CHAPTER 3

What Could We Watch?

Regardless of the type of site you're running, there are many things you can track: the actions visitors took, the experiences they had, how well they were able to use the site, what they hoped to accomplish, and most importantly, *whether your business benefited in some way from their visits*.

Here's a quick overview of some of the things you'd like to know, and the tools you'll use to collect that knowledge.

What we'd like to know	Tool set to use
How much did visitors benefit my business?	Internal analytics
Where is my traffic coming from?	External analytics
What's working best (and worst)?	Usability testing, A/B testing
How good is my relationship with my market?	Customer surveys, community monitoring
How healthy is my infrastructure?	Performance monitoring
How am I doing against my competitors?	Search, external testing
Where are my risks?	Search, alerting
What are people saying about me?	Search, community monitoring
How are my site and content being used elsewhere?	Search, external analytics

We're now going to look at many of the individual metrics you can track on your website. If you're unfamiliar with how various web monitoring technologies work, you may want to skip to Chapter 4 and treat this chapter as a reference you can return to as you're defining your web monitoring strategy.

How Much Did Visitors Benefit My Business?

When you first conceived your website, you had a goal in mind for your visitors. Whether that was a purchase, a click on some advertising you showed them, a contribution they made, a successful search result, or a satisfied subscriber, the only thing

that really counts now is how well your site helps them accomplish the things you hoped they'd do.

This may sound obvious, but it's overlooked surprisingly often. Beginner web operators focus on traffic *to* the site rather than business outcomes *of* the site.

Conversion and Abandonment

All sites have some kind of goal, and only a percentage of visitors accomplish that goal. The percentage of visitors that your site converts to contributors, buyers, or users is the most important metric you can track. Analyzing traffic by anything other than these goals and outcomes is misleading and dangerous. Visits mean nothing unless your visitors accomplish the things you want them to.

This is so important, we'll say it again: *analyzing web activity by anything other than outcomes leads to terrible mistakes*.

Your site's ability to make visitors do what you wanted is known as conversion. It's usually displayed as a funnel, with visitors arriving at the top and proceeding through the stages of a transaction to the bottom, as shown in Figure 3-1.

By adjusting your site so that more visitors achieve desired goals—and fewer of them leave along the way—you improve conversion. Only once you know that your site can convert visitors should you invest any effort in driving traffic to it.

What to watch: Conversion rates; pages that visitors abandon most.

Click-Throughs

Some sites *expect* people to leave, provided that they're going to a paying advertiser's site. That's how bills get paid. If your site relies on third-party ad injection (such as Google's AdWords or an ad broker) then click-through data is the metric that directly relates to revenue.

A media site's revenue stream is a function of click-through rates and the money advertisers pay for those clicks, measured in cost per mil (CPM), the cost for a thousand visitors. Even if the site is showing sponsored advertising (rather than pay-per-click advertising) for a fixed amount each month, it's important to track click-throughs to prove to sponsors that their money is well spent.

What to watch: The ratio of ads served to ads clicked (click-through ratio); clicks by visitors (to compare to ad network numbers and claims); demographic data and correlation to click-through ratio; and CPM rates from advertisers.

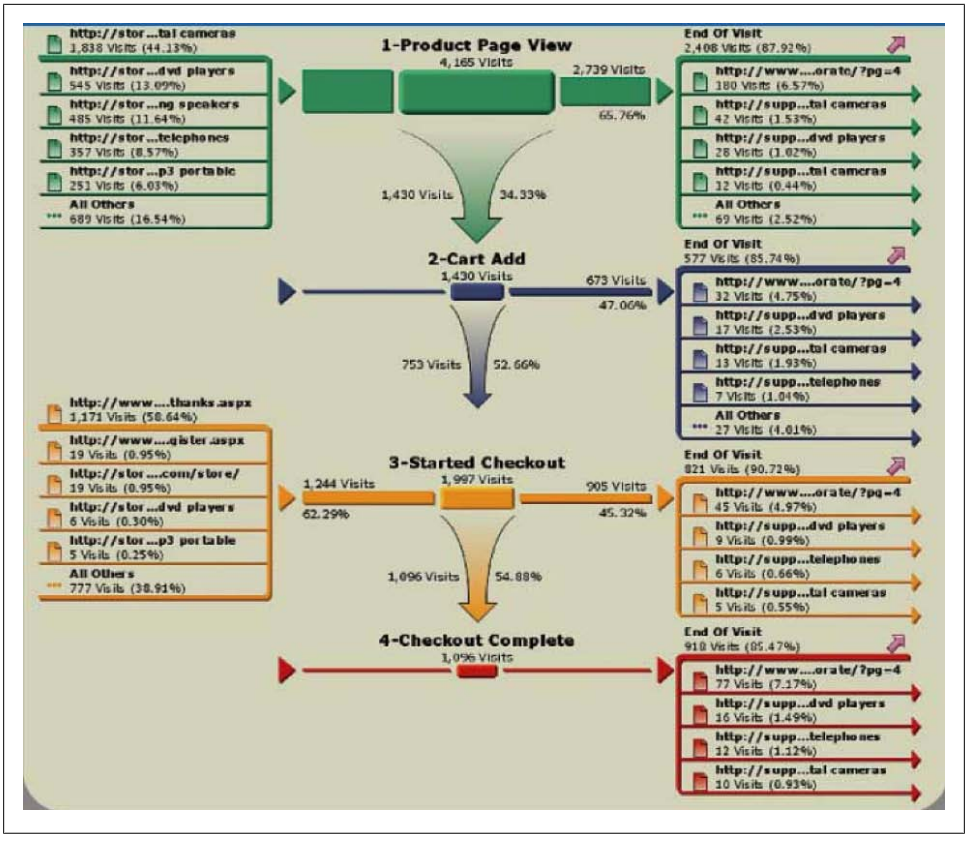

Figure 3-1. A typical e-commerce conversion funnel

Offline Activity

Many actions that start on the Web end elsewhere. These conversions are hard to associate with their ultimate conclusions: the analytics tool can't see the purchase that started online if it ends in a call center, as shown in Figure 3-2.

You can, however, still track conversions that end offline to some degree. Provide a dedicated phone number for calls that begin on the website, then measure call center order volumes alongside requests for contact information. You'll see how much traffic the site is driving to a call center in aggregate, as shown in Figure 3-3.

With a bit more work, you can get a much better idea of offline outcomes. To do this, you'll need an enterprise-class analytics solution that has centralized data warehousing capabilities. First, provide a unique code to visitors that they can then provide to call center operators in return for a discount, as shown in Figure 3-4. Then use this information to associate calls with web visits.

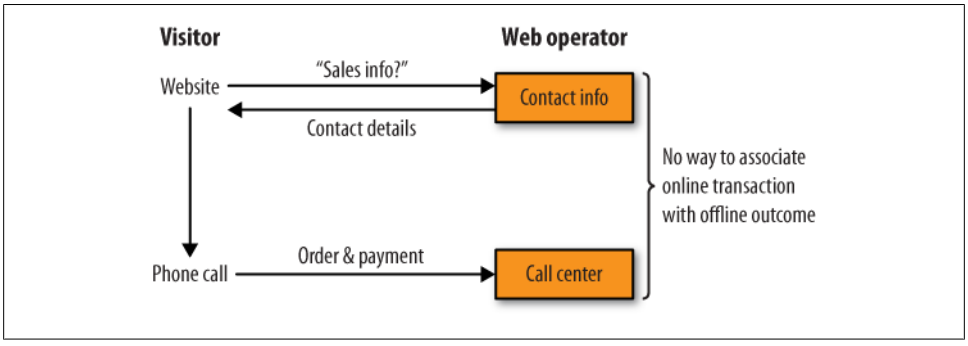

Figure 3-2. A standard web visit with an offline component

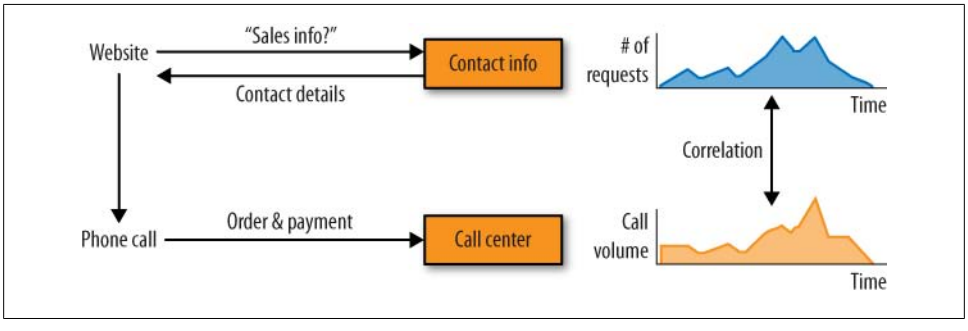

Figure 3-3. Visual correlation of online and offline data sources by time

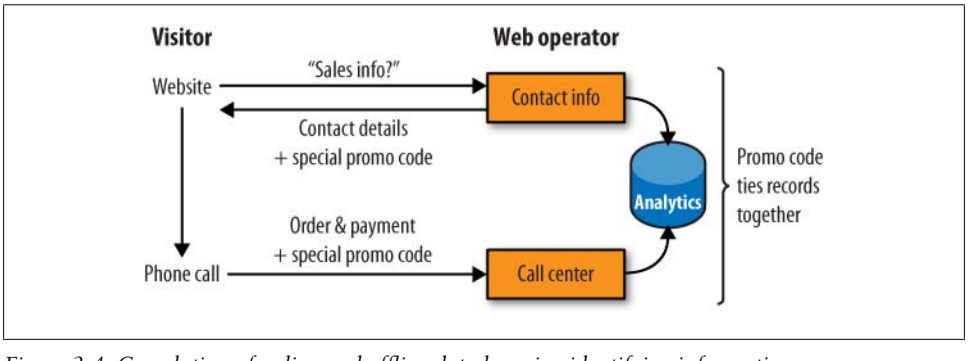

Figure 3-4. Correlation of online and offline data by using identifying information

The need to tie visits to outcomes is one reason many companies are deploying web applets that invite visitors to click to chat with sales or support personnel—it's far easier to track the effectiveness of a website when all the outcomes happen within view of the analytics system.

Figure 3-5. A postcard for FarmsReach.com to be distributed offline includes a custom URL used to tie the marketing message to an online outcome

Real-world beginnings can also lead to online ends, presenting many of the same problems: a tag on a Webkinz toy leads a child to an online portal; an Akoha card prompts someone to create an online account and propagate the card. The postcard pictured in Figure 3-5 contains a unique URL that's tied to an online marketing campaign, allowing an analytics team to compare the effectiveness of this message to others and to optimize offline campaign components.

Standard analytics tools ignore offline components at the beginning or end of a transaction without additional work. If your business has an offline component, you're going to need to get your hands on a system that allows you to integrate your data. Until you can get a system that automates this process, expect to spend a lot of time manually consolidating information.

