, my group does not. I have talked before about how data from the Google toolbar could be pretty noisy as well. You can see an example of how noisy this is by installing Alexa. If you do, you see a definite skew towards Webmaster sites. I know that my site does not get as much traffic as many other sites, and it might register higher on Alexa because of this bias.

Eric Enge: Right. A site owner could start prompting people to install the Google toolbar whenever they come to their site.

Matt Cutts: Right. Are you sure you don't want to install a Google toolbar, Alexa, and why not throw in Compete and Quantcast? I am sure Webmasters are a little savvier about that, then the vast majority of sites. So, it's interesting to see that there is usually a Webmaster bias or SEO bias, with many of these usage-based tools.

Eric Enge: Let's move on to the hidden text. There are a lot of legitimate ways people can use hidden text, and, there are of course ways they can illegitimately use hidden text.

It strikes me that many of these kinds of hidden text are hard to tell apart. You can have someone who is using a simple CSS display:none scenario, and perhaps they are stuffing keywords, but maybe they do this with a certain amount of intelligence, making it much harder to detect than the site [you recently wrote about](#). So, tell me about how you deal with these various forms of hidden text?

Matt Cutts: Sure. I don't know if you saw the blog post recently where somebody tried to layout many different ways to do hidden text and ended up coming up with [14 different techniques](#). It was kind of a fun blog post, and I forwarded it to somebody and said: "hey, how many do we check for"? There were at least a couple that wasn't strictly hidden text, but it was still an interesting post.

Certainly, there are some cases where people do deceptive or abusive things with hidden text, and, those are the things that get our users most angry. If your web counter displays a single number, that's just a number, a single number. Probably, users aren't going to complain about that to Google, but if you have 4,000 stuffed words down at the bottom of the page that's clearly the sort of thing that if the user realizes it's down at the bottom of the page, they get angry about it.

Interestingly enough they get angry about it whether it helped or not. I saw somebody do a [blog post recently](#) that had a complaint about six words of hidden text and how they showed up for the query "access panels". In fact, the hidden text didn't even include the word access panels, just a variant of that phrase.

Eric Enge: I am familiar with the post.

Matt Cutts: I thought it was funny that this person had gotten really offended at six words of hidden text, and complained about a query which had only one word out of



the two. So, you do see a wide spectrum, where people really dislike a ton of hidden keyword stuff. They might not mind a number, but even with as little as six words, we do see complaints about that. So, our philosophy has tried to be not to find any false positives, but to try to detect stuff that would qualify as keyword stuffing, or gibberish, or stitching pages, or scraping, especially put together with hidden text.

We use a combination of algorithmic and manual things to find hidden text. I think Google is alone in notifying Webmasters about relatively small incidences of hidden text because that is something where we'll try to drop an email to the Webmaster, and alert them in Webmaster Central. Typically, you'd get a relatively short-term penalty from Google, maybe 30 days for something like that. But, that can certainly go up over time, if you continue to leave the text on your page.

Eric Enge: Right. So, a 30 days penalty in this sort of situation, is that getting removed from the index, or just de-prioritizing their rankings?

Matt Cutts: Typically with hidden text, a regular person can look at it and instantly tell that it is hidden text. There are certainly great cases you could conjure up where that is not the case, but the vast majority of the time it's relatively obvious. So, for that, it would typically be a removal for 30 days.

Then, if the site removes the hidden text or does a reconsideration request directly after that it could be shorter. But, if they continue to leave up that hidden text then that penalty could get longer.

We have to balance what we think is best for our users. We don't want to remove resources from our index longer than we need to it, especially if it's relatively high quality. But, at the same time, we do want to have a clean index and protect the relevance of it.

Eric Enge: Right. Note that Accesspanels.net has removed the hidden text and they are still ranked no. 1 in Google for the query "access panels".

I checked this a few days ago, and the hidden text had been removed. The site has a "last updated" indicator at the bottom of the page, and it was just the day before I checked.

Matt Cutts: That's, we probably shouldn't get into too much detail about individual examples, but that one got our attention and is winding its way through the system.

Eric Enge: Right. When reporting web spam, writing a blog post in a very popular blog and getting a lot of peoples' attention to it is fairly effective. But, also Webmaster tools allows you to do submissions there, and it gets looked at pretty quickly, doesn't it?

Matt Cutts: It does. We try to be pretty careful about the submissions that we get to our spam report form. We've always been clear that the first and primary goal with those is to look at those to try to figure out how to improve our algorithmic quality. But, it is definitely the case that we look at many of those manually as well, and so you can imagine if you had a complaint about a popular site because it had hidden text, we would certainly check that out. For example, the incident we discussed just a minute ago, someone had checked it out earlier today and noticed the hidden text is already gone.

We probably won't bother to put a postmortem penalty on that one, but, it's definitely the case that we try to keep an open mind and look at spam reports, and reports across the web not just on big blogs, but also on small blogs.

We try to be pretty responsive and adapt relatively well. That particular incident was interesting, but I don't think that the text involved was actually affecting that query since it was different words.

Eric Enge: Right. Are there hidden text scenarios that are harder for you to discern whether or not they are spam versus something like showing just part of a site's terms and conditions on display, or dynamic menu structures? Are there any scenarios where it's really hard for you to tell whether it is spam or not?

Matt Cutts: I think Google handles the vast majority of idioms like dynamic menus and things like that very well. In almost all of these cases, you can construct interesting examples of hidden text. Hidden text, like many techniques, is on a



spectrum. The vast majority of the time, you can look and you can instantly tell that it is malicious, or it's a huge amount of text, or it's not designed for the user. Typically we focus our efforts on the most important things that we consider to be a high priority. The keyword stuffed pages with a lot of hidden text, we definitely give more attention.

