FOURTH AMENDMENT SEARCH AND THE POWER OF THE HASH

Richard P. Salgado[*]

Replying to Orin S. Kerr, *Searches and Seizures in a Digital World*, 119 HARV. L. REV. 531 (2005).

I. THE POWER OF HASHING

Hashing is a powerful and pervasive technique used in nearly every examination of seized digital media. The concept behind hashing is quite elegant: take a large amount of data, such as a file or all the bits on a hard drive, and use a complex mathematical algorithm to generate a relatively compact numerical identifier (the hash value) unique to that data. Examiners use hash values throughout the forensics process, from acquiring the data, through analysis, and even into legal proceedings. Hash algorithms are used to confirm that when a copy of data is made, the original is unaltered and the copy is identical, bit-for-bit. That is, hashing is employed to confirm that data analysis does not alter the evidence itself. Examiners also use hash values to weed out files that are of no interest in the investigation, such as operating system files, and to identify files of particular interest.

It is clear that hashing has become an important fixture in forensic examinations. What is not clear is whether the use of hashing implicates the Fourth Amendment, and if so, how. Is there a search when an examiner simply calculates the hash value of a hard drive in the data duplication process? Is there any Fourth Amendment moment if an examiner uses hashing to identify files on the media that are of interest but that are not within the scope of the search warrant or authority? What if those files constitute digital contraband? Does hash-based examination for child pornography fall outside the Fourth Amendment, allowing investigators throughout the country to look for child pornography on any seized electronic storage device, regardless of the nature of the investigation, without a warrant?

This Reply examines those questions within the framework of Professor Kerr's "exposure" theory of search. I conclude that when used as part of the imaging process, hashing reveals no meaningful hints as

---

[*] Lecturer in Law, Stanford Law School; Certified Instructor, SANS Institute; Member, Board of Directors, Honeynet Project; Senior Legal Director, Yahoo!, Inc. (The statements expressed herein should not be taken as a position or endorsement of Yahoo!, Inc. or its subsidiaries and may not reflect the opinion of their affiliates, joint ventures, or partners). Thanks to Kyle French for his thoughtful insights.

to the contents of the imaged media and thus is not a "search" for the purposes of the Fourth Amendment. I also conclude that in the analysis process, the use of hashing to find only files that constitute contraband does not constitute a Fourth Amendment search under the rulings of *United States v. Place*,[1] *United States v. Jacobsen*,[2] and *Illinois v. Caballes*.[3] There are important differences between the type of contraband in those cases — narcotics — and digital contraband that may make the discussion somewhat academic, at least at this juncture.

II. HASHING AND DIGITAL FORENSICS

Hashing is the process of taking an input data string (the bits on a hard drive, for example), and using a mathematical function to generate a (usually smaller) output string.[4] For example, one could take a digital wedding photo from a hard drive and calculate the hash value of the photo. Hash values can also be calculated for other data sets, including the contents of a DVD, USB drive, or an entire hard drive.

A. Key Properties of Hashing Algorithms

For computer forensics purposes, a good hashing algorithm will result in a hash value with two important properties. First, the hash value will be, for all practical purposes, uniquely associated with the input. No other file will have the same hash value as the wedding photo, except a file that is identical, bit-for-bit. If one altered the wedding photo by changing so little as one bit, the hash value of the photo would be different as well.[5] The chance of two different inputs "colliding," to use the language of cryptanalysts, is astronomically small.[6] Although research has shown cracks in MD-5 and SHA-1, two com-

---

[1] 462 U.S. 696 (1983).

[2] 466 U.S. 109 (1984).

[3] 125 S. Ct. 834 (2005).

[4] *See* BRUCE SCHNEIER, APPLIED CRYPTOGRAPHY 30 (2d ed. 1996).

[5] The name of the file and other file attributes are not included in calculating the hash. A file by any name will hash the same.

[6] SCHNEIER, *supra* note 4, at 429. The range of values generated from commonly used hash algorithms is huge. For example, the prolific algorithm MD-5 can generate more than 340,000,000,000,000,000,000,000,000,000,000,000,000,000 (that's 340 billion, billion, billion, billion) possible values. The widely used SHA-1 algorithm generates a range of values over four billion times larger than that. Thus, although there is a finite number of possible hash values and an infinite number of possible data inputs, the odds of a collision are infinitesimally small.

monly used algorithms,[7] there is more than reasonable assurance that two different inputs will not have the same hash value.[8]

A second property is that the hash algorithm works in only one direction. One can calculate a hash value from input, but cannot derive the input from the hash value.[9] The hash value of the wedding photo cannot be "reversed" to generate the photo itself. Of course, by virtue of the first property, if an unknown file has a hash value identical to that of another known file, then you know that the first file is the same as the second.

