Secure Spread Spectrum Watermarking for Multimedia

Ingemar J. Cox, *Senior Member, IEEE,* Joe Kilian, F. Thomson Leighton, and Talal Shamoon, *Member, IEEE*

***Abstract*— This paper presents a secure (tamper-resistant) algorithm for watermarking images, and a methodology for digital watermarking that may be generalized to audio, video, and multimedia data. We advocate that a watermark should be constructed as an independent and identically distributed (i.i.d.) Gaussian random vector that is imperceptibly inserted in a spread-spectrum-like fashion into the perceptually *most* significant spectral components of the data. We argue that insertion of a watermark under this regime makes the watermark robust to signal processing operations (such as lossy compression, filtering, digital-analog and analog-digital conversion, requantization, etc.), and common geometric transformations (such as cropping, scaling, translation, and rotation) provided that the original image is available and that it can be succesfully registered against the transformed watermarked image. In these cases, the watermark detector unambiguously identifies the owner. Further, the use of Gaussian noise, ensures strong resilience to multiple-document, or collusional, attacks. Experimental results are provided to support these claims, along with an exposition of pending open problems.**

***Index Terms*— Intellectual property, fingerprinting, multimedia, security, steganography, watermarking.**

I. Introduction

THE PROLIFERATION of digitized media (audio, image, and video) is creating a pressing need for copyright enforcement schemes that protect copyright ownership. Conventional cryptographic systems permit only valid keyholders access to encrypted data, but once such data is decrypted there is no way to track its reproduction or retransmission. Therefore, conventional cryptography provides little protection against data piracy, in which a publisher is confronted with unauthorized reproduction of information. A digital watermark is intended to complement cryptographic processes. It is a visible, or preferably invisible, identification code that is permanently embedded in the data and remains present within the data after any decryption process. In the context of this work, data refers to audio (speech and music), images (photographs and graphics), and video (movies). It does not include ASCII representations of text, but does include text represented as an image. Many of the properties of the scheme presented in this work may be adapted to accommodate audio and video implementations, but the algorithms here specifically apply to images.

A simple example of a digital watermark would be a visible "seal" placed over an image to identify the copyright owner (e.g., [2]). A visible watermark is limited in many ways. It marrs the image fidelity and is susceptible to attack through direct image processing. A watermark may contain additional information, including the identity of the purchaser of a particular copy of the material. In order to be effective, a watermark should have the characteristics outlined below.

Unobtrusiveness: The watermark should be perceptually invisible, or its presence should not interfere with the work being protected.

Robustness: The watermark must be difficult (hopefully impossible) to remove. If only partial knowledge is available (for example, the exact location of the watermark in an image is unknown), then attempts to remove or destroy a watermark should result in severe degradation in fidelity before the watermark is lost. In particular, the watermark should be robust in the following areas.

- *Common signal processing:* The watermark should still be retrievable even if common signal processing operations are applied to the data. These include, digital-to-analog and analog-to-digital conversion, resampling, requantization (including dithering and recompression), and common signal enhancements to image contrast and color, or audio bass and treble, for example.
- *Common geometric distortions (image and video data):* Watermarks in image and video data should also be immune from geometric image operations such as rotation, translation, cropping and scaling.
- *Subterfuge attacks (collusion and forgery):* In addition, the watermark should be robust to collusion by multiple individuals who each possess a watermarked copy of the data. That is, the watermark should be robust to combining copies of the same data set to destroy the watermarks. Further, if a digital watermark is to be used in litigation, it must be impossible for colluders to combine their images to generate a different valid watermark with the intention of framing a third party.

Manuscript received January 14, 1996; revised January 24, 1997. Portions of this work were reprinted, with permission, from the Proceedings of the IEEE Conference on Image Processing, 1996, and from the Proceedings of the First International Conference on Data Hiding (Springer-Verlag, 1996). The associate editor coordinating the reivew of this manuscript and approving it for publication was Prof. Sarah Rajala.

I. J. Cox and J. Kilian are with NEC Research Institute, Princeton, NJ 08540 USA (e-mail: ingemar@research.nj.nec.com; joe@research.nj.nec.com).