What to watch: Call center statistics by time; call center data combined with analytics records; on-demand chat usage; lead generation sent to CRM (Customer Relationship Management) applications.

User-Generated Content

If your site thrives on user-generated content (UGC), contribution is key. You need to know how many people are adding to the site, either as editors or commenters, and whether your contributors are creating the content that your audience wants.

UGC contribution can be thought of as the ratio of media consumed to media created. It takes three major forms: new content, editing, and responses. All three are vital to a dynamic site, but too much or too little of any of them can be a bad sign.

- Too many comments are telltale signs of vitriol and infighting that distracts from the main post and discourages more casual visitors.
- Too much new content from a few users suggests spamming or scripting.
- Frequent editing of a single item may signal partisan disagreements.

If you're running a collaborative site, you care about how valuable new content is. Many of the metrics you'll track depend on the platform you're using. For example, if you're running a wiki, you care about incipient links (the links within a wiki entry that point to another, as-yet-unwritten, entry).

If an entry has no incipient links, it's orphaned and not well integrated into the rest of the site. On the other hand, if it has many incipient links that haven't yet been completed, the site isn't generating new content quickly enough. Finally, when many people click on a particular incipient link only to find that the destination page doesn't yet exist that's probably the page you should write next.

We'll return to specific metrics for various kinds of community sites in the community monitoring section of the book.

What to watch: Read-to-post ratio; difference between "average" and "super" users; patterns of down-voting and "burying" new posts; high concentrations of a few frequent contributors; content with high comment or edit rates; incipient link concentration; "sentiment" in the form of brand-specific comments or blog responses.

Subscriptions

Some media sites offer premium subscriptions that give paying customers more storage, downloadable content, better bandwidth, and so on. This can be the main revenue source for analyst firms, writers, and large media content sites such as independent video producers.

Additional bandwidth costs money, so subscriptions need to be monitored for cost to ensure that the premium service contributes to the business as a whole. For example, if users get high-bandwidth game downloads, how much traffic are they consuming? If they buy content through PayPal, what's the charge for payment? In these situations, analytics becomes a form of accounting, and you may need to collect information from networking equipment such as gigabytes of traffic delivered.

What to watch: Subscription enrollment monitored as a transaction goal; subscription resource usage such as bandwidth or storage costs.

Billing and Account Use

If you're running a subscription website—such as a SaaS application—then your subscribers pay a recurring fee to use the application. This is commonly billed per month, and may be paid for by the individual user or as a part of a wider subscription from an employer.

It's essential to track billing and account use, not only because it shows your revenues, but also because it can pinpoint users who are unlikely to renew. You'll need to define what constitutes an "active" user, and watch how many of your users are no longer active. You'll also want to watch the rate of nonrenewal to measure churn.

Small business invoicing outfit Freshbooks lets users invoice only a few customers for free; automated video production service Animoto limits the length of video clips it generates for free; and picture site Flickr constrains the volume of uploads in a month. This "velvet rope" pricing strategy encourages users to upgrade to the paid service. If you take this approach, treat the act of converting free users to paying users as a conversion process in a transactional site.

What to watch: Monthly revenue, number of paying subscribers; "active" versus "idle" users; volume of incidents per subscriber company; churn (nonrenewal) versus new enrollments.

Where Is My Traffic Coming From?

Once a website benefits from visitors in some way, it's time to worry about where those visitors are coming from. Doing so lets you:

- Encourage sites that send you traffic, either by contacting them, sponsoring them, or inviting them to become an affiliate of some sort.
- Advertise on sites that send you visitors who convert, since visitors they send your way are more likely to do what you want once they reach your site.
- Measure affiliate referrals as part of an affiliate compensation program.
- Understand the organic search terms people use to find you and adjust your marketing, positioning, and search engine optimization accordingly.
- Verify that paid search results have a good return on investment.
- Find the places where your customers, your competitors, and the Internet as a whole are talking about you so you can join the conversation.

The science of getting the right visitors to your site is a combination of Affiliate Marketing, Search Engine Marketing, and Search Engine Optimization, which are beyond the scope of this book.

Referring Websites

When a web browser visits a website, it sends a request for a page. If the user linked to that page from elsewhere, the browser includes a referring URL. This lets you know who's sending you traffic.

The HTTP standard actually calls a referring URL a "referer," which may have been a typo on the part of the standard's authors. We'll use the more common spelling of "referrer" here.

If you know the page that referred visitors, you can track those visits back to the site that sent them and see what's driving them to you. Remember, however, that you need to look not only at who's sending you visitors, but also at who's sending you the ones that *convert*.

Referring URLs, once a mainstay of analytics, are becoming less common in web requests. JavaScript within web pages or from Flash plug-ins may not include it, and desktop clients may not preserve the referring URL. In other words, you don't always know where visitors came from.

What to watch: Traffic volume by referring URL; goal conversion by URL.

Inbound Links from Social Networks

An increasing number of visitors come to you from social networks. If your media site breaks a news story or offers popular content, social communities will often link to it. This includes not only social news aggregators like reddit or Digg, but also bloggers, comment threads, and sites like Twitter.

These traffic sources present their own challenges for monitoring.

- The links that sent you visitors may appear in comment threads or transient conversations that you can't link back to and examine because they've expired.
- The traffic may come from desktop clients (Twitter users, for example, employ Tweetdeck and Twhirl to read messages without a web browser). These clients omit referring URLs, making visit sources hard to track.
- The social network may require you to be a member, or at the very least, to log in to see the referring content.
- The author of the link may not be affiliated with the operator of the social network, so you may have no recourse if you're misrepresented.

Most of the challenges of social network traffic come from difficulties in tracking down the original source of a message. Sometimes, you simply won't see a referrer. But other times, you'll still see the referring URL but will need to interpret it differently.

For example, a referral from *www.reddit.com/new/* means that the link came from the list of new stories submitted to reddit. Over time, that link will move off the New Stories section of reddit, so you won't be able to find the source submission there. But you do know that the referral was the result of a reddit submission. Other social network referrals contain similar clues:

- *Microblogging websites*, such as Twitter, FriendFeed, or Identi.ca may tell you which person mentioned you, but also whether the inbound link came from someone's own Twitter page (*/home*) or a page they were reading that belongs to someone else.
- *URL-shortening services*, such as tinyurl, is.gd, bit.ly, or snipurl may hide some of the referral traffic. These usually rely on an HTTP redirect and won't show up in the analytics data, but some providers such as bit.ly offer their own analytics.
- Referrals from *microblogging search* show that a visitor learned about your site when searching Twitter feeds. Links from other microblog aggregation (such as hash tag sites) are signs that you're a topic online.
- Referrals from *mail clients*, such as Yahoo! Mail, Hotmail, or Gmail mean someone forwarded your URL via email.
- Referrals from *web-based chat* clients, like meebo.com, are a sign that people are discussing your site in instant messages.
- *Homepage portal* referral URLs can be confusing, and you need to look within the URL to understand the traffic source. For example, a URL ending in */ig/* came from an iGoogle home page; */reader/view* came from Google Reader; and */notebook/* came from Google Notebook. Some analytics packages break down these referrals automatically.

In other words, all referring sites aren't equal. It's not enough to differentiate and analyze referring sites by name; you have to determine the nature of the referrer. Different types of referring sites require different forms of analysis.

Since social networks and communities contain UGC, referrals may also show you when people aren't just linking to you, but are instead presenting your content as their own without proper attribution. Tracking social network referrals is an important part of protecting your intellectual property.

People will often mention your content elsewhere but not link to you directly. This will generally result in users searching for the name of your site and the content in question, which will make referral traffic look less significant while overinflating the amount of search and direct traffic, particularly navigational search.

Community managers need to identify the source of the traffic so they can engage the people who brought them the attention and mitigate the inevitable comment battles. And site designers need to make sure that new visitors become returning visitors.

What to watch: Referring sites and tools by group; sudden changes in unique visitors and enrollments from those groups; search results from social aggregators and microblogs.

Visitor Motivation

Knowing how visitors got to you doesn't always tell the whole story. Sometimes the only way to get inside a visitor's head is to ask her, using surveys and questions on the site. Such approaches are generally lumped into the broader category of *voice of the customer* (VOC).

Asking visitors what they think can yield surprising results. In the early days of travel sites, for example, site owners noticed an extremely high rate of abandonment just before paying for hotels. The sites tried many things to improve conversion—new page layouts, special offers, and so on, but it was only when they asked visitors, through pop-up surveys, why they were leaving that they realized the problem: many users were just checking room availability, with no intention of buying.

Use any travel site today and you'll likely see messages encouraging visitors to sign up to be notified of specials or changes in availability. This is a direct result of the insights gained through VOC approaches.

What to watch: What were users trying to accomplish? Did they plan to make a purchase? Where did they first hear about the site? What other products or services are they considering? What demographic do they fit into?

What's Working Best (and Worst)?

You can always do better. Even with large volumes of inbound traffic and a site that guides visitors to the outcomes you want, there's work to be done: filling shopping carts fuller, emphasizing the best campaigns, ensuring users find things quickly and easily, and so on. One of the main uses of web analytics is optimization.

Site Effectiveness

Your site converts visitors, and you have visitors coming in. What could be better than that? For starters, they could buy more each time they check out. A site that convinces visitors to purchase more than what they initially intended is an effective site.

Many e-commerce sites suggest related purchases or offer package deals. A bookstore might try to bundle a book the visitor is buying with another by offering savings, or try to show what else buyers of that book also bought. A hosting company could try to sell a multiyear contract for a discount. And an airline might refer ticket buyers to a partner rental company or try to add in travel insurance.

The total shopping cart value and the acceptance of these upselling attempts are essential metrics for e-commerce sites. You can treat upselling as a second funnel, and you should track upsold goods independently from the initial purchase whenever possible. Because upselling adds to an existing transaction, you should experiment with it.

Effectiveness isn't just for transactional sites, however. For example, on a collaborative site, how many visitors subscribe to a mailing list or an RSS feed? On a static web portal, how many people visit the "About" page or check for contact information?

What to watch: Percentage of upselling attempts that work; total cart value.

Ad and Campaign Effectiveness

Department store magnate John Wanamaker is supposed to have said, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." Yesterday's chain-smoking ad executive, promoter of subjective opinions, espoused the value of hard-to-measure "brand awareness" at three-martini lunches.

Not so online. Every penny spent on advertising can be linked back to how much it benefits the business. Today, the ideal marketer is an analytical tyrant constantly searching for the perfect campaign, more at home with a spreadsheet than a cocktail. The main reason for this shift is the hard, unavoidable truth of campaign analytics.

Referring sites aren't the only method of categorizing inbound traffic. People don't surf the web by randomly entering URLs to see if they exist. With the exception of organic traffic, most visitors arrive because of a campaign. This may be an online campaign—banner ads, sponsorship, or paid content—but it may also be a part of an offline campaign such as a movie trailer or radio spot, or simply good word of mouth and an informal community.

