Jessica Fridrich (Ed.)

LNCS 3200

Information Hiding

6th International Workshop, IH 2004
Toronto, Canada, May 2004
Revised Selected Papers

Springer

Jessica Fridrich (Ed.)

Information Hiding

6th International Workshop, IH 2004
Toronto, Canada, May 23-25, 2004
Revised Selected Papers

Springer

Volume Editor

Jessica Fridrich
SUNY Binghamton, Department of Electrical and Computer Engineering
Binghamton, NY 13902-6000, USA
E-mail: fridrich@binghamton.edu

Library of Congress Control Number: Applied for

CR Subject Classification (1998): E.3, K.6.5, K.4.1, K.5.1, D.4.6, E.4, C.2, H.4.3, H.3, H.5.1

ISSN 0302-9743
ISBN 3-540-24207-4 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 11371847 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science 3200

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Preface

It is an honor and great pleasure to write a preface for this postproceedings of the 6th International Workshop on Information Hiding. In the past 10 years, the field of data hiding has been maturing and expanding, gradually establishing its place as an active interdisciplinary research area uniquely combining information theory, cryptology, and signal processing.

This year, the workshop was followed by the Privacy Enhancing Technologies workshop (PET) hosted at the same location. Delegates viewed this connection as fruitful as it gave both communities a convenient opportunity to interact.

We would like to thank all authors who submitted their work for consideration. Out of the 70 submisions received by the program committee, 25 papers were accepted for publication based on their novelty, originality, and scientific merit. We strived to achieve a balanced exposition of papers that would represent many different aspects of information hiding. All papers were divided into eight sessions: digital media watermarking, steganalysis, digital forensics, steganography, software watermarking, security and privacy, anonymity, and data hiding in unusual content. This year, the workshop included a one-hour rump session that offered an opportunity to the delegates to share their work in progress and other brief but interesting contributions.

The program committee consisted of Ross J. Anderson (University of Cambridge, UK), Jan Camenisch (IBM Zurich Research Laboratory, Switzerland), Christian Collberg (University of Arizona, USA), Ingemar J. Cox (University College London, UK), John McHugh (SEI/CERT, USA), Ira S. Moskowitz (Naval Research Laboratory, USA), Job Oostveen (Philips Research, Netherlands), Richard C. Owens (University of Toronto), Fabien A.P. Petitcolas (Microsoft Research, UK), Andreas Pfitzmann (Dresden University of Technology, Germany), Mike Reiter (Carnegie Mellon University, USA), and Jessica Fridrich (SUNY Binghamton, USA).

The following external reviewers participated in the review process: Richard Clayton (University of Cambridge, UK), Farid Ahmed (The Catholic University of America, USA), Dogan Kesdogan (Aachen University of Technology, Germany), Hany Farid (Dartmouth College, USA), Deepa Kundur (Texas A&M University, USA), Slava Voloshinovsky (CUI, University of Geneva, Switzerland), Fernando Perez-Gonzales (University of Vigo, Spain), Nasir Memon (Polytechnic University, USA), Scott Craver (Princeton University, USA), Li Wu Chang (Naval Research Laboratory, USA), Lisa Marvel (University of Delaware, USA), Frederic Deguillaume (CUI, University of Geneva, Switzerland), Andrei Serjantov (University of Cambridge, UK), Rainer Böhme (Dresden University of Technology, Germany), Andreas Westfeld (Dresden University of Technology, Germany), George Danezis (University of Cambridge, UK), Sandra Steinbrecher (Dresden University of Technology, Germany), Phil Sallee (Booz Allen Hamilton, USA), Richard E. Newman (University of Florida, USA), Paul Syverson (Naval Research Laboratory, USA), John McDermott (Naval Research Laboratory, USA), Dagmar Schönfeld (Dresden