F. T. Leighton is with the Mathematics Department and Laboratory for Computer Science, The Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: ftl@math.mit.edu).

T. Shamoon is with InterTrust STAR Laboratory, Sunnyvale, CA 94086 USA (e-mail: talal@intertrust.com).

Publisher Item Identifier S 1057-7149(97)08460-1.

Universality: The same digital watermarking algorithm should apply to all three media under consideration. This is potentially helpful in the watermarking of multimedia products. Also, this feature is conducive to implementation of audio and image/video watermarking algorithms on common hardware.

Unambiguousness: Retrieval of the watermark should unambiguously identify the owner. Furthermore, the accuracy of owner identification should degrade gracefully in the face of attack.

There are two parts to building a strong watermark: the *watermark structure* and the *insertion strategy*. In order for a watermark to be robust and secure, these two components must be designed correctly. We provide two key insights that make our watermark both robust and secure: We argue that the watermark be placed explicitly in the perceptually most significant components of the data, and that the watermark be composed of random numbers drawn from a Gaussian $(N(0,1))$ distribution.

The stipulation that the watermark be placed in the perceptually significant components means that an attacker must target the fundamental structural components of the data, thereby heightening the chances of fidelity degradation. While this strategy may seem counterintuitive from the point of view of steganography (how can these components hide any signal?), we discovered that the significant components have a *perceptual capacity* that allows watermark insertion without perceptual degradation. Further, most processing techniques applied to media data tend to leave the perceptually significant components intact. While one may choose from a variety of such components, in this paper, we focus on the perceptually significant *spectral* components of the data. This simultaneously yields high perceptual capacity and achieves a uniform spread of watermark energy in the pixel domain.

The principle underlying our watermark structuring strategy is that the mark be constructed from independent, identically distributed (i.i.d.) samples drawn from a Gaussian distribution. Once the significant components are located, Gaussian noise is injected therein. The choice of this distribution gives resilient performance against collusion attacks. The Gaussian watermark also gives our scheme strong performance in the face of quantization, and may be structured to provide low false positive and false negative detection. This is discussed below, and elaborated on in [13].

Finally, note that the techniques presented herein do not provide proof of content ownership on their own. The focus of this paper are algorithms that insert messages into content in an extremely secure and robust fashion. Nothing prevents someone from inserting another message and claiming ownership. However, it is possible to couple our methods with strong authentication and other cryptographic techniques in order to provide complete, secure and robust owner identification and authentication.

Section III begins with a discussion of how common signal transformations, such as compression, quantization, and manipulation, affect the frequency spectrum of a signal. This discussion motivates our belief that a watermark should be components. Of course, the major problem then becomes how to imperceptibly insert a watermark into perceptually significant components of the frequency spectrum. Section III-A proposes a solution based on ideas from spread spectrum communications. In particular, we present a watermarking algorithm that relies on the use of the original image to extract the watermark. Section IV provides an analysis based on possible collusion attacks that indicates that a binary watermark is not as robust as a continuous one. Furthermore, we show that a watermark structure based on sampling drawn from multiple i.i.d Gaussian random variables offers good protection against collusion. Ultimately, no watermarking system can be made perfect. For example, a watermark placed in a textual image may be eliminated by using optical character recognition technology. However, for common signal and geometric distortions, the experimental results of Section V suggest that our system satisfies most of the properties discussed in the introduction, and displays strong immunity to a variety of attacks in a collusion resistant manner. Finally, Section VI discusses possible weaknesses and potential enhancements to the system and describes open problems and subsequent work.

II. Previous Work

Several previous digital watermarking methods have been proposed. Turner [25] proposed a method for inserting an identification string into a digital audio signal by substituting the "insignificant" bits of randomly selected audio samples with the bits of an identification code. Bits are deemed "insignificant" if their alteration is inaudible. Such a system is also appropriate for two-dimensional (2-D) data such as images, as discussed in [26]. Unfortunately, Turner's method may easily be circumvented. For example, if it is known that the algorithm only affects the least significant two bits of a word, then it is possible to randomly flip *all* such bits, thereby destroying any existing identification code.