Analytics applications can segment incoming traffic by campaign to measure how much they helped the bottom line. While it's harder to measure the effectiveness of offline advertising, you can still get good results with unique URLs that get press coverage (such as *http://www.watchingwebsites.com/booklink*).

What to watch: Which campaigns are working best, segmented by campaign ID, media ID, or press release.

Findability and Search Effectiveness

Users conduct site searches to find what they're after. Rather than browsing through several hierarchies of a directory, users prefer to type in what they're looking for and choose from the results.

Yet many site owners often overlook internal search metrics in their analyses. You need to know if your users are finding the results they're after quickly so you can better label and index your site.

There are many commercial search engines, as well as robust open source engines like Lucene (*http://lucene.apache.org/java/docs/index.html*), that developers can integrate into a site. Hosted search engines like Google (*www.google.com/sitesearch/*) can also be embedded within a site and configured to only return results from the site itself. Internal and third-party search engines can generate reports (for example, *http://blog.foofactory.fi/2008/08/interactive-query-reporting-with-lucene.html*) on what visitors are searching for; if this search data is tied into analytics, we can measure search effectiveness.

What to watch: How many searches ended with another search? With a return to the home page? With abandonment? What are the most popular search terms whose sessions have a significantly higher abandonment rate? Which search terms lead to a second search term?

Trouble Ticketing and Escalation

An increase in call center activity and support email messages are sure signs of a broken site. Site operators need to track the volume of trouble tickets related to the website, and ideally relate those trouble tickets to the user visits that cause them in order to speed up problem diagnosis.

There are a number of products that can record visits, flagging those that had problems and indexing them by the identities of the visitors or the errors that occurred. We'll look at these tools in depth in Chapter 6, when we consider how visitors interacted with the site, but for now, recognize that capturing a record of what actually happened makes it far easier to fix errors and prove who or what is at fault. You can also use records of visits as scripts for testing later on.

What to watch: Number of errors seen in logs or sent to users by the server; number of calls into the call center; errored visits with navigation path and replay.

Content Popularity

Media sites are about content. The successful ones put popular content on the home page, alongside ad space for which they charge a premium. Knowing what works best is essential, but it's also complex. Popular stories may be one-hit wonders—trashy content that draws visitors who won't stay. Who you attract with your content, and what they do afterward, is an important part of what works best. In other words, content popularity has to tie back to site goals, as well as stickiness metrics, rather than just page views.

But what about the fleeting popularity of transient content such as breaking news? This is a more difficult problem—stories grow stale over time. To balance fresh content with community popularity, social news aggregators like reddit, Slashdot, and Digg count the number of *upvotes* (how many people liked the content) and divide them by the content's age (based on the notion that content gets "stale"). It's not really this simple—other factors, such as the rate of upvoting, make up each site's proprietary ranking

algorithms. Upvoting of this kind also shows the voting scores to the community, making it more likely to be seen and voted on.

While social networking is a relatively recent phenomenon, its origins can be traced back to K. Eric Drexler's concept of "filtered hyperlinks," which Drexler describes as "a system that enables users to automatically display some links and hide others (based on user-selected criteria)," which "implies support for what may be termed social software, including voting and evaluation schemes that provide criteria for later filtering." Drexler's paper was first published at Hypertext 87 (*http://www.islandone.org/Foresight/WebEnhance/HPEK1.html*).

What to watch: Content popularity by number of visitors; bounce rate; outcomes such as enrollment; ad click-through.

Usability

No site will succeed if it's hard to use. Focus groups and prerelease testing can identify egregious errors before you launch, but there's no substitute for watching a site in production.

There are high-end products to capture and replay every user's visit, but even if you are on a tight budget you can use JavaScript-based tools to monitor click patterns and understand where a user's mouse moved. Some of these are built into free tools like Google Analytics. By combining these with careful testing of different page designs, you can maximize the usability of an application.

Whenever visitors link to help or support pages, track the pages from which they came, which will point you to the source of their problems.

What to watch: Click patterns on key pages, particularly abandonment points and landing pages; referring URLs on support and feedback pages; form abandonment analysis; visitor feedback surveys.

User Productivity

While usability focuses on whether someone could understand how to do something the way it was intended, user productivity looks at whether visitors could accomplish their tasks quickly and without errors. Every website operator should care whether visitors can accomplish goals, but for SaaS sites this is particularly important, as users may spend their entire workday interacting with the application.

With the growth of the Web as a business platform, people are using browsers for tasks such as order entry or account lookups. If someone's employees are using your web application to accomplish tasks, you need to measure the rate at which those tasks are

completed. This could be the number of orders entered in an hour or how long it takes to process an account.

The business that pays for a SaaS subscription cares about its employees' productivity. If you release a version of your SaaS website on which employees take twice as long to enter an order, you're sure to hear about it soon from their frustrated employer. On the other hand, if your website lets employees look up twice as many accounts an hour as an in-house alternative, your sales team should use this as a differentiator when talking to future customers.

What to watch: Time to accomplish a unit of work; tasks completed per session; errors made; changes in productivity across releases.

Community Rankings and Rewards

Sometimes, what's working on a website isn't the site itself, it's key contributors. Much of Wikipedia is edited by a small, loyal staff of volunteers; users flag inappropriate content on community sites like Craigslist, and power users promote interesting content on social news aggregators.

What to watch: Top contributors; contributions by user; specific "rewards" for types of contribution.

How Good Is My Relationship with My Visitors?

Once you've got your site in order, traffic is flowing in, and you're making the most of all of your visitors, it's time to be sure your relationship with them is long and fruitful.

In the early days of the Web, the main measure of engagement with your visitors was loyalty—how often they returned to your site. Today's users receive messages from a wide range of social networks, RSS (Really Simple Syndication) feeds, email subscriptions, and newsgroups, all of which push content out to them without them first asking for it.

As a result, visits don't count as much. The definition of loyalty needs to be amended for this two-way relationship. It's not just about how often your visitors return; it's about how willing they are to let you contact them and how frequently they act on your advances and offers.

Loyalty

The best visitors are those who keep coming back. Thanks to browser cookies, most web analytics applications show the ratio of new to returning visitors. Strike a healthy balance here: get new blood so you can grow, but encourage existing visitors to return so they become regular buyers or contributors.

For this, you need to look at two additional metrics. One is the average time between visits, which shows you how much a part of your audience's daily life you are. The other is the number of users who no longer engage with the site. Since users don't usually terminate an account, this is measured by the time since their last login.

What to watch: Ratio of new to returning visitors; average time between visits; time since last login; rate of attrition or disengagement.

Enrollment

Reaching people when they visit isn't enough. Visitors you're allowed to contact are the holy grail of online marketing. Their love is more than loyalty—it's permission.

Enrollment is valuable because consumers are increasingly skeptical of web marketing. Browsers run ad-blocking software. Mail clients hide pictures. Some portals let users hide advertising. An enrolled visitor is reachable despite all of these obstacles.

Enrollment also provides better targeting. You can ask subscribers for demographic information such as gender, interests, and income, then tailor your messages—and those of your advertisers—to your audience.

What to watch: Signups; actual enrollments (email messages sent that didn't bounce); email churn (addresses that are no longer valid); RSS subscription rates.

Reach

It's great to have people enroll, but it's even better to actually be able to reach them. Whether through email subscriptions, alerts, or RSS feeds, *reach* is the measurement of how many enrolled visitors actually see your messages.

In the case of email, this may be the number of people who opened the message. For RSS feeds, it's the number that actually looked at the update you sent them. For video, it could be the number that played the content beyond a certain point. Figure 3-6 shows the FeedBurner report for subscribers (the number enrolled) and reach (the number that actually saw a message) for a blog.

Reach is a far more meaningful measure of subscription, since it discounts "stale" enrollments and shows how well your outbound messages, blogs, and alerts result in action.

What to watch: Reach of email recipients; reach of RSS subscribers.

How Healthy Is My Infrastructure?

Slow page loads or excessive downtime can undermine even the best-designed, most effective, easiest-to-use website. While web analytics shows you *what* people are doing

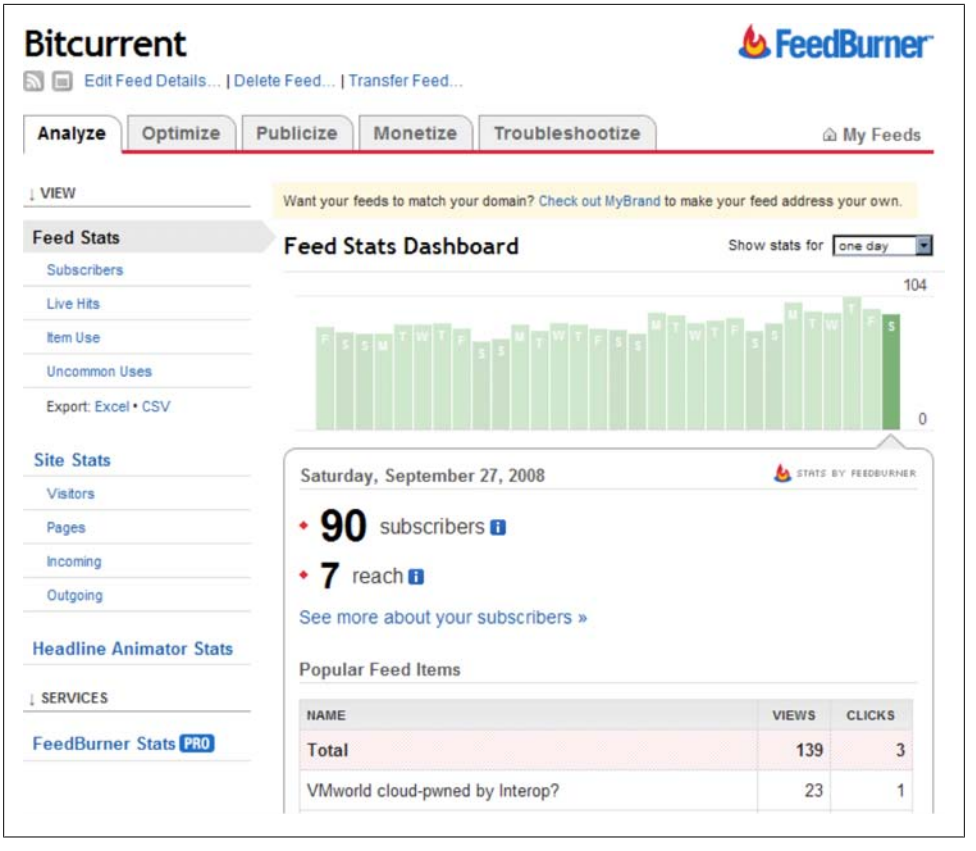

Figure 3-6. Google's FeedBurner Feed Stats

on your site; end user monitoring shows you whether they *could* do it—and how quickly they did it.