University of Technology, Germany), Tim McChesney (Naval Research Laboratory, USA), Karen Spärck Jones (University of Cambridge, UK), Sebastian Clauß (Dresden University of Technology, Germany), Sorina Dumitrescu (McMaster University, Canada), Elke Franz (Dresden University of Technology, Germany), Edward Carter (University of Arizona, USA), Andrew Huntwork (University of Arizona, USA), Saumya Debray (University of Arizona, USA), Kelly Heffner (University of Arizona, USA), Ginger Myles (University of Arizona, USA), Clark Thomborson (University of Auckland, New Zealand), Jasvir Nagra (University of Auckland, New Zealand), Viktor Raskin (Purdue University, USA), Nicholas Hopper (Carnegie Mellon University, USA), Aweke Lemma (Philips Digital Systems Laboratories, The Netherlands), Gerhard Langelaar (Philips Digital Systems Laboratories, The Netherlands), Frans Willems (Technical University of Eindhoven, The Netherlands), Fons Bruekers (Philips Research, The Netherlands), Arno van Leest (Philips Research, The Netherlands), Michiel van der Veen (Philips Research, The Netherlands), and Ton Kalker (Hewlett-Packard, USA).

This year, for the first time this workshop had two program chairs, one for multimedia watermarking and steganography (myself) and the second for anonymous communication, covert channels, and privacy (Mike Reiter). I would like to thank Mike for helping me with the review process and managing the communication with authors.

The general chair Richard C. Owens and his assistant Alison Bambury did a wonderful job organizing the event. Many thanks to them for such a tasteful selection of a comfortable meeting place. The workshop was held at The Radisson located on the Ontario Waterfront. In the evening of the second day, the attendees had an opportunity to relax at a dinner cruise while admiring the Ontario city silhouette lit by fireworks for Victoria Day.

Special thanks belong to Tim Olson from Microsoft Conference Management Services. The submission of papers and reviews as well as notification of authors and reviewers was greatly simplified both for the authors and program committee members.

Finally, I would like to thank The Information and Privacy Commissioner/Ontario, The Centre for Innovation Law Policy, and Bell University Laboratories for their sponsorship of this workshop.

September 2004

Jessica Fridrich
SUNY Binghamton
New York, USA

Lecture Notes in Computer Science

For information about Vols. 1–3247

please contact your bookseller or Springer

Vol. 3358: J. Cao, L.T. Yang, M. Guo, F. Lau (Eds.), Parallel and Distributed Processing and Applications. XXIV, 1058 pages. 2004.

Vol. 3356: G. Das, V.P. Gulati (Eds.), Intelligent Information Technology. XII, 428 pages. 2004.

Vol. 3348: A. Canteaut, K. Viswanathan (Eds.), Progress in Cryptology - INDOCRYPT 2004. XIV, 431 pages. 2004.

Vol. 3347: R.K. Ghosh, H. Mohanty (Eds.), Distributed Computing and Internet Technology. XX, 472 pages. 2004.

Vol. 3341: R. Fleischer, G. Trippen (Eds.), Algorithms and Computation. XVII, 935 pages. 2004.

Vol. 3340: C.S. Calude, E. Calude, M.J. Dinneen (Eds.), Developments in Language Theory. XI, 431 pages. 2004.

Vol. 3339: G.I. Webb, X. Yu (Eds.), AI 2004: Advances in Artificial Intelligence. XXII, 1272 pages. 2004. (Subseries LNAI).

Vol. 3338: S.Z. Li, J. Lai, T. Tan, G. Feng, Y. Wang (Eds.), Advances in Biometric Person Authentication. XVIII, 699 pages. 2004.

Vol. 3337: J.M. Barreiro, F. Martin-Sanchez, V. Maojo, F. Sanz (Eds.), Biological and Medical Data Analysis. XI, 508 pages. 2004.

Vol. 3336: D. Karagiannis, U. Reimer (Eds.), Practical Aspects of Knowledge Management. X, 523 pages. 2004. (Subseries LNAI).

Vol. 3334: Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, E.-p. Lim (Eds.), Digital Libraries: International Collaboration and Cross-Fertilization. XX, 690 pages. 2004.

Vol. 3333: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004. XXXV, 785 pages. 2004.

Vol. 3332: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004. XXXVI, 1051 pages. 2004.

Vol. 3331: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004. XXXVI, 667 pages. 2004.

Vol. 3329: P.J. Lee (Ed.), Advances in Cryptology - ASIACRYPT 2004. XVI, 546 pages. 2004.

Vol. 3328: K. Lodaya, M. Mahajan (Eds.), FSTTCS 2004: Foundations of Software Technology and Theoretical Computer Science. XVI, 532 pages. 2004.