Caronni [6] suggests adding *tags*—small geometric patterns—to digitized images at brightness levels that are imperceptible. While the idea of hiding a spatial watermark in an image is fundamentally sound, this scheme may be susceptible to attack by filtering and redigitization. The fainter such watermarks are, the more susceptible they are such attacks and geometric shapes provide only a limited alphabet with which to encode information. Moreover, the scheme is not applicable to audio data and may not be robust to common geometric distortions, especially cropping.

Brassil *et al.* [4] propose three methods appropriate for document images in which text is common. Digital watermarks are coded by 1) vertically shifting text lines, 2) horizontally shifting words, or 3) altering text features such as the vertical endlines of individual characters. Unfortunately, all three proposals are easily defeated, as discussed by the authors. Moreover, these techniques are restricted exclusively to images containing text.

Tanaka *et al.* [19], [24] describe several watermarking schemes that rely on embedding watermarks that resemble quantization noise. Their ideas hinge on the notion that quan-

first scheme injects a watermark into an image by using a predetermined data stream to guide level selection in a predictive quantizer. The data stream is chosen so that the resulting image looks like quantization noise. A variation on this scheme is also presented, where a watermark in the form of a dithering matrix is used to dither an image in a certain way. There are several drawbacks to these schemes. The most important is that they are susceptible to signal processing, especially requantization, and geometric attacks such as cropping. Furthermore, they degrade an image in the same way that predictive coding and dithering can.

In [24], the authors also propose a scheme for watermarking facsimile data. This scheme shortens or lengthens certain runs of data in the run length code used to generate the coded fax image. This proposal is susceptible to digital-to-analog and analog-to-digital attacks. In particular, randomizing the least significant bit (LSB) of each pixel's intensity will completely alter the resulting run length encoding. Tanaka *et al.* also propose a watermarking method for "color-scaled picture and video sequences". This method applies the same signal transform as the Joint Photographers Expert Group (JPEG) (discrete cosine transform of 8×8 subblocks of an image) and embeds a watermark in the coefficient quantization module. While being compatible with existing transform coders, this scheme may be susceptible to requantization and filtering and is equivalent to coding the watermark in the LSB's of the transform coefficients.

In a recent paper, Macq and Quisquater [18] briefly discuss the issue of watermarking digital images as part of a general survey on cryptography and digital television. The authors provide a description of a procedure to insert a watermark into the least significant bits of pixels located in the vicinity of image contours. Since it relies on modifications of the least significant bits, the watermark is easily destroyed. Further, their method is restricted to images, in that it seeks to insert the watermark into image regions that lie on the edge of contours. Bender *et al.* [3] describe two watermarking schemes. The first is a statistical method called *patchwork*. Patchwork randomly chooses n pairs of image points, (a_i, b_i), and increases the brightness at a_i by one unit while correspondingly decreasing the brightness of b_i. The expected value of the sum of the differences of the n pairs of points is then $2n$, provided certain statistical properties of the image are true.

The second method is called "texture block coding," wherein a region of random texture pattern found in the image is copied to an area of the image with similar texture. Autocorrelation is then used to recover each texture region. The most significant problem with this technique is that it is only appropriate for images that possess large areas of random texture. The technique could not be used on images of text, for example, nor is there a direct analog for audio.

Rhoads [21] describes a method that adds or subtracts small random quantities from each pixel. Addition or subtraction is determined by comparing a binary mask of L bits with the LSB of each pixel. If the LSB is equal to the corresponding mask bit, then the random quantity is added, otherwise it is subtracted. The watermark is subtracted by first computing and then by examining the sign of the difference, pixel by pixel, to determine if it corresponds to the original sequence of additions and subtractions. This method does not make use of perceptual relevance, but it is proposed that the high frequency noise be prefiltered to provide some robustness to lowpass filtering. This scheme does not consider the problem of collusion attacks.