B. Key Uses of Hashing Algorithms in the Digital Forensics Process

These properties make hash algorithms highly useful throughout the forensics process. Often one of the first steps of the process is to generate an image of the media to be analyzed.[10] Because hash values uniquely identify the underlying data, a technician who generates an image can use a hash algorithm to ensure that the copying was done accurately, to the bit. To do this, the technician will calculate the hash value of the entire original drive,[11] generate an image, and then calculate the hash value of the entire image.[12] If the two hash values match, the imaging process worked and the duplicate is an exact replica of the original; if they don't match, then the imaging can be attempted again. Similarly, if additional forensics images are created, the hash value generated initially becomes the touchstone for ensuring those copies are and remain duplicates of the original.[13]

During the analysis phase, analysts often must sort through massive amounts of data to gather the important information and disregard the unimportant. Hashing can automate much of this work. There are now libraries of hash values calculated for common programs and files.[14] Many digital analysis tools can be configured to cal-

---

[7] *See* Xiaoyun Wang, Yiqun Lissa Yin & Hongbo Yu, *Finding Collisions in the Full SHA-1*, http://www.infosec.sdu.edu.cn/paper/sha1-crypto-auth-new-2-yao.pdf (last visited Jan. 17, 2006); Xiaoyun Wang & Hongbo Yu, *How To Break MD5 and Other Hash Functions* (2005), http://www.infosec.sdu.edu.cn/paper/md5-attack.pdf.

[8] Research shows that, under controlled and artificial circumstances, it is possible to engineer two different files with the same hash value. Some effort has been made to design a tool that can create collisions. It is extremely unlikely that collisions would happen in the wild, much less in the context of digital media imaging and forensics.

[9] SCHNEIER, *supra* note 4, at 30.

[10] *See* MICHAEL G. SOLOMON, DIANE BARRETT & NEIL BROOM, COMPUTER FORENSICS JUMPSTART: COMPUTER FORENSICS BASICS 85 (2004).

[11] The process of calculating the hash value of data does not alter the data itself.

[12] *See* SOLOMON ET AL., *supra* note 10, at 83–84.

[13] *See id.* at 69.

[14] The National Software Reference Library (NSRL), run by the National Institute of Science and Technology, has a huge set of hash values of software commonly found on computers. In

culate separate hash values of each individual file on the imaged media, and match the values against known hash sets. Depending on what the examiner is looking for, the forensic tool may include the matching files in a report or exclude them. For example, an investigator having no interest in the word processing software on a seized hard drive can load the hash set for common word processing programs and exclude those files from analysis. In this respect, the use of hash values minimizes the amount of data that the investigator sees. Using Professor Kerr's terminology, the unwanted data is not "exposed" in the forensics work.

By the same token, hash sets can be used to look for particular programs or files on the media. For example, an examiner may need to look for signs of malicious code on the media (perhaps in anticipation of a suspect's claim that the computer was infected with a worm that gave somebody else control over it). The examiner can load hash sets for known worms, viruses, and Trojan horses to look for signs that corroborate or rebut the defense.[15] Likewise, an examiner may use hashing to look for steganography[16] software and similar programs for evidence the user employs data-hiding techniques.

Although examinations are not necessarily flawed without hashing, the tool provides a reliable and inexpensive means to address critical data acquisition and examination goals. It also presents an interesting and important series of Fourth Amendment issues. In the next Part, I look at the Fourth Amendment implications of hashing, particularly in light of Professor Kerr's "exposure" theory of search.

III. HASHING AND THE FOURTH AMENDMENT

A. Hashing for Integrity Validation

Professor Kerr proposes a rule whereby a Fourth Amendment search occurs when any data on a hard drive or information about that data, no matter how little, is exposed to human observation.[17] Could Professor Kerr's proposal mean that a search occurs no later

---

June 2005, the NSRL had over 10.5 million unique hash values using different algorithms covering over 31.7 million files. National Software Reference Library, http://www.nsrl.nist.gov/ (last visited Jan. 17, 2006).