Vol. 3326: A. Sen, N. Das, S.K. Das, B.P. Sinha (Eds.), Distributed Computing - IWDC 2004. XIX, 546 pages. 2004.

Vol. 3323: G. Antoniou, H. Boley (Eds.), Rules and Rule Markup Languages for the Semantic Web. X, 215 pages. 2004.

Vol. 3322: R. Klette, J. Žunić (Eds.), Combinatorial Image Analysis. XII, 760 pages. 2004.

Vol. 3321: M.J. Maher (Ed.), Advances in Computer Science - ASIAN 2004. XII, 510 pages. 2004.

Vol. 3320: K.-M. Liew, H. Shen, S. See, W. Cai (Eds.), Parallel and Distributed Computing: Applications and Technologies. XXIV, 891 pages. 2004.

Vol. 3316: N.R. Pal, N.K. Kasabov, R.K. Mudi, S. Pal, S.K. Parui (Eds.), Neural Information Processing. XXX, 1368 pages. 2004.

Vol. 3315: C. Lemaître, C.A. Reyes, J.A. González (Eds.), Advances in Artificial Intelligence – IBERAMIA 2004. XX, 987 pages. 2004. (Subseries LNAI).

Vol. 3314: J. Zhang, J.-H. He, Y. Fu (Eds.), Computational and Information Science. XXIV, 1259 pages. 2004.

Vol. 3312: A.J. Hu, A.K. Martin (Eds.), Formal Methods in Computer-Aided Design. XI, 445 pages. 2004.

Vol. 3311: V. Roca, F. Rousseau (Eds.), Interactive Multimedia and Next Generation Networks. XIII, 287 pages. 2004.

Vol. 3309: C.-H. Chi, K.-Y. Lam (Eds.), Content Computing. XII, 510 pages. 2004.

Vol. 3308: J. Davies, W. Schulte, M. Barnett (Eds.), Formal Methods and Software Engineering. XIII, 500 pages. 2004.

Vol. 3307: C. Bussler, S.-k. Hong, W. Jun, R. Kaschek, D.. Kinshuk, S. Krishnaswamy, S.W. Loke, D. Oberle, D. Richards, A. Sharma, Y. Sure, B. Thalheim (Eds.), Web Information Systems – WISE 2004 Workshops. XV, 277 pages. 2004.

Vol. 3306: X. Zhou, S. Su, M.P. Papazoglou, M.E. Orlowska, K.G. Jeffery (Eds.), Web Information Systems – WISE 2004. XVII, 745 pages. 2004.

Vol. 3305: P.M.A. Sloot, B. Chopard, A.G. Hoekstra (Eds.), Cellular Automata. XV, 883 pages. 2004.

Vol. 3303: J.A. López, E. Benfenati, W. Dubitzky (Eds.), Knowledge Exploration in Life Science Informatics. X, 249 pages. 2004. (Subseries LNAI).

Vol. 3302: W.-N. Chin (Ed.), Programming Languages and Systems. XIII, 453 pages. 2004.

Vol. 3299: F. Wang (Ed.), Automated Technology for Verification and Analysis. XII, 506 pages. 2004.

Vol. 3298: S.A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), The Semantic Web – ISWC 2004. XXI, 841 pages. 2004.

Vol. 3296: L. Bougé, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2004. XXV, 530 pages. 2004.

Vol. 3295: P. Markopoulos, B. Eggen, E. Aarts, J.L. Crowley (Eds.), Ambient Intelligence. XIII, 388 pages. 2004.

Vol. 3294: C.N. Dean, R.T. Boute (Eds.), Teaching Formal Methods. X, 249 pages. 2004.

Vol. 3293: C.-H. Chi, M. van Steen, C. Wills (Eds.), Web Content Caching and Distribution. IX, 283 pages. 2004.

Vol. 3292: R. Meersman, Z. Tari, A. Corsaro (Eds.), On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops. XXIII, 885 pages. 2004.

Vol. 3291: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE. XXV, 824 pages. 2004.

Vol. 3290: R. Meersman, Z. Tari (Eds.), On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE. XXV, 823 pages. 2004.