Koch, Rindfrey, and Zhao [14] propose two general methods for watermarking images. The first method, attributed to Scott Burgett, breaks up an image into 8×8 blocks and computes the discrete cosine transform (DCT) of each of these blocks. A pseudorandom subset of the blocks is chosen, then, in each such block, a triple of frequencies is selected from one of 18 predetermined triples and modified so that their relative strengths encode a one or zero value. The 18 possible triples are composed by selection of three out of eight predetermined frequencies within the 8×8 DCT block. The choice of the eight frequencies to be altered within the DCT block is based on a belief that the "middle frequencies...have moderate variance," i.e. they have similar magnitude. This property is needed in order to allow the relative strength of the frequency triples to be altered without requiring a modification that would be perceptually noticeable. Superficially, this scheme is similar to our own proposal, also drawing an analogy to spread spectrum communications. However, the structure of their watermark is different from ours, and the set of frequencies is not chosen based on any direct perceptual significance, or relative energy considerations. Further, because the variance between the eight frequency coefficients is small, one would expect that their technique may be sensitive to noise or distortions. This is supported by the experimental results that report that the "embedded labels are robust against JPEG compression for a quality factor as low as about 50%." By comparison, we demonstrate that our method performs well with compression quality factors as low as 5%. An earlier proposal by Koch and Zhao [15] used not triples of frequencies but pairs of frequencies, and was again designed specifically for robustness to JPEG compression. Nevertheless, they state that "a lower quality factor will increase the likelihood that the changes necessary to superimpose the embedded code on the signal will be noticeably visible." In a second method, designed for black and white images, no frequency transform is employed. Instead, the selected blocks are modified so that the relative frequency of white and black pixels encodes the final value. Both watermarking procedures are particularly vulnerable to multiple document attacks. To protect against this, Zhao and Koch propose a *distributed* 8×8 block created by randomly sampling 64 pixels from the image. However, the resulting DCT has no relationship to that of the true image and consequently may be likely to cause noticeable artifacts in the image and be sensitive to noise.

In addition to direct work on watermarking images, there are several works of interest in related areas. Adelson [1] describes a technique for embedding digital information in an analog signal for the purpose of inserting digital data into an analog TV signal. The analog signal is quantized into one of two disjoint ranges ($\{0, 2, 4, \ldots\}$, $\{1, 3, 5, \ldots\}$, for example) that

Adelson's method is equivalent to watermark schemes that encode information into the LSB's of the data or its transform coefficients. Adelson recognizes that the method is susceptible to noise and therefore proposes an alternative scheme wherein a 2×1 Hadamard transform of the digitized analog signal is taken. The differential coefficient of the Hadamard transform is offset by zero or one unit prior to computing the inverse transform. This corresponds to encoding the watermark into the least significant bit of the differential coefficient of the Hadamard transform. It is not clear that this approach would demonstrate enhanced resilience to noise. Furthermore, like all such LSB schemes, an attacker can eliminate the watermark by randomization.

Schreiber *et al.* [22] describe a method to interleave a standard NTSC signal within an enhanced definition television (EDTV) signal. This is accomplished by analyzing the frequency spectrum of the EDTV signal (larger than that of the NTSC signal) and decomposing it into three subbands (L, M, H for low-, medium- and high-frequency, respectively). In contrast, the NTSC signal is decomposed into two subbands, L and M. The coefficients, M_k, within the M band are quantized into m levels and the high frequency coefficients, H_k, of the EDTV signal are scaled such that the addition of the H_k signal plus any noise present in the system is less than the minimum separation between quantization levels. Once more, the method relies on modifying least significant bits. Presumably, the midrange rather than low frequencies were chosen because these are less perceptually significant. In contrast, the method proposed here modifies the *most* perceptually significant components of the signal.

Finally, it should be noted that existing techniques are generally not resistant to collusion attacks by multiple documents.

III. Watermarking in the Frequency Domain

In order to understand the advantages of a frequency-based method, it is instructive to examine the processing stages that an image (or sound) may undergo in the process of copying, and to study the effect that these stages could have on the data, as illustrated in Fig. 1. In the figure, "transmission" refers to the application of any source or channel code, and/or standard encryption technique to the data. While most of these steps are information lossless, many compression schemes (JPEG, MPEG, etc.) are lossy, and can potentially degrade the data's quality, through *irretrievable* loss of information. In general, a watermarking scheme should be resilient to the distortions introduced by such algorithms.

Lossy compression is an operation that usually eliminates perceptually nonsalient components of an image or sound. Most processing of this sort takes place in the frequency domain. In fact, data loss usually occurs among the high-frequency components.