Availability and Performance

The most basic metrics for web health are availability (is it working?) and performance (how fast is it?), sometimes referred to collectively as performability. These can be measured on a broad, site-wide basis by running synthetic tests at regular intervals; or they can be measured for every visit to every page with real user monitoring (RUM).

In general, availability (the time a site is usable) is communicated as a percentage of tests that were able to retrieve the page correctly. Performance (how long the user had to wait to interact with the site) is measured in seconds to load a page for a particular segment of visitors.

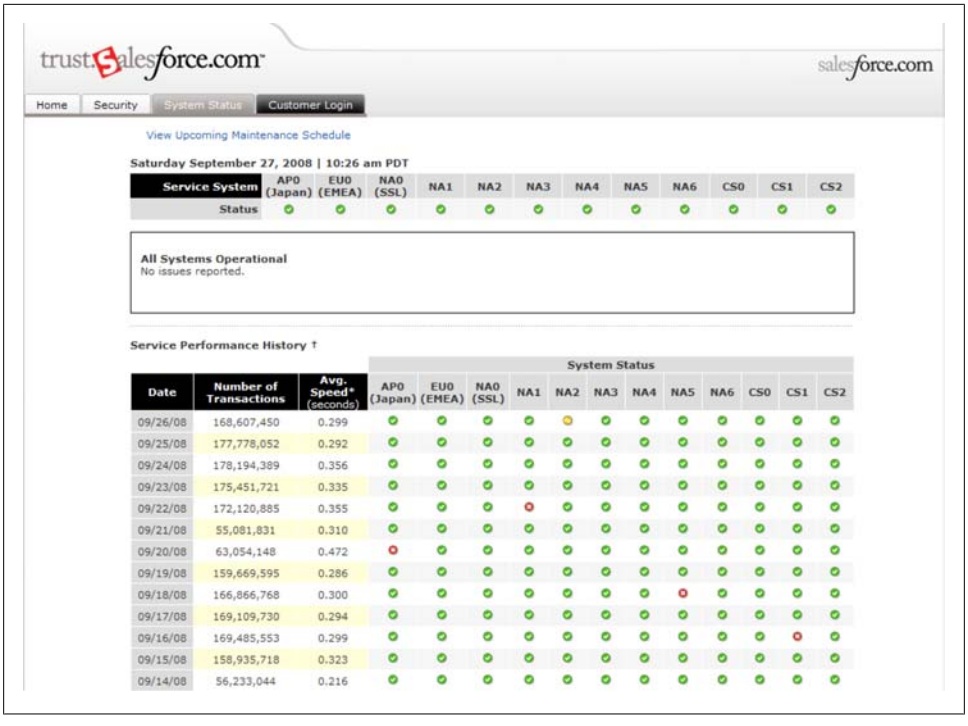

Figure 3-7. Salesforce.com's System Status dashboard

What to watch: Availability from locations where visitors drive revenue; page load time for uncached and cached pages; end-to-end and host time at various traffic volumes; changes in performance and availability over time or across releases.

Service Level Agreement Compliance

If people pay to use your site, you have an implied contract that you'll be available and usable. While this may not be a formal Service Level Agreement (SLA), you should have internal guidelines for how much delay is acceptable. Some SaaS providers, such as Salesforce.com, show current uptime information to subscribers and use this as a marketing tool (see Figure 3-7).

A properly crafted SLA includes not only acceptable performance and availability, but also time windows, stakeholders, and which functions of the website are covered.

Your SLAs may also depend on your partners, so you may have to consider other SLAs when measuring your own. If, for example, your site is part of a web supply chain, you need to measure the other links in that chain to know whether you're at fault when an SLA is missed.

SLAs Are Complex Things

Some web users currently have formal SLAs with their providers, and as more organizations use the Web as the primary channel for business, SLAs will become commonplace. There are many factors to consider when defining an SLA, which is one of the reasons they tend to be either ponderously detailed or uselessly simple.

- From whose perspective are you measuring the SLA?
- Are you measuring the website as a whole or its individual servers?
- Are you watching a single page or an entire workflow or business process?
- Are you measuring from inside your firewall, outside your firewall, or from where your customers are located?
- What clients and operating systems are you using to measure performance?
- Are you watching actual users or simulating their visits?
- Are you measuring the average performance or a percentile (the worst five percent, for example)?
- Does your SLA apply around the clock or only during business hours? Whose business hours?

There's no one correct answer to these questions, but organizations need to know what they're measuring and what they're not. In Figure 3-7, for example, what is Salesforce really measuring? Will they report that a North American instance is not working properly if West Coast users are doing fine but East Coast users are having performance issues?

Measure and report the metrics that comprise an SLA in a regular fashion to both your colleagues and your customers.

What to watch: SLA metrics against agreed-upon targets; customers or subscribers whose SLAs were violated.

Content Delivery

Measuring the delivery of simple, static advertising was once straightforward: if the browser received the page, the user saw the ad. With the advent of interactive advertising and video, however, delivery to the browser no longer means visibility to the user.

Content delivery is important for media companies. A Flash ad may be measured for its delivery to the browser, whether its sites were within the visible area of the browser, and whether the visitor's mouse moved over it. Users may need to interact with the content—by rolling over the ad, clicking a sound button, and so on. Then the ad plays and the user either clicks on the offer or ignores it. This means *each interactive ad has its own abandonment process*.

The provider that served the ad tracks this. The Flash content sends messages back about engagement, abandonment, and conversion. As a result, media site operators don't need to treat this content differently from static advertising.

While rich media often requires custom analytics, there's one kind of embedded media that's quickly becoming mainstream: Web video. David Hogue of Fluid calls it "the new JPEG," a reflection of how commonplace it is on today's sites.

While there are a variety of companies specializing in video analytics (such as Visible Measures, divinity Metrics, Streametrics, TubeMogul, and Traackr), embedded video is quickly becoming a part of more mainstream web analytics packages. It is also becoming a feature of many content delivery network (CDN) offerings that specialize in video.

Most for-pay analytics offerings available today allow a video player to send messages back to the analytics service when key events, such as pausing or rewinding, take place, as shown in Figure 3-8.

Embedded video serves many purposes on websites. Sometimes it's the reason for the site itself—the content for which visitors come in the first place. Sometimes it's a form of advertising, tracked by the ad network that served it. Sometimes it's a lure to draw the visitor deeper into the site. In each case, what you measure will depend on your site.

What to watch: Content engagement; attention; completion of the media; pauses.

Capacity and Flash Traffic: When Digg and Twitter Come to Visit

When a community suddenly discovers content that it likes, the result is a flash crowd. A mention by a popular blogger, breaking news, or upvoting on a social news aggregator can send thousands of visitors to your website in seconds.

For most websites, capacity and bandwidth is finite. When servers get busy and networks get congested, performance drops. The problem with flash crowds is that they last for only a few hours or days—after that, any excess capacity you put in place is wasted. One of the attractions of CDNs and on-demand computing infrastructure is the ability to "burst" to handle sudden traffic without making a long-term investment in bandwidth or hardware.

When you're on the receiving end of a flash crowd, there's a lot to do. Marketing needs to engage the one-time visitors, making them loyal and encouraging them to subscribe or return. IT operators need to ensure that there's enough capacity, working with service providers or provisioning additional resources if applicable. And community managers need to identify the source of the traffic so they can nurture and prolong the attention.

While flash crowds create dramatic bursts of traffic, a gradual, sustained increase in traffic can sneak up on you and consume all available capacity. You need to monitor

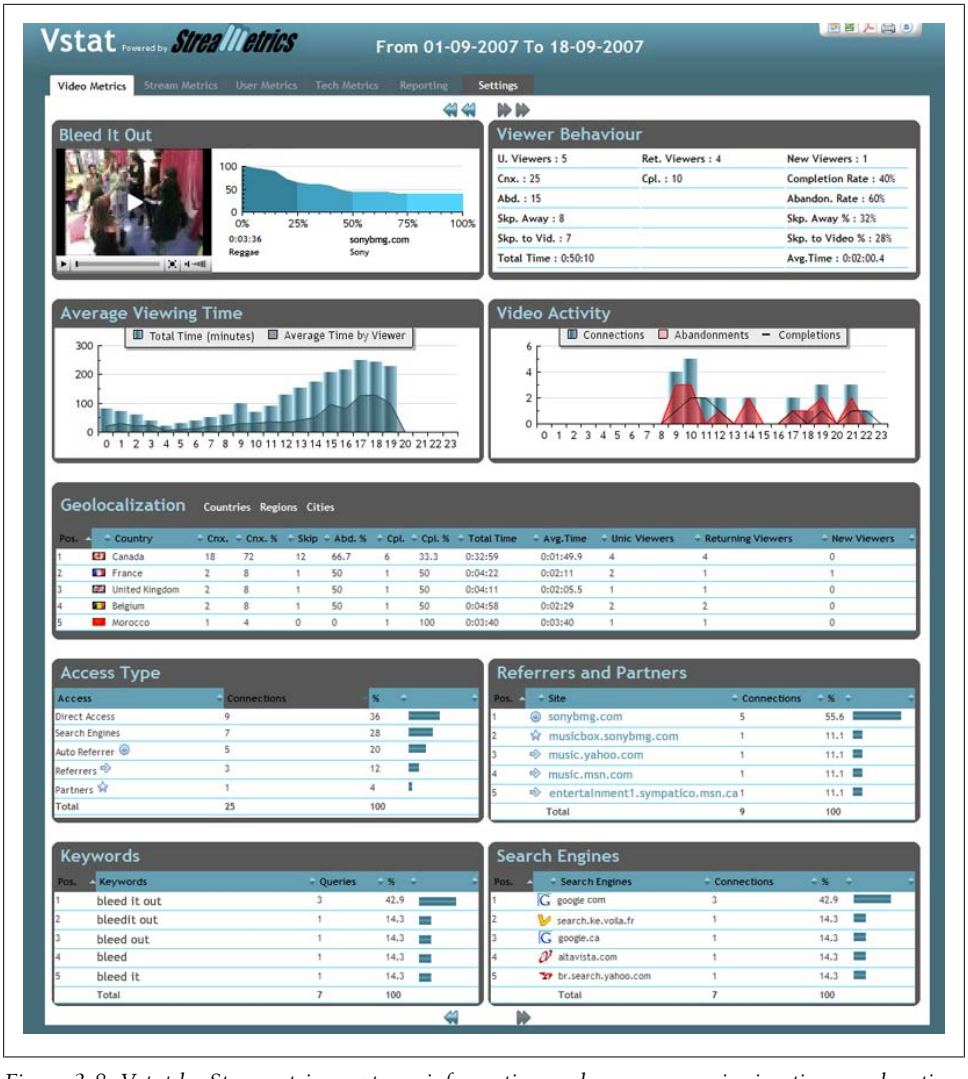

Figure 3-8. Vstat by Streametrics captures information such as average viewing time, geolocation, referrers, and so on

long-term increases in page latency or server processing or decreases in availability that may be linked to increased demand for your website.