[15] There are better ways to look for malware, but this is one technique. *See* Susan W. Brenner & Brian Carrier with Jef Henninger, *The Trojan Horse Defense in Cybercrime Cases*, 21 SANTA CLARA COMPUTER & HIGH TECH. L.J. 1, 36–37 (2004); *see also* United States v. Bass, 411 F.3d 1198, 1200 (10th Cir. 2005) (defendant claimed a virus was responsible for downloaded child pornography).

[16] Steganography refers to the process of hiding data within other data, such as hiding text in an image.

[17] Orin S. Kerr, *Searches and Seizures in a Digital World*, 119 HARV. L. REV. 531, 547–48 (2005).

than when the imaging technician views the hash values to verify that the process was completed successfully? Is exposure of the media's hash value a constitutional moment? After all, the hash value is derived from each bit on the drive and its placement relative to the other bits. There is essentially no chance that any other hard drive would have the same hash value. It is a value intimately associated with all the data on the media.

The Supreme Court decision in *Arizona v. Hicks*,[18] involving a search of an apartment, sheds an interesting light on this question. In *Hicks*, the Court concluded that there was a search for Fourth Amendment purposes when an officer, legitimately on the premises, moved a turntable to view a serial number on the bottom.[19] At first glance, the facts in *Hicks* may seem analogous to a hash calculation: the serial number serves to identify the turntable, just as the hash value serves to identify the data on the hard drive. Finding the serial number required that the agent manipulate the turntable, just as calculating a hash value requires accessing the data on the hard drive.

But those similarities are superficial. In *Hicks*, the item that the agent was searching for and found, the serial number, was already in existence and was located on the turntable. A hash value, on the other hand, is something that the agent will derive; it is not something the agent will look for stored on the hard drive.[20] Moreover, although the hash is generated from the data on the drive, it will normally reveal nothing about that information.[21] It does not suggest what the information is or lay bare any of its characteristics or attributes. Although a "search is a search, even if it happens to disclose nothing but the bottom of a turntable,"[22] hash calculations disclose nothing. As a predictor of data, the hash value is no more useful than a random number.

Though one may argue that there is a reasonable expectation of privacy in the hash value of the drive, and thus its exposure constitutes a search, the true degree of intrusion into private matters is, at most, de minimis. Generating the hash value of the image serves the purpose of allowing the analysis of a massive quantity of data to proceed with great confidence that the data were collected correctly and will not be tainted by the forensic process. A defendant hardly can be heard to claim a constitutional right to a lesser standard of evidence

---

[18] 480 U.S. 321 (1987).

[19] *Id.* at 324–25.

[20] Professor Kerr notes that some metadata stored on a drive could be revealed as part of the imaging, and suggests that to the extent these data are exposed, there is a "search." Kerr, *supra* note 17, at 560 n.130. Like the serial number in *Hicks*, and unlike the hash value, the metadata themselves are taken from the media.

[21] There may be some rare edge cases where this is not true. One such case is discussed *infra* section III.B.2.

[22] *Hicks*, 480 U.S. at 325.

handling.[23] Because hashing is "minimally intrusive" and is driven by "operational necessities," there is little constitutional significance.[24]

B. Hashing for Data Reduction and Data Exposure

1. Data Reduction. — Hashing can greatly speed up forensic analysis by acting as a sorting mechanism. When a technician uses hashing to exclude files, not only can the examination conclude more efficiently, but it is also less intrusive. Files that are known to be outside of the scope of the legitimate search (and for which hash values are available) can be carved from the examination altogether. Under Professor Kerr's "exposure" theory of search, so long as the list of excluded files found on the media was not exposed, the data would not have been "searched" in this data reduction process. This conclusion is also consistent with the recognition that it is permissible to sort through reams of data in order to find and search only the information within the scope of the warrant.[25] Indeed, because the process is largely automated, the privacy interest of the searched party is better protected than it would be if the sorting were done by human hands as it is done with paper archives.

2. Data Exposure. — A technician can also match the media hash list against a known-file hash list to find particular files. This could include looking for evidence that the computer had been compromised by malicious code, such as a worm, virus, or Trojan horse. Custom-made known-file hash lists provide nearly unlimited flexibility to look for unusual files that would not be found in a public hash set. A customized list could include hash values for files stolen in an intrusion, or pirated software not yet released generally. Under Professor Kerr's analysis, use of hashing in this manner would constitute a Fourth Amendment search, because the results expose certain contents of the media.[26]

The more interesting, and controversial, question is whether hash-based examination may be used to find "digital contraband" without a

---

[23] *Cf.* United States v. Jacobsen, 466 U.S. 109, 119 & n.16 ("Protecting the risk of misdescription hardly enhances any legitimate privacy interest, and is not protected by the Fourth Amendment.").