After receipt, an image may endure many common transformations that are broadly categorized as geometric distortions or signal distortions. Geometric distortions are specific to images and video, and include such operations as rotation, translation, scaling and cropping. By manually determining a

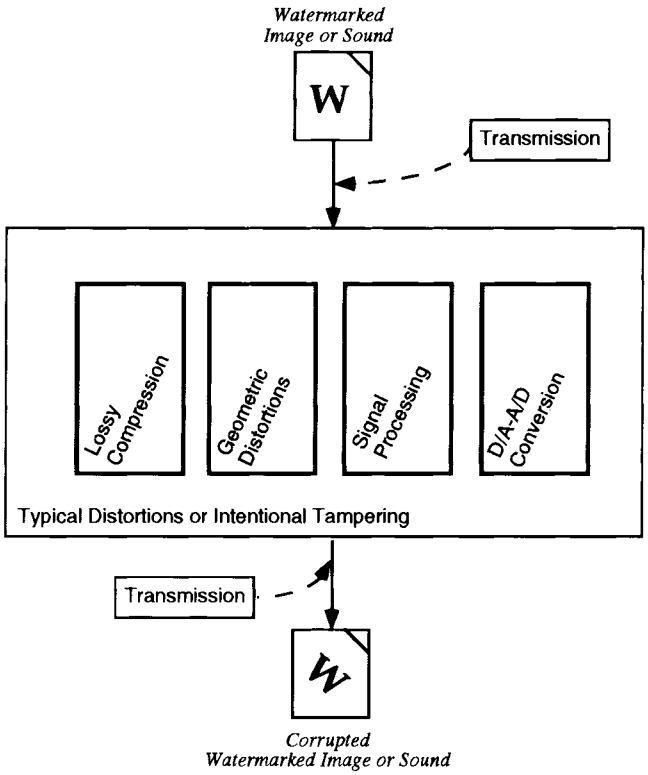

Fig. 1. Common processing operations that a media document could undergo.

original and the distorted watermark, it is possible to remove any two or three-dimensional (3-D) affine transformation [8]. However, an affine scaling (shrinking) of the image leads to a loss of data in the high-frequency spectral regions of the image. Cropping, or the cutting out and removal of portions of an image, leads to irretrievable loss of image data, which may seriously degrade any spatially based watermark such as [6]. However, a frequency-based scheme spreads the watermark over the whole spatial extent of the image, and is therefore less likely to be affected by cropping, as demonstrated in Section V-E.

Common signal distortions include digital-to-analog and analog-to-digital conversion, resampling, requantization, including dithering and recompression, and common signal enhancements to image contrast and/or color, and audio frequency equalization. Many of these distortions are nonlinear, and it is difficult to analyze their effect in either a spatial- or frequency-based method. However, the fact that the original image is known allows many signal transformations to be undone, at least approximately. For example, histogram equalization, a common nonlinear contrast enhancement method, may be removed substantially by histogram specification [10] or dynamic histogram warping [7] techniques.

Finally, the copied image may not remain in digital form. Instead, it is likely to be printed, or an analog recording made (onto analog audio or video tape). These reproductions introduce additional degradation into the image that a watermarking scheme must be robust to.

The watermark must not only be resistant to the inadvertent

be immune to intentional manipulation by malicious parties. These manipulations can include combinations of the above distortions, and can also include collusion and forgery attacks, which are discussed in Section IV-E.

A. Spread Spectrum Coding of a Watermark

The above discussion illustrates that the watermark should *not* be placed in perceptually insignificant regions of the image (or its spectrum), since many common signal and geometric processes affect these components. For example, a watermark placed in the high-frequency spectrum of an image can be easily eliminated with little degradation to the image by any process that directly or indirectly performs lowpass filtering. The problem then becomes how to insert a watermark into the most perceptually significant regions of the spectrum in a fidelity preserving fashion. Clearly, any spectral coefficient may be altered, provided such modification is small. However, very small changes are very susceptible to noise.

To solve this problem, the frequency domain of the image or sound at hand is viewed as a *communication channel*, and correspondingly, the watermark is viewed as a signal that is transmitted through it. Attacks and unintentional signal distortions are thus treated as noise that the immersed signal must be immune to. While we use this methodology to hide watermarks in data, the same rationale can be applied to sending any type of message through media data.