Analytics is a good place to start: IT operators should correlate periods of poor performance with periods of high traffic to bandwidth- or processor-intensive parts of the site. If there's a gradual increase in the volume of downloads, you should plan for additional bandwidth. Similarly, if there's a rise in transaction processing, you may need more servers.

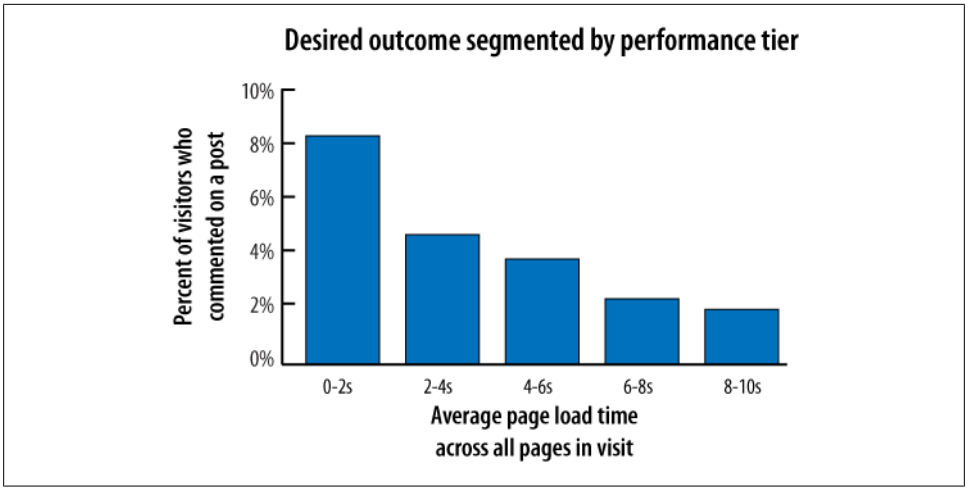

Figure 3-9. Segmenting conversion rates by tiers of web performance

Too often, IT operations and web analytics teams don't talk. The result is last-minute additions to capacity rather than planned, predictable spending.

What to watch: Sudden increases in requests for content; referring URLs; search engine results that reference the subject or the company; infrastructure health metrics; growth in large-sized content or requests for processor-intensive transactions

Impact of Performance on Outcomes

While you can measure the impact of visitors on performance, it's equally important to measure the impact of performance on visitors. Poor performance has a direct impact on outcomes like conversion rate, as well as on user productivity. Google and Amazon both report a strong correlation between increased delay and higher bounce rates or abandonment, and responsive websites encourage users to enter a "flow state" of increased productivity, while slow sites encourage distraction.

The relationship between performance and conversion can be measured on an individual basis, by making performance a metric that's tracked by analytics and by segmenting conversion rates for visitors who had different levels of page latency, as shown in Figure 3-9.

However, this can be hard to do unless you have a way of combining web analytics with the experience of individual end users.

Another way to understand the impact of performance is to compare aggregate page latency with aggregate conversion metrics, as shown in Figure 3-10. To do this properly, you need to eliminate any other factors that may be affecting conversion, such as promotions, daily spikes, or seasonal sales increases.

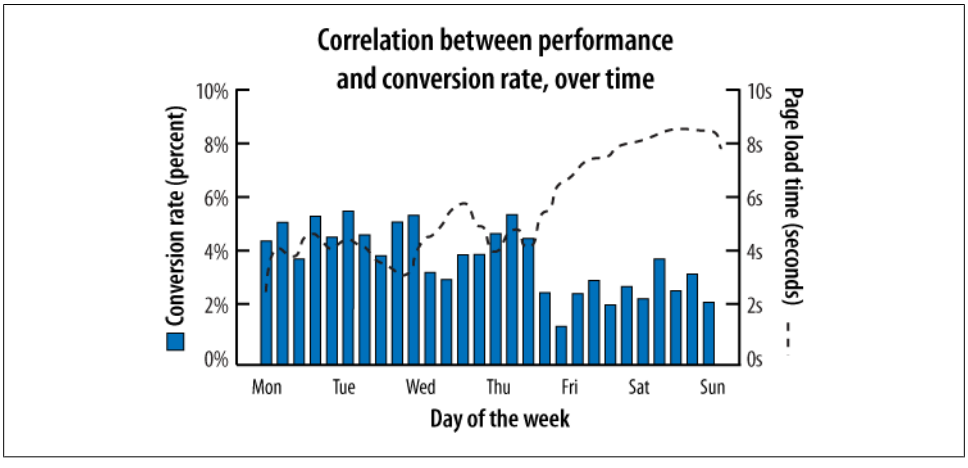

Figure 3-10. Aggregate view of conversion rate alongside site performance

The fundamental question we want to answer is: *does a slow user experience result in a lower conversion rate or a reduced amount of upselling?*

What to watch: Conversion rates segmented by tiers of page latency; daily performance and availability summaries compared with revenue and conversion; revenue loss due to downtime.

Traffic Spikes from Marketing Efforts

Marketing campaigns should drive site traffic. You need to identify the additional volume of visitors to your site not only for marketing reasons, but also to understand the impact that marketing promotions have on your infrastructure and capacity.

Properly constructed campaigns have some unique identifier—a custom URL, a unique referrer ID, or some other part of the initial request to the site—that lets you tie it back to a campaign. This is used to measure ad and campaign effectiveness. You can use the same data to measure traffic volumes in technical terms—number of HTTP sessions, number of requests per second, megabits per second of data delivered, availability, and so on.

Pay particular attention to first-time visitors. They place a greater load on the network (because their browsers have yet to cache large objects) and on applications because of enrollment, email registration, and other functions that occur when a visitor first arrives.

What to watch: Traffic by marketing campaign alongside infrastructure health metrics, such as availability or performance, on as granular a level as possible (ideally per-visit). A summary similar to the one shown in Figure 3-11 is ideal.

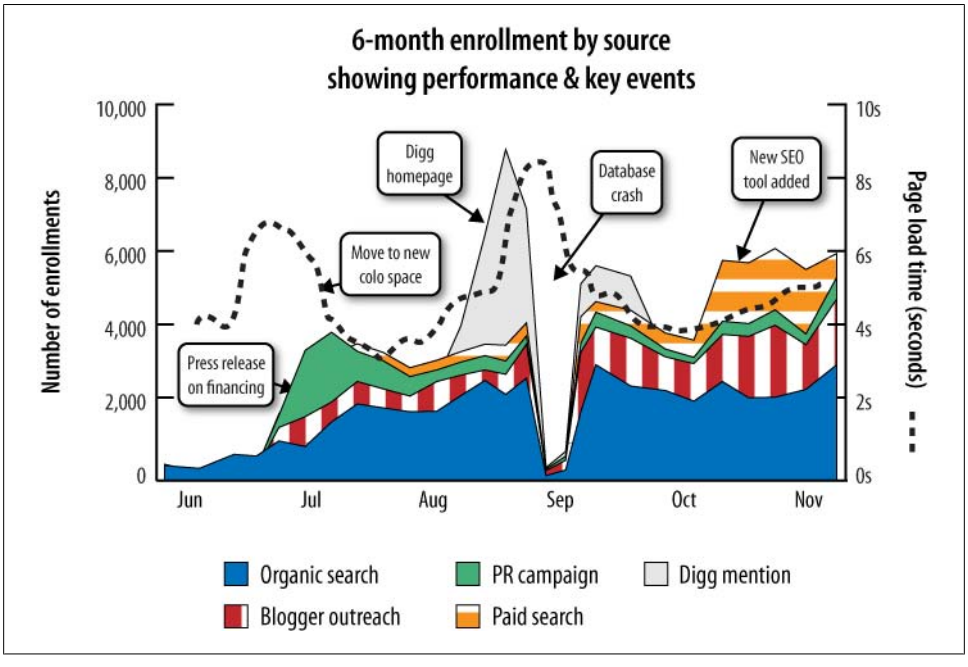

Figure 3-11. A "state of social media" report by month alongside performance information

Seasonal Usage Patterns

If your business is highly seasonal, you need to understand historical usage patterns. The fates of firms like Hallmark and ProFlowers are tied to specific holidays—indeed, at ProFlowers, so much of the company's transactions happen on Valentine's Day that they refer to it internally as "V-day."

Seasonal usage isn't really a new metric, but it's a requirement for monitoring in general. If you're allowed to, collect at least five quarters of data so you can compare each month to the same month of the previous year. You're doing this for two reasons: to understand usage trends so you can plan for capacity changes, and to confirm that you're meeting long-term SLAs.

You only need to store aggregate data, such as hourly performance and availability, for this long. Indeed, many governments and privacy organizations are looking more closely at the long-term storage of personal information, and some sites have a deletion policy that may limit your ability to capture long-term trends.

What to watch: Page views; performance, availability, and volume of CPU-intensive tasks (like search or checkout) on a daily basis for at least 15 months.

How Am I Doing Against the Competition?

You want visitors. Unfortunately, so do your competitors. If visitors aren't on your website, you want to know where they're going, why they're going there, and how you stack up against the other players in your industry.

In addition to monitoring your own website and the communities that affect your business, you also need to watch your competition.

Site Popularity and Ranking

On the Web, popularity matters. When it comes to valuations, most startups and media outlets are judged by their monthly unique visitor count, which is considered a measure of a site's ability to reach people. Relevance-based search engine rankings reinforce this, because sites with more inbound links are generally considered more authoritative.

Some websites pay marketing firms to funnel traffic to them. Artificial inflation of visits to the site does nothing to improve conversion rates. These paid visitors seldom turn into buyers or contributors, but they do raise the site's profile with ranking services such as comScore, which may eventually get the site noticed by others. That said, raw traffic volumes are a spurious metric for comparing yourself to others.

Several services, such as Alexa and Compete.com, try to estimate site popularity. Use this data with caution. Alexa, for example, collects data on site activity from browser toolbars, then extrapolates it to the population as a whole. Unfortunately, this approach has many limitations, including problems with SSL visibility and concerns that the sample population doesn't match the Web as a whole (see *http://www.techcrunch.com/2007/11/25/alexas-make-believe-internet/*, which points out that according to Alexa, YouTube overtook all of Google at one point).

Compete.com has different methods for determining rankings, and even its estimates don't map cleanly to actual traffic. Figure 3-12 shows a comparison of actual traffic volumes and third-party traffic estimates.

Any mention of accuracy begs the question, "What is actual traffic?" Due to the differences in measurement methodologies across various ranking sites, there's bound to be a difference in traffic estimates. Rough trends should, however, be representative of what's going on, and comparing several companies using the same tools and definitions is a valid comparison.