Vol. 3289: S. Wang, K. Tanaka, S. Zhou, T.W. Ling, J. Guan, D. Yang, F. Grandi, E. Mangina, I.-Y. Song, H.C. Mayr (Eds.), Conceptual Modeling for Advanced Application Domains. XXII, 692 pages. 2004.

Vol. 3288: P. Atzeni, W. Chu, H. Lu, S. Zhou, T.W. Ling (Eds.), Conceptual Modeling – ER 2004. XXI, 869 pages. 2004.

Vol. 3287: A. Sanfeliu, J.F. Martínez Trinidad, J.A. Carrasco Ochoa (Eds.), Progress in Pattern Recognition, Image Analysis and Applications. XVII, 703 pages. 2004.

Vol. 3286: G. Karsai, E. Visser (Eds.), Generative Programming and Component Engineering. XIII, 491 pages. 2004.

Vol. 3285: S. Manandhar, J. Austin, U.B. Desai, Y. Oyanagi, A. Talukder (Eds.), Applied Computing. XII, 334 pages. 2004.

Vol. 3284: A. Karmouch, L. Korba, E.R.M. Madeira (Eds.), Mobility Aware Technologies and Applications. XII, 382 pages. 2004.

Vol. 3283: F.A. Aagesen, C. Anutariya, V. Wuwongse (Eds.), Intelligence in Communication Systems. XIII, 327 pages. 2004.

Vol. 3282: V. Guruswami, List Decoding of Error-Correcting Codes. XIX, 350 pages. 2004.

Vol. 3281: T. Dingsøyr (Ed.), Software Process Improvement. X, 207 pages. 2004.

Vol. 3280: C. Aykanat, T. Dayar, İ. Körpeoğlu (Eds.), Computer and Information Sciences - ISCIS 2004. XVIII, 1009 pages. 2004.

Vol. 3278: A. Sahai, F. Wu (Eds.), Utility Computing. XI, 272 pages. 2004.

Vol. 3275: P. Perner (Ed.), Advances in Data Mining. VIII, 173 pages. 2004. (Subseries LNAI).

Vol. 3274: R. Guerraoui (Ed.), Distributed Computing. XIII, 465 pages. 2004.

Vol. 3273: T. Baar, A. Strohmeier, A. Moreira, S.J. Mellor (Eds.), <<UML>> 2004 - The Unified Modelling Language. XIII, 454 pages. 2004.

Vol. 3271: J. Vicente, D. Hutchison (Eds.), Management of Multimedia Networks and Services. XIII, 335 pages. 2004.

Vol. 3270: M. Jeckle, R. Kowalczyk, P. Braun (Eds.), Grid Services Engineering and Management. X, 165 pages. 2004.

Vol. 3269: J. Lopez, S. Qing, E. Okamoto (Eds.), Information and Communications Security. XI, 564 pages. 2004.

Vol. 3268: W. Lindner, M. Mesiti, C. Türker, Y. Tzitzikas, A. Vakali (Eds.), Current Trends in Database Technology - EDBT 2004 Workshops. XVIII, 608 pages. 2004.

Vol. 3267: C. Priami, P. Quaglia (Eds.), Global Computing. VIII, 377 pages. 2004.

Vol. 3266: J. Solé-Pareta, M. Smirnov, P.V. Mieghem, J. Domingo-Pascual, E. Monteiro, P. Reichl, B. Stiller, R.J. Gibbens (Eds.), Quality of Service in the Emerging Networking Panorama. XVI, 390 pages. 2004.

Vol. 3265: R.E. Frederking, K.B. Taylor (Eds.), Machine Translation: From Real Users to Research. XI, 392 pages. 2004. (Subseries LNAI).

Vol. 3264: G. Paliouras, Y. Sakakibara (Eds.), Grammatical Inference: Algorithms and Applications. XI, 291 pages. 2004. (Subseries LNAI).

Vol. 3263: M. Weske, P. Liggesmeyer (Eds.), Object-Oriented and Internet-Based Technologies. XII, 239 pages. 2004.

Vol. 3262: M.M. Freire, P. Chemouil, P. Lorenz, A. Gravey (Eds.), Universal Multiservice Networks. XIII, 556 pages. 2004.

Vol. 3261: T. Yakhno (Ed.), Advances in Information Systems. XIV, 617 pages. 2004.