[24] *Cf. Hicks*, 480 U.S. at 327 ("We have held that [a seizure] can [be justified on less than probable cause] where, for example, the seizure is minimally intrusive and operational necessities render it the only practicable means of detecting certain types of crime.").

[25] *Cf.* United States v. Brooks, 427 F.3d 1246, 1250–52 (10th Cir. 2005).

[26] A related question arises if the hash list generated from the seized media contains entries for files outside the search authority and is viewed by the investigators. Arguably, to the extent such an entry is compared against a known-file hash list and a match is discovered, information is revealed from the hash. Under Professor Kerr's proposal, this too may constitute exposure and trigger Fourth Amendment requirements.

warrant or exception to the warrant requirement.[27] The question stems from three Supreme Court decisions, each of which involved narcotics detection: *United States v. Place*, *United States v. Jacobsen*, and *Illinois v. Caballes*. In *Caballes*, the Court held that the use of a "well-trained narcotics-detection dog — one that 'does not expose non-contraband items that otherwise would remain hidden from public view[]' — during a lawful traffic stop, generally does not implicate legitimate privacy interests."[28] (*Place* similarly involved use of a narcotics-sniffing dog at an airport.[29]) In *Jacobsen*, the Court held that a field chemical test "that merely discloses whether or not a particular substance is cocaine does not compromise any legitimate interest in privacy."[30]

All three decisions are premised on the doctrine that unless government conduct compromises a "legitimate" interest in privacy, there is no search subject to the Fourth Amendment warrant requirement.[31] Because there is no legitimate interest in possessing something that is illegal to possess (for example, contraband), there is similarly no search when an investigator uses a tool that reveals only contraband.[32] In concluding that there was no "search," the key focus, particularly in *Caballes*, was on two main questions: (1) is the item to be detected illegal to possess, and (2) will the detection tool reveal only that item?[33]

The decisions could have significant implications for digital media forensics, particularly in light of powerful hash-based examinations. It is clear that digital child pornography constitutes contraband in the *Caballes*-sense that it is illegal to possess.[34] Further, hash-based examination provides a very effective and efficient means to identify only particular files. Does this suggest that law enforcement could, as a routine practice in all forensics labs when searching digital media, also conduct a hash search for child pornography without regard to the confines of the search authority?

Could every examination of every electronic storage device include

---

[27] *See, e.g.*, Ric Simmons, *The Two Unanswered Questions of* Illinois v. Caballes*: How To Make the World Safe for Binary Searches*, 80 TUL. L. REV. (forthcoming 2006); Michael Adler, Note, *Cyberspace, General Searches, and Digital Contraband: The Fourth Amendment and the Net-Wide Search*, 105 YALE L.J. 1093 (1996).

[28] *Illinois v. Caballes*, 125 S. Ct. 834, 838 (2005) (citation omitted) (quoting United States v. Place, 462 U.S. 696, 707 (1983)).

[29] *Place*, 462 U.S. at 707.

[30] *United States v. Jacobsen*, 466 U.S. 109, 123 (1984).

[31] *Caballes*, 125 S. Ct. at 837.

[32] *Id.*; *Jacobsen*, 466 U.S. at 124 n.24.

[33] *See Caballes*, 125 S. Ct. at 838; *Jacobsen*, 466 U.S. at 123 ("Congress has decided — and there is no question about its power to do so — to treat the interest in 'privately' possessing cocaine as illegitimate; thus governmental conduct that can reveal whether a substance is cocaine, and no other arguably 'private' fact, compromises no legitimate privacy interest.")

[34] *See Caballes*, 125 S. Ct. at 839.

a hash-based search for child pornography, even in, for example, a tax evasion case where there is no predicate to believe child pornography is present on the media? In 2004, the Regional Computer Forensics Laboratories, run largely by the FBI, conducted over 1300 examinations.[35] No doubt that many were accompanied by a warrant or valid consent allowing a search for evidence of child pornography, but many involved other investigations such as homicides, terrorism, and health-care fraud.[36] Certainly we benefit from an aggressive battle against the scourge of child pornography. Yet there would be something very creepy about an expansive and unrestrained search through media, even though properly in the hands of law enforcement, for offending images.