Traffic estimates work well for broad, competitive analysis across large sites, but fail with low-traffic sites. Alexa, for example, doesn't track estimated reach beyond the top 100,000 websites. This data is still valuable for determining basic growth in an industry—if the large sites with which you compete are growing at 15% a month, a 15% traffic increase on your own site means you're merely holding your own.

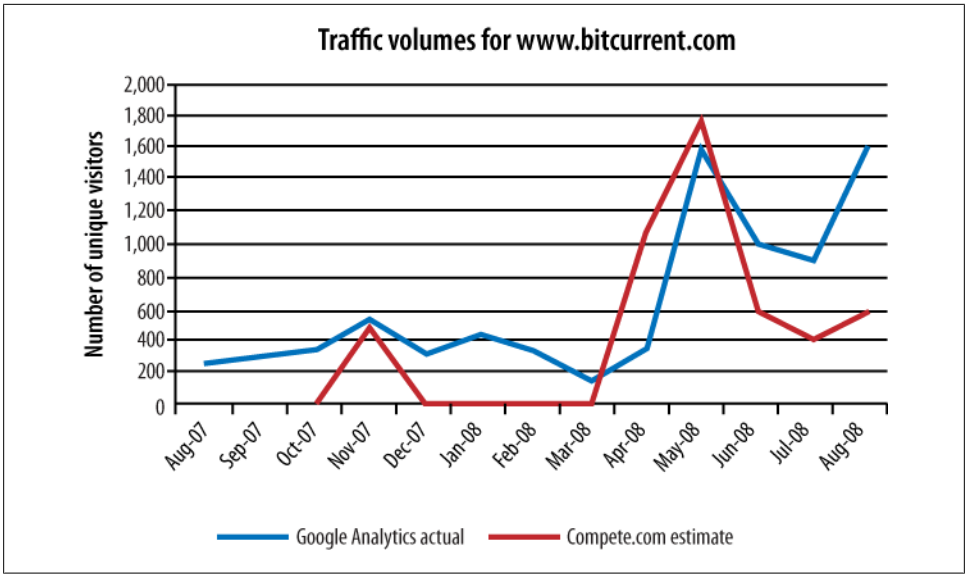

Figure 3-12. Unique visitors compared to Compete.com estimates

There are other ways to measure popularity. You can count how many people type a URL into a search engine—what's known as *navigational search*. This happens a lot. From July to September 2007, Compete.com reported that roughly 17 percent of all searches were navigational searches (*http://blog.compete.com/2007/10/17/navigational-search-google-yaho0-msn/*). We can analyze search terms with tools like Google Trends or Google Insights (*www.google.com/trends* or *http://www.google.com/insights/search/*) and get some idea of relative site popularity. Insights also shows searches by geography, so it can be a useful tool for identifying new markets.

The opposite of navigational search is type-in traffic, where users type a search term like "pizza" into the address bar. This behavior is one of the reasons firms like Marchex buy domains like pizza.com. See John Battelle's article on the subject at *http://battellemedia.com/archives/002118.php*.

Technorati and BlogPulse also show the popularity of sites and topics, as shown in Figure 3-13, although their focus is on blogging.

What to watch: Page views; unique daily visitors; new visitors; Google PageRank; Google Trends and Google Insights; incoming links; reach on panel sites like Compete.com or toolbar ranking sites like Alexa. If you're a media site or portal that has to report traffic estimates as part of your business, ComScore and Nielsen dominate traffic measurement, with Quantcast and Hitwise as smaller alternatives.

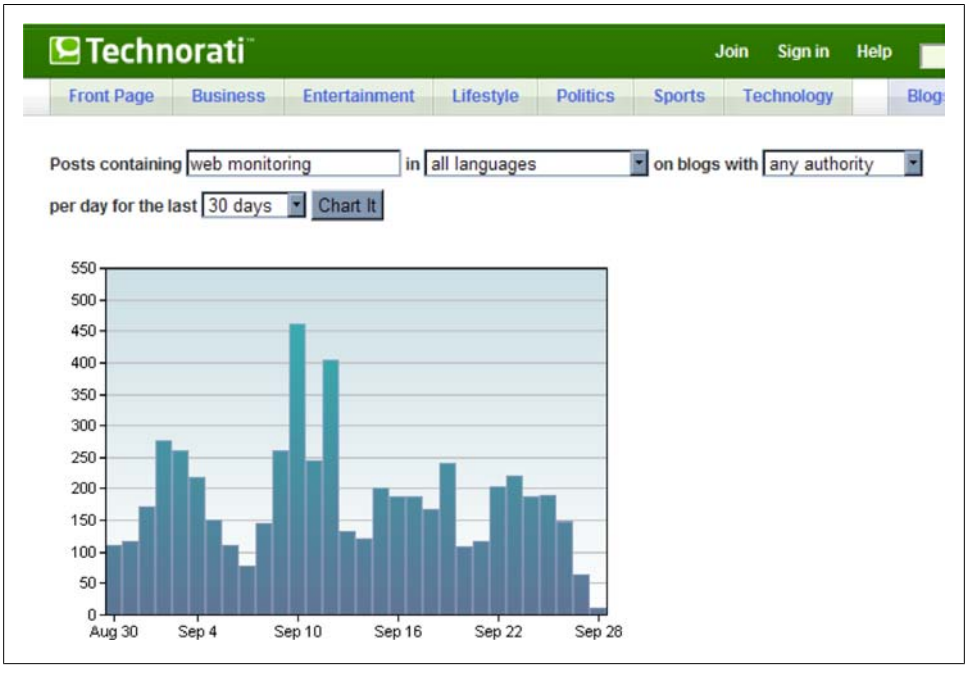

Figure 3-13. Popularity ranking for the term "web monitoring" using Technorati's BlogPulse service

Many "popularity" sites have limited accuracy. They are based on visits from a survey population that has installed a toolbar, and often don't correctly represent actual traffic.

How People Are Finding My Competitors

Your competitors are fighting you for all those visitors. Because they're probably using Google's AdWords, you can find out a good deal about what terms they're using and how much they're spending on searches. Major search engines share data on who's buying keywords and how much they're paying for them—this is a side effect of their keyword bidding model. Services like Spyfu collect this data and show the relative ad spending and keyword popularity of other companies.

Knowing which organic terms are leading visitors to your competitors helps you understand what customers are looking for and how they're thinking about your products or services. On the other hand, using a competitor's web domain, you can find out what search terms the market thinks apply to your product category and change your marketing accordingly.

What to watch: Organic and inorganic search results for key competitors.

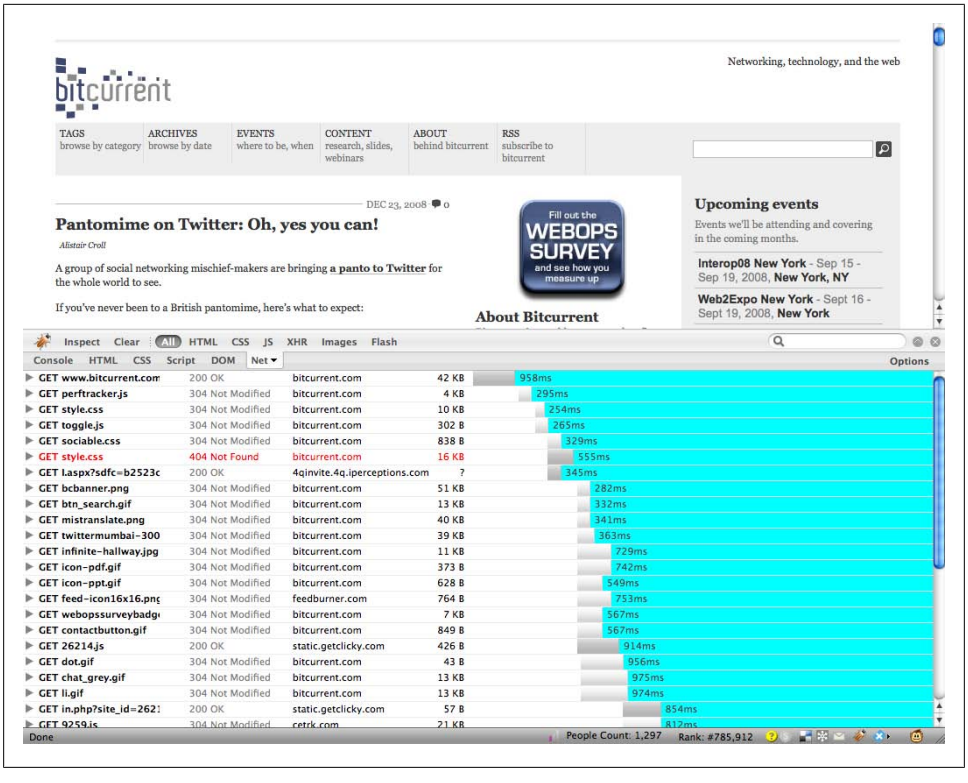

Figure 3-14. Firebug cascading performance analysis diagram of Bitcurrent.com

Relative Site Performance

Now that you have some idea of what traffic your competitors are receiving, how people are finding them, and what they're spending for traffic, you should see how well they're performing. While you can't look at their internal operations, you can compare their performance and availability to yourself and to industry benchmarks.

Synthetic testing companies like Keynote and Gomez publish scorecards of web performance (for example, *http://scorecards.keynote.com* and *http://benchmarks.gomez.com*) at regular intervals that range from simple latency comparisons to detailed multifactor studies.

While these reports give you a good indication of what "normal" is, they're less useful for a narrow market segment. If you're not part of a benchmark index, you can still get a sense of your competitors' performance.

If you simply want to measure a competitor's site, using a browser plug-in like Firebug (*http://getfirebug.com/*) can be enough to analyze page size and load times (see Figure 3-14). This has the added benefit of showing which analytics and monitoring plugins your competitors are using.

If you want to compare your performance to others' over time, you may want to set up your own tests of their site using a synthetic testing service. You may even set up a transactional benchmark that can show the difference in performance across similar workflows.

Don't just consider the latency of individual pages when comparing yourself to others. A competitor may have an enrollment process that happens in only three steps instead of your five-step process, so consider the overall performance of the task.

Be careful when testing competitors' websites, though: some may have terms of service that prohibit you from testing them. You may also want to run your tests from a location or an IP address that isn't linked back to your organization. Not that we'd ever condone such a thing.

What to watch: Industry benchmarks; competitors' page load times; synthetic tests of competitors where feasible. Within these, track site availability, response time, and consistency (whether the site has the same response time from various locations or at various times of the day).

Competitor Activity

No discussion of competitive monitoring would be complete without discussing alerts and search engines. You can use Google Alerts to be notified whenever specific keywords, such as a competitor's brand name, appear on the Web. Additionally, a number of software tools will crawl competitors' websites and flag changes.

What to watch: Alerts for competitor names and key executives online; changes to competitors' pages with business impact such as pricing information, financing, media materials, screenshots, and executive teams.

Where Are My Risks?