We originally conceived our approach by analogy to spread spectrum communications [20]. In spread spectrum communications, one transmits a narrowband signal over a much larger bandwidth such that the signal energy present in any single frequency is undetectable. Similarly, the watermark is spread over very many frequency bins so that the energy in any one bin is very small and certainly undetectable. Nevertheless, because the watermark verification process knows the location and content of the watermark, it is possible to concentrate these many weak signals into a single output with high signal-to-noise ratio (SNR). However, to destroy such a watermark would require noise of high amplitude to be added to *all* frequency bins.

Spreading the watermark throughout the spectrum of an image ensures a large measure of security against unintentional or intentional attack: First, the location of the watermark is not obvious. Furthermore, frequency regions should be selected in a fashion that ensures severe degradation of the original data following any attack on the watermark.

A watermark that is well placed in the frequency domain of an image or a sound track will be practically impossible to see or hear. This will always be the case if the energy in the watermark is sufficiently small in any single frequency coefficient. Moreover, it is possible to increase the energy present in particular frequencies by exploiting knowledge of masking phenomena in the human auditory and visual systems. Perceptual masking refers to any situation where information in certain regions of an image or a sound is occluded by perceptually more prominent information in another part of the scene. In digital waveform coding, this frequency domain

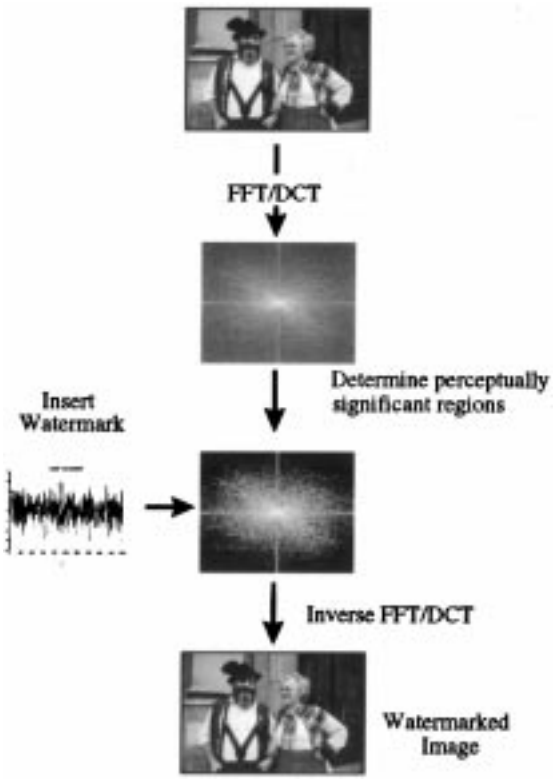

Fig. 2. Stages of watermark insertion process.

extensively to achieve low bit rate encoding of data [9], [12]. It is known that both the auditory and visual systems attach more resolution to the high-energy, low-frequency, spectral regions of an auditory or visual scene [12]. Further, spectrum analysis of images and sounds reveals that most of the information in such data is located in the low-frequency regions.

Fig. 2 illustrates the general procedure for frequency domain watermarking. Upon applying a frequency transformation to the data, a *perceptual mask* is computed that highlights perceptually significant regions in the spectrum that can support the watermark without affecting perceptual fidelity. The watermark signal is then inserted into these regions in a manner described in Section IV-B. The precise magnitude of each modification is only known to the owner. By contrast, an attacker may only have knowledge of the possible range of modification. To be confident of eliminating a watermark, an attacker must assume that each modification was at the limit of this range, despite the fact that few such modifications are typically this large. As a result, an attack creates visible (or audible) defects in the data. Similarly, unintentional signal distortions due to compression or image manipulation, must leave the perceptually significant spectral components intact, otherwise the resulting image will be severely degraded. This is why the watermark is robust.

In principle, any frequency domain transform can be used. However, in the experimental results of Section VI we use a Fourier domain method based on the DCT [16], although we are currently exploring the use of wavelet-based schemes as a variation. In our view, each coefficient in the frequency domain

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.