Vol. 3260: I.G.M.M. Niemegeers, S.H. de Groot (Eds.), Personal Wireless Communications. XIV, 478 pages. 2004.

Vol. 3259: J. Dix, J. Leite (Eds.), Computational Logic in Multi-Agent Systems. XII, 251 pages. 2004. (Subseries LNAI).

Vol. 3258: M. Wallace (Ed.), Principles and Practice of Constraint Programming – CP 2004. XVII, 822 pages. 2004.

Vol. 3257: E. Motta, N.R. Shadbolt, A. Stutt, N. Gibbins (Eds.), Engineering Knowledge in the Age of the Semantic Web. XVII, 517 pages. 2004. (Subseries LNAI).

Vol. 3256: H. Ehrig, G. Engels, F. Parisi-Presicce, G. Rozenberg (Eds.), Graph Transformations. XII, 451 pages. 2004.

Vol. 3255: A. Benczúr, J. Demetrovics, G. Gottlob (Eds.), Advances in Databases and Information Systems. XI, 423 pages. 2004.

Vol. 3254: E. Macii, V. Paliouras, O. Koufopavlou (Eds.), Integrated Circuit and System Design. XVI, 910 pages. 2004.

Vol. 3253: Y. Lakhnech, S. Yovine (Eds.), Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems. X, 397 pages. 2004.

Vol. 3252: H. Jin, Y. Pan, N. Xiao, J. Sun (Eds.), Grid and Cooperative Computing - GCC 2004 Workshops. XVIII, 785 pages. 2004.

Vol. 3251: H. Jin, Y. Pan, N. Xiao, J. Sun (Eds.), Grid and Cooperative Computing - GCC 2004. XXII, 1025 pages. 2004.

Vol. 3250: L.-J. (LJ) Zhang, M. Jeckle (Eds.), Web Services. X, 301 pages. 2004.

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004. (Subseries LNAI).

Table of Contents

Session 7 - Anonymity
Session Chair: Andreas Pfitzmann (Dresden University of Technology)

Session 8 - Data Hiding in Unusual Content
Session Chair: Christian Collberg (University of Arizona)

An Implementation of, and Attacks on, Zero-Knowledge Watermarking

Scott Craver, Bede Liu, and Wayne Wolf

Department of Electrical Engineering
Princeton University

Abstract. A problem of considerable theoretical interest in digital watermarking is that of asymmetric, or *zero-knowledge* watermarking. In this problem, we wish to embed a watermark in a piece of multimedia and later prove that we have done so, but without revealing information that can be used by an adversary to remove the signal later.
In this paper we develop a watermarking system based on the ambiguity attack method outlined in [14], constructing a vector-based watermarking system applicable to images, audio and video. An example of image watermarking is provided. We also outline some important attacks and thus important design principles for asymmetric watermarking systems.

1 Introduction

A problem of considerable interest in recent years is the *asymmetric watermarking problem:* can we prove that a watermark has been embedded in a piece of multimedia without providing the information that would allow its removal?

Several methods to solve this problem have been proposed, some of which have since been broken, and some of which possess properties we would like to remove, such as the need for a trusted third party. Still other ideas are embryonic, such as the watermarking approach based on ambiguity attacks outlined but not implemented in [14]. An overview of various asymmetric watermarking techniques can be found in [3].

In this paper we develop a watermarking system based on the ambiguity attack method, mechanizing the process of constructing counterfeit watermarks for a blind detector. Real and counterfeit watermarks are then used in a zero-knowledge proof to show that at least one of a set of watermarks is valid. This yields a general algorithm for vector-based watermarking applicable to images, audio and video multimedia.

We also discuss some of the design principles we have encountered in the development of this system. In analyzing zero-knowledge watermarking systems, we find that some are vulnerable to ambiguity attacks, which we continue to stress are a serious problem in asymmetric watermarking systems and not to be ignored. This problem highlights general design philosophies regarding what a watermark is meant to mean, and what we are attempting to prove when passing data through a detector.

J. Fridrich (Ed.): IH 2004, LNCS 3200, pp. 1–12, 2004.

2 The Still Very Serious Problem of Ambiguity Attacks

Ambiguity attacks, once an easily preventable curiosity, become a critical problem in zero-knowledge watermarking, perhaps partially because they are mistakenly regarded as trivial application issues [5,2,6]. However, if one does not carefully design a watermarking scheme to rule out these attacks, they may never be preventable. We provide several examples of such attacks on existing systems.