Any online presence carries risks. As soon as you engage your market via the Web, aggrieved customers and anonymous detractors can attack you publicly. You may also expose yourself to legal liability and have to monitor your website for abusive content left by others.

Trolling and Spamming

On the Web, everyone's got an opinion—and you probably don't agree with all of them. Any website that offers comment fields, collaboration, and content sharing will become a target for two main groups of mischief-makers: spammers and trolls.

Spammers want to pollute your site with irrelevant content and links to sites. They want to generate inbound links to their sites from as many places as possible in an effort to raise their site's rankings or influence search engines. This is only getting worse.

According to Akismet Wordpress's stats found at *http://www.akismet.com/stats*, SPAM comments have been on a steady rise since they started tracking spammy entries in early 2006.

To combat this, many search engines' web crawlers ignore any links that have a specific *nofollow* tag in them, and blogs that allow commenters to add links routinely mark them with this tag. Nevertheless, blog comment spam is a major source of activity on sites; not only does it need to be blocked, but it must be accounted for in web analytics, since spammers' scripts don't help the business, but they may count as visits.

In early 2005, Google developed the nofollow tag for the rel attribute of HTML link and anchor elements. Google does not consider links with the nofollow tag for the purposes of PageRank (*http://en.wikipedia.org/wiki/Spamdexing*).

Trolls are different beasts entirely. Wikipedia describes trolling as a "deliberate violation of the implicit rules of Internet social spaces," and defines trolls as people who are "deliberately inflammatory on the Internet in order to provoke a vehement response from other users." Since Wikipedia is one of the largest community sites on the Internet, it has its own special troll problem, and a page devoted to Wikipedia trolls (*http://meta.wikimedia.org/wiki/What_is_a_troll%3F*).

While common wisdom says to ignore the activity of trolls and to block spammers, you should still care about them for several reasons:

- They make the site less appealing for legitimate users.
- They consume resources, such as computing, bandwidth, and storage.
- If your site contains spammy content, search engines may consider it less relevant, and your search rankings will drop.
- You may be liable for harmful, offensive, or copyrighted content others post on your site.

Your antispam software will probably provide reports on the volume of spam it has blocked, as shown in Figure 3-15.

Dealing with spammers and trolls is the job of a community manager. In systems that require visitors to log in before posting, spam is easier to control, since the majority of spam comes from automated scripts run on a hijacked machine rather than from users with validated accounts.

How do you detect spammers and trolls? One way, employed by most antispam tools, is to examine the content they leave, which may contain an excessive number of links or specific keywords. A second approach is to look at their behavior. Spammers and trolls may comment on many posts with similar content, move quickly between topic

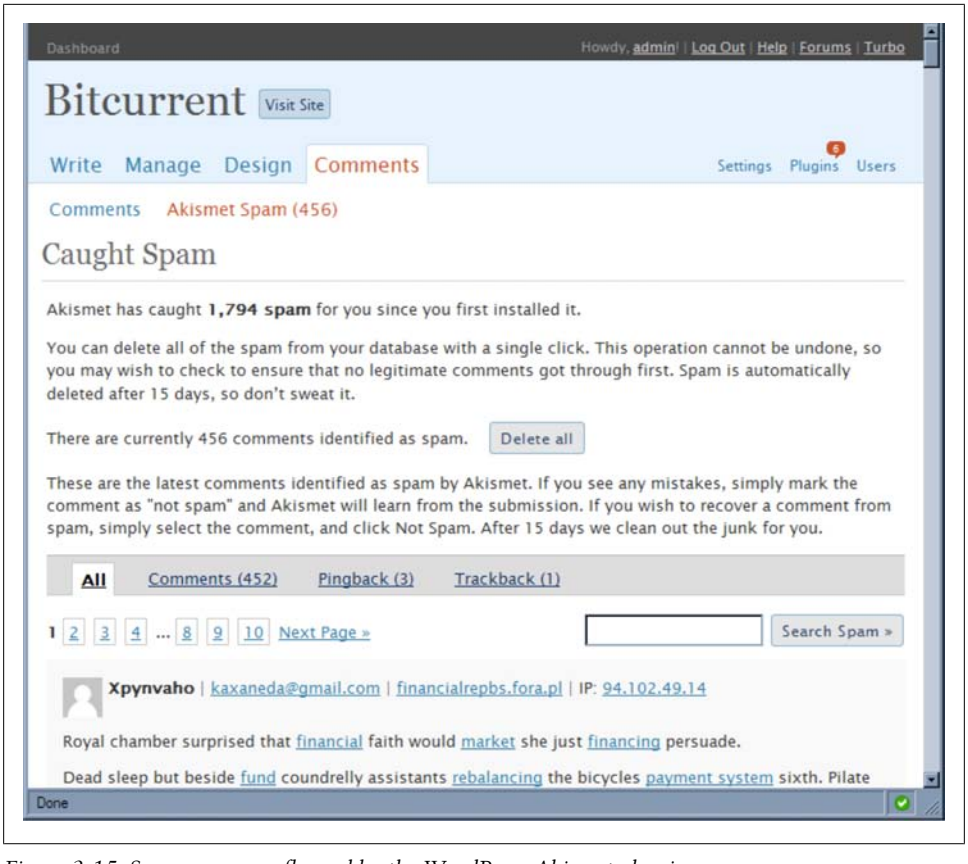

Figure 3-15. Spam messages flagged by the WordPress Akismet plug-in

areas, or befriend many community members without having that friendship reciprocated.

A far more effective approach is to harness the power of the community itself. Sites like Craigslist invite visitors to flag inappropriate content so that editors and community managers can intervene. You can capture the rate of flagged content as a metric of how much troll and spammer activity your website is experiencing.

Since every site has slightly different metrics and interaction models, you will likely have to work with your engineering team to build tools that track these unwanted visitors.

What to watch: Number of users that exhibit unwanted behaviors; percent of spammy comments; traffic sources that generate spam; volume of community flags.

Copyright and Legal Liability

If your site lets users post content, you may have to take steps to ensure that this content isn't subject to copyright from other organizations. Best practices today are to ask users to confirm that they are legally permitted to post the content, and to provide links for someone to report illegal content.

It's hard to say what's acceptable in a quickly changing online world. Services like Gracenote and MusicDNS can recognize copies of music, regardless of format, and help license holders detect and enforce copyright. Yet the Center for Social Media (*http://www.centerforsocialmedia.org/files/pdf/CSM_Recut_Reframe_Recycle_report.pdf*) points out that "a substantial amount of user-generated video uses copyrighted material in ways that are eligible for fair use consideration, although no coordinated work has yet been done to understand such practices through the fair use lens."

So as a web operator, you need to track the content users upload and quickly review, and possibly remove, content that has been flagged by the user community.

What to watch: Users who upload significantly more content than others; most popular content; content that has been flagged for review.

Fraud, Privacy, and Account Sharing

If your site contains personally identifiable information, worry about privacy. Safeguarding your visitors' data is more than just good practice—in much of the world, it's a legal obligation.

As with other site-specific questions, you will probably need to work with the development team or your security experts to flag breaches in privacy or cases of fraud. The best thing you can do is to be sure you've got plenty of detailed logfiles on hand that can be searched when problems arise.

One type of fraud that web operators should watch directly, however, is *account sharing*. Many applications—particularly SaaS and paid services—are priced per seat. Subscribers may be tempted to share their accounts, particularly with "utility" services such as online storage, real estate listing services, or analyst firm portals.

We've seen one case of a user who shared his paid account to an analyst's website with his development team, in violation of the site's terms of service. The user carefully coordinated his logins so employees never used the account at the same time. The fraudulent use of the account was only discovered when a web administrator noticed that the user seemed to be traveling from Sunnyvale to Mumbai and back every day.

Pinpointing account fraud can be challenging, but there are some giveaways:

- Accounts that log in while a user is already logged in (concurrent use).
- Accounts that log in from several geographic regions relatively quickly.
- A high variety of browser user agents associated with one account.

- Users who log in despite being terminated from the company that purchased the account.

To track down violators, generate reports that identify these accounts, and provide this data to sales or support teams who can contact offenders, offer to upsell their accounts, or even demand additional payment for truly egregious violations.

What to watch: Number of concurrent-use logins per account; number of states from which a user has logged in; number of different user agents seen for an account.

What Are People Saying About Me?

On an increasingly social web, your marketing has migrated beyond your website into chat rooms, user groups, social networking applications, news aggregators, and blogs. To properly understand what the Internet thinks of you, you need to watch the Web beyond your front door.

Your primary tool for this task is search. Hundreds of automated scripts crawl the Web constantly, indexing what they find. And most community sites offer some form of internal content search these days. A variant on traditional search, known as *persistent search* or *prospective search*, combs the Web for new content that matches search terms, then informs you of it.

Google Alerts dominates the prospective search market, with some other services, such as Rollyo, using competing search engines like Yahoo!. HubSpot offers integrated marketing to small businesses by combining topical searches, lead tracking, and similar functions, as shown in Figure 3-16. There are also many community listening platforms—Radian6, Techrigy, ScoutLabs, Sysomos, Keenkong and so on—that help community managers monitor their online buzz.

Figure 3-16. HubSpot ties community traffic to web analytics

In all of these models, you subscribe to a keyword across various types of sites (blogs, mailing lists, news aggregators, etc.), then review the results wherever someone is talking about things that matter to your organization.

Site Reputation

In the era of search, nothing matters more than what Google thinks of you. As Figure 3-17 shows, if you're the best result for a particular search term, you don't even pay for your advertising.

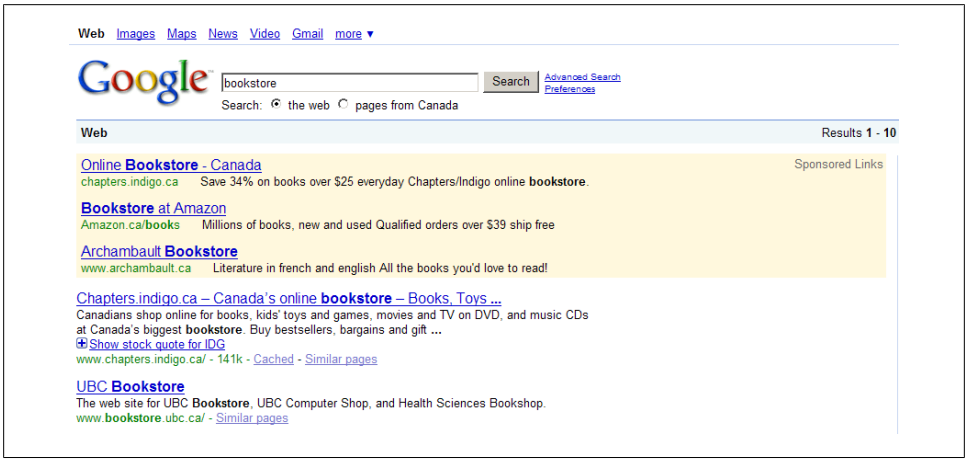

Figure 3-17. Google's ranking of Chapters as a bookstore means the company doesn't have to pay for advertising the way Amazon does in this search

Google's PageRank is a measure of how relevant and significant Google thinks your website is, and encompasses factors such as the number of inbound links to the site and the content on the site itself. The PageRank algorithm Coca-Cola is closely guarded and constantly evolving to stay ahead of unscrupulous site promoters.