Eric Enge: Right.

Matt Cutts: So, we do view a lot of different spam or quality techniques as being on a spectrum. And, the best advice that I can give to your readers is probably to ask a friend to look at their site, it's easy to do a Ctrl+A, it's easy to check things with cascading style sheets off, and stick to the more common idioms, the best practices that lots of sites do rather than trying to do an extremely weird thing that could be misinterpreted even by a regular person as being spamming.

Eric Enge: Fair enough. There was a scenario that we reported a long while ago involving a site that was buying links, and none of those links were labeled.

There was a very broad pattern of it, but the one thing that we noticed and thought was a potential signal was that the links were segregated from the content, they were either in the right rail or the left rail, and the main content for the pages were in the middle. The links weren't integrated with the site, there was no labeling of them, but they were relevant. That's an example of a subtle signal, so, it must be challenging to think about how much to make out of that kind of a signal.

Matt Cutts: We spend every day, all day, pretty much steeped in looking at high-quality content and low-quality content. I think our engineers and different people who are interested in web spam are relatively attuned to the things that are pretty natural and pretty organic. I think it's funny how you'll see a few people talking about how to fake being natural or fake being organic.

It's really not that hard to really be natural and to really be organic, and sometimes the amount of creativity you put into trying to look natural could be much better used just by developing a great resource, or a great guide, or a great hook that gets people interested. That will attract completely natural links by itself, and those are always going to be much higher quality links because they are really editorially chosen. Someone is really linking to you because they think you've got a great site or really good content.

Eric Enge: I think you have a little bit of the Las Vegas gambling syndrome too. When someone discovers that they have something that appears to have worked, they want to do more, and then they want to do more, and then they want to do more, and it's kind of hard to stop. Certainly, you don't know where the line is, and there is only one way to find the line, which is to go over it.

Matt Cutts: Hopefully the guidelines that we give on the Webmaster guidelines are relatively common sense. I thought it was kind of funny that we responded to community feedback and recommended that people avoid excessive reciprocal links, of course, some of these do happen naturally, but, people started to worry and wonder about what the definition of excessive was. I thought it was kind of funny because within one response or two responses people were like saying "if you are using a ton of automated scripts to send out spam emails that strikes me as excessive."

People pretty reasonably and pretty quickly came to a fairly good definition of what is excessive and that's the sort of thing where we try to give general guidance so that people can use their own common sense. Sometimes people help other people to know where roughly those lines are, so they don't have to worry about getting too close to them.

Eric Enge: The last question is a link related question. You can get a thunderstorm of links in different ways. You get on the front page of Digg, or you can be written up in the New York Times, and suddenly a whole bunch of links pours into your site. There are patterns help by Google that talk about temporal analysis, for example, if you are acquiring links at a certain rate, and suddenly it changes to a very high rate.

That could be a spam signal, right. Correspondently, if you are growing at a high rate, and then that rate drops off significantly, that could be a poor quality signal. So, if you are a site owner and one of these things happens to you, do you need to be concerned about how that will be interpreted?

Matt Cutts: I would tell the average site owner not to worry because, in the same



way that you spend all day thinking about links, and pages, and what's natural and what's not, it's very common for a few things to get on the front page of Digg. It happens dozens of times a day, and so getting there might be a really unique thing for your website, but it happens around the web all the time. And so, a good search engine needs to be able to distinguish the different types of linking patterns, not just by real content, but breaking news and things like that.

I think we do a pretty good job of distinguishing between real links and links that are maybe a little more artificial, and we are going to continue to work on improving that. We'll just keep working to get even smarter about how we process links we see going forward in time.

Eric Enge: You might have a very aggressive guy that is going out there and he knows how to work to Digg system, and he is getting on the front page of Digg every week or so. They would end up with a very large link growth over a short period of time, and that's what some of the pundits would advise you to do.

Matt Cutts: That's an interesting case. I think at least with that situation you've still got to have something that's compelling enough that it gets people interested in it somehow.

Eric Enge: You have appealed to some audience.

Matt Cutts: Yeah. Whether it's a Digg technology crowd or a Techmeme crowd, or Reddit crowd, I think different things appeal to different demographics. It was interesting that at Search Engine Strategies in San Jose you saw Greg Boser pick on the viral link builder approach. But, I think one distinguishing factor is that with a viral link campaign, you still have to go viral. You can't guarantee that something will go viral, so the links you get with these campaigns have some editorial component to them, which is what we are looking for.

Eric Enge: People have to respond at some level or you are not going to get anywhere.

Matt Cutts: Right. I think it's interesting that it's relatively rare these days for people to do a query and find completely off-topic spam. It absolutely can still happen, and if you go looking for it you can find off-topic spam, but it's not a problem that most people have on a day-to-day basis these days. Over time, in Web spam, we start to think more about bias and skew and how to rank things that are on topic appropriately. Just like we used to think about how we return the most relevant pages instead of off-topic stuff.

The fun thing about working at Google is, the challenges are always changing and you come into work and there are always new and interesting situations to play with. So, I think we will keep working on trying to improve search quality and improve how we handle links, and handle reputation, and in turn, we are trying to work with Webmasters who want to return the best content and try to make great sites so they can be successful

Eric Enge: Thank you very much.

Matt Cutts: Always a pleasure to talk to you Eric.

Share Post



Tags

[Google](#) / [Matt Cutts](#)

Leave a Reply

Enter your comment here...

This site uses Akismet to reduce spam. [Learn how your comment data is processed.](#)