Aside from this visceral reaction, however, *Caballes* and its line do present some opportunity for hash-based examinations for contraband. But how good a fit is it?

Arguably, the hash tool is less like the narcotics-detecting dog of *Caballes*, and more like the heat imaging device used to detect marijuana in a home in *Kyllo v. United States*.[37] After all, in many cases the computer will have been used in and seized from a personal residence. But that is about where the similarity ends. Significantly, in distinguishing *Caballes* from *Kyllo*, the Court did not rely on the fact that *Caballes* involved a traffic stop and *Kyllo* a home.[38] Rather, the Court held that unlike a dog sniff, the heat imaging tool of *Kyllo* was not discerning enough to reveal only illicit activities.[39] Hash-based file detection works extraordinarily well in identifying only those files that are exact matches, and is far more akin to *Caballes* in this material feature than *Kyllo*.

It would seem that *Caballes* would allow for the routine use by government of hash-based contraband detection in any search of a digital storage device, regardless of the scope of the search authority. But it may not be as simple as that. For hash-based examinations to work, the underlying known-file hash lists must contain hash values for images that are illegal to possess. It is one thing to conclude that

---

[35] FBI REG'L COMPUTER FORENSICS LAB. PROGRAM, FISCAL YEAR 2004 ANNUAL REPORT 6, *available at* http://www.rcfl.gov/downloads/documents/RCFL_Nat_Annual04.pdf.

[36] *Id.* at 8.

[37] 533 U.S. 27 (2001).

[38] Nothing in the analysis of *Caballes* suggested that it was location-dependent, or that location separated *Caballes* from *Kyllo*. That is not to say there is no room left to argue that the degree of intrusiveness is greater when the drug dog sniffs outside a home or apartment. *Cf.* United States v. Thomas, 757 F.2d 1359, 1366 (2d Cir. 1985) (use of dog to sniff outside suspect's apartment constituted a search). In most digital forensics examinations, however, the hash-matching process will be conducted entirely off-site, with no additional intrusion of the home or meaningful delay in the examination.

[39] *Caballes*, 125 S. Ct. at 838.

child pornography is contraband; it is quite another to conclude that a particular image to be included in a hash set is child pornography.

The definition of child pornography cannot be set out as a chemical formula, unlike drug contraband, and no legislative body has declared particular images to be contraband, much less blessed a hash set. Instead, the definitions describe the attributes that make an image contraband.[40] It would seem that populating a hash set requires exercise of discretion that is not required when teaching a dog to detect cocaine or developing a chemical test to react to particular narcotics.

There are, however, known series of confirmed child pornography. These are images in which the person depicted has been identified as underage, and perhaps has even provided sworn statements and other evidence to that effect. Some of the images have been the subject of adjudication.[41] There is little doubt that these images are child pornography. Is it enough under *Caballes* that the facts as developed in judicial proceedings so strongly support the conclusion these images fit within the statutory definition of child pornography, or must that decision be made by the legislature?

At least theoretically, *Caballes* and its line may very well provide a basis for using hash-based tools to find child pornography on any digital media properly in the control of law enforcement. The technology for running hash-based examinations is well established in the most common of forensics tools. The more difficult problem is creating a hash set that contains only images that are illegal to possess. Without that, and in spite of its phenomenal ability to reveal only files it is asked to reveal, the tool becomes more like the *Kyllo* device; it exposes that to which there is a legitimate privacy interest.

CONCLUSION

The hash algorithm has afforded digital media forensic analysis a highly reliable and efficient means to ensure that the integrity of the digital evidence collected remains uncompromised. It also provides a means to discard from the examination the irrelevant, and focus in on the important, while exposing little, if any, ancillary information. The interests of law enforcement are served, as are the legitimate privacy interests of the subject. The use of hashing thus does a laudable job at advancing the purposes of the Fourth Amendment. As collections of known hash values grow and are incorporated into widely used forensics tools, however, the potential for hashing to intrude into private areas also grows.

---

[40] *See* 18 U.S.C.A. § 2256(8) (West 2000 & Supp. 2005) (defining "child pornography").

[41] *See, e.g.*, United States v. Venson, 82 F. App'x 330 (5th Cir. 2003) (prosecution for possessing and receiving child pornography).