Other sites, such as Technorati, use their own approaches, based on similar factors such as the number of inbound links in the past six months.

What to watch: Google PageRank; Technorati ranking; StumbleUpon rating; other Internet ranking tools.

Trends

Google Trends and Yahoo! Buzz show the popularity of search terms, and Google Insights, shown in Figure 3-18, breaks them down over time. If you want to understand the relative popularity of content on the Internet in order to optimize the wording of your site or to downplay aging themes, these tools are useful for more than just competitive analysis.

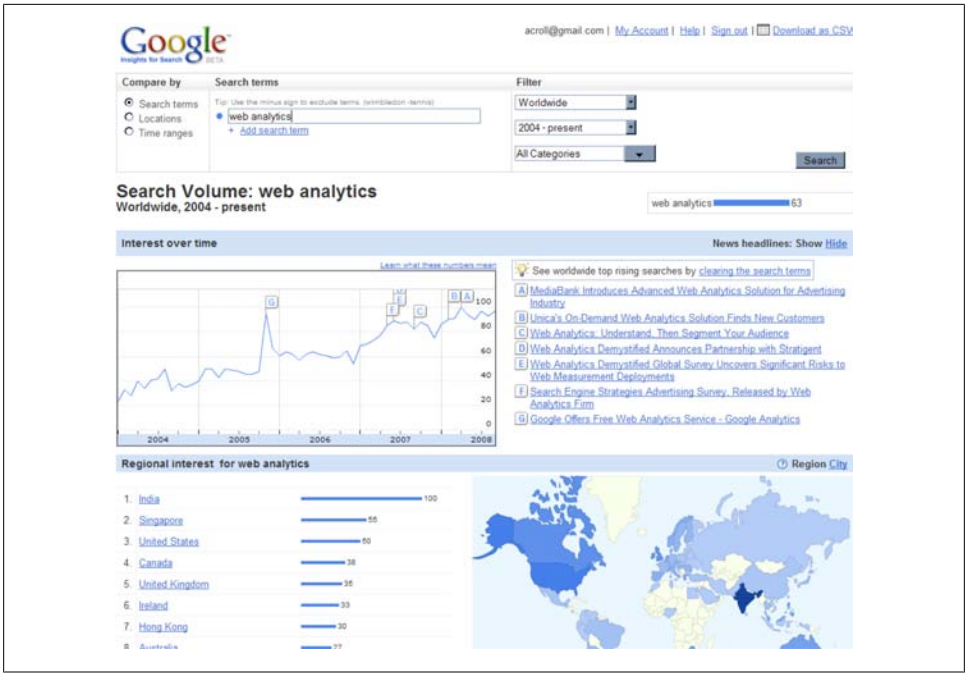

Figure 3-18. Google Insights, showing spread and growth of searches for "web analytics" worldwide

What to watch: Mentions of yourself and your competitors over time; product names; subject matter you cover.

Social Network Activity

Many social networks have a built-in search you can use to see whether your company, site, or products are being discussed, as shown in Figure 3-19.

What to watch: Search results for your company name, URL, product names, executives, and relevant keywords across social sites like Digg, Summize, and Twitter, as well as any that are relevant to your particular industry or domain.

How Are My Site and Content Being Used Elsewhere?

Other people are using your site, and you don't know it. They may be doing so as part of a mashup. They may be running search engine crawlers to index your content. Or they may be competitors checking up on you. Whatever the case, you need to track and monitor them.

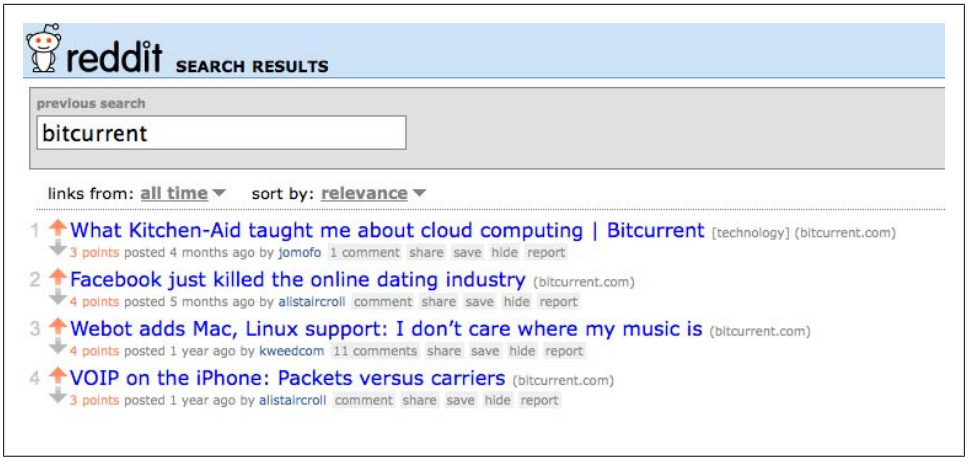

Figure 3-19. A search for a specific topic using reddit's internal search

API Access and Usage

Your site may offer formal web services or Application Programming Interfaces (APIs) to let your users access your application programmatically through automated scripts. Most sites have at least one trivially simple web service—the RSS feed, which is simply an XML file retrieved via HTTP.

Most real web services are more "heavyweight" than syndication feeds, but even the most basic web services need to be monitored. If letting people extend your web application with their own code is important to your business, monitor the APIs to ensure they are reliable and that people are using them in appropriate ways.

Some RSS management tools, like FeedBurner, will showcase "unusual" uses of your feed automatically—for example, someone who is pulling an RSS feed into Yahoo! Pipes for postprocessing, as shown in Figure 3-20.

If your web services include terms of service that limit how much someone can use them, track offending users before a greedy third-party developer breaks your site. Note that many analytics packages that rely on JavaScript can't be used to track API calls because there's no way to reliably embed JavaScript in the content they deliver.

What to watch: Traffic volume and number of requests for each API you offer; number of failed authentications to the API; number of API requests by developer; top URLs by traffic and request volume.

Mashups, Stolen Content, and Illegal Syndication

Your site's data can easily appear online in a mashup. By combining several sites and services, web users can create a new application, often without the original sites

Figure 3-20. FeedBurner's Uncommon Uses report can show you ways in which others are using your RSS feed

knowing it. Tools like Yahoo! Pipes or Microsoft's Popfly make it easy to assemble several services without any programming knowledge.

Some mashups are well intentioned, even encouraged. Google makes it easy for developers to embed and augment their maps in third-party sites. Other repurposing of content, such as republishing blog posts or embedding photos from third-party sites, may not be so innocent. If someone else reposts your content, your search ranking goes down, and if someone else embeds links to media you're hosting, you pay for the bandwidth.

If this is happening to you, you'll see referring URLs belonging to the mashup page, and you can track back to that URL to determine where the traffic is coming from and take action if needed. You may have to look in server logs or on a sniffer, because if the mashup is pulling in a component like a video or an image, you won't see any sign of it in JavaScript-based monitoring.

Although the term "Sniffer" is a registered trademark of Network General Corporation, it is also used by many networking professionals to refer to packet capture devices in general.

Try to treats mashups as business opportunities, not threats. If you have interesting content, find a way to deliver it that benefits both you and the mashup site.

What to watch: URLs with a high volume of requests for an object on your site without the normal entry path; search results containing your unique content.

Integration with Legacy Systems

Some SaaS applications may connect to their subscribers' enterprise software through dedicated links between the subscriber's data center and the SaaS application in order to exchange customer, employee, and financial data. Relying on these kinds of third-party services can affect the performance and availability of your website. Travel search sites, for example, are highly dependent on airline and hotel booking systems for their search results.

While not directly related to web monitoring, the performance of these third-party connections must be tested and reported as part of an SLA. Excessive use of enterprise APIs or long delays from third-party services may degrade the performance of the website. Keep nonweb activity in mind when monitoring the web-facing side of the business, and find ways to track private API calls to enterprise clients.

What to watch: Volume and performance of API calls between the application and enterprise customers or data partners.

The Tools at Our Disposal

Clearly, there are many metrics and KPIs to gather in order to understand and improve your online presence. Fortunately, there's a wide range of tools available for watching yourself online. The trick is to use the right technologies to collect the metrics that matter the most to your business.

At the broadest level, there are three categories of monitoring technology you can use to understand web activity. You can *collect* what users do from various points in the web connection; you can use *search* engines that crawl and index the web, and may send you alerts for changes or keywords; and you can run scripts that *test* your site directly.

Collection Tools

There are many ways to collect visitor information, depending on how much access you have to your servers, the features of those visitors' browsers, and the kind of data you wish to collect.

Collection can happen on your own machines, on intermediate devices that collect a copy of traffic, through a browser toolbar, or on the visitor's browser. It may also happen with third-party services like FeedBurner (for RSS feeds) or Mashery (for APIs and web services) that proxy your content or manage your APIs.

The volume of data collected in these ways grows proportionally with traffic. Collecting data may slow down your servers, and may also pose privacy risks. You also need to consider what various collection methods can see, since they all have different per-

spectives. An inline sniffer can't see client-side page load time, for example; similarly, client-side JavaScript can't see server errors.

Search Systems

Another way to monitor your online presence is through the use of search engines that run scripts—called *crawlers*—that visit web pages and follow links on sites to collect and index the data they find.

There are hundreds of these crawlers on the Web. Some of them feed search giants like Google, Yahoo!, and MSN. More specialized crawlers also look for security problems, copyright violations, plagiarism, contact information, archiving, and so on.

Crawlers can't index the entire Web. Many websites are closed to crawlers, either because they require a login, because of their dynamic nature, or because they've blocked crawlers from indexing some of the content. This is common for news sites and blog comment threads. As a result, you need to use site-specific internal search tools alongside global search engines for complete coverage of web activity.

While we may run searches to see what's happening online, a more practical way to manage many search queries is to set up alerts when certain keywords arise or when specific pages change.

Testing Services

In addition to collecting visitor data and setting up searches and alerts, you can run tests against websites to measure their health or check their content. These tests can simulate specific browsers or run from several geographic regions to pinpoint problems with a specific technology or location.

Testing services can also watch your competitors or monitor third-party sites, like payment or mapping, on which your main application depends. There's a wide range of such services available, from open source, roll-your-own scripting to global testing networks that can verify a site's health from almost anywhere and run complex, multistep transactions.

Ultimately, the many metrics we've looked at above, using the collection, search, and testing approaches outlined here, give us complete web visibility. That visibility amounts to four big questions, which we'll look at next.