2.1 The Basic Attack

The basic attack is very simple: find an arbitrary signal that sets off the watermark detector. Then, claim that this signal is a watermark. This is often very easy, and it can be performed by inspection of both the detector structure and the multimedia to be attacked. For example, for a correlator detector, we can construct a counterfeit signal consisting of the multimedia itself, attenuated and perhaps processed to disguise it. This signal will naturally correlate with the original signal.

In symmetric systems, this is prevented by requiring a watermark to be the output of a secure hash $w = h(\text{seed})$. Now, an arbitrary signal can not easily be found for which a seed can be presented; and so the seed is evidence that a watermark was legitimately added, rather than found in place. This is a well known, simple and effective remedy, perhaps enough to relegate this attack to the domain of "implementation issues."

The problem now is that in an asymmetric system, a watermark owner cannot simply produce the seed or the watermark as proof of its authenticity. The parameters of the new problem can disallow the usual remedy, allowing this simple but serious vulnerability to appear.

2.2 An Example

A straightforward example is proposed in [7], in which a wartermark is embedded in data and a randomly selected subset of coefficients is revealed. For security, this subset is immersed in a random vector with each coefficient of a public watermark being either a coefficient of the secret vector or a random value. This signal is detectable as long as the embedding is sufficiently strong.

How do we prevent someone from presenting as a watermark any vector that correlates with the data, or encodes an arbitrary message? Proving that a randomly selected subset, immersed within a random vector, is drawn from a legal watermark is indeed difficult. To be fair, however, the authors do not propose this scheme specifically for proof of ownership applications, in which ambiguity attacks are a problem. In other so-called "digital rights management" applications, these attacks less important.

2.3 Another Example

A watermarking system proposed in [15] outlines a method of secure, blinded correlation. In this system, multimedia data and watermarks are represented as vectors, and detection consists of correlation followed by thresholding—all of which the authors are able to perform in a blinded domain. Given a commitment of an image vector I and a commitment of a watermark vector w, one can compute their dot-product and compare this result to a threshold without revealing w.

Thus, one is able to prove in zero knowledge (and this the authors establish rigorously) that a watermark signal w sets off a detector $D(w, I)$. The authors provide both a blind and non-blind case, although we will focus on the less ambiguous case of blind watermarking.

The simple vulnerability of this protocol, by itself, is that anyone can find a signal w for which $D(w, I) = 1$. The attacker has access to the signal itself and $D(I, I) = 1$ for a correlator detector. Likewise, the attacker can use all sorts of signals derived from I. Under the blinding conditions of the protocol, there is no way of determining if this trivial form of cheating is taking place, so anyone can prove the presense of a watermark in anything.

Knowledge of such a vector w is therefore not valuable information and does not need to be proven by any protocol. By analogy, imagine a zero-knowledge proof that one knows a factor of an integer n. Anyone can pass this test because everybody knows a factor of n. What is valuable is a proof that one knows a *nontrivial* factor or a legally constructed watermark.

Thus, the basic ambiguity attack cannot be prevented by the standard remedy. The authors in [15] propose a trusted third party to prevent ambiguity; images are registered with an authority who performs the watermarking, computes the blinded versions of mark and image, and provides a verifiable certificate to the image owner. Images are not allowed to be registered if they are "similar" to one previously registered. Given a third party of such capabilities, however, do we need asymmetric watermarking at all? The trusted third party can simply perform the detection itself.

2.4 Discussion

The important problem with asymmetric watermarking is not that these attacks are possible; but that *whenever they are possible, they are difficult to prevent.* In our experience analyzing and desigining asymmetric watermarking systems, we find that a successful scheme should be designed from the start with resistance to ambiguity. Only by luck can we expect an ambiguity-related flaw to be patchable in implementation.

We also observe that this class of vulnerabilities highlights an important semantic issue regarding watermarking in general: we are not trying to prove that a watermark signal is *detectable;* we are trying to prove that a signal has been *embedded.* These are very different notions, and ambiguity attacks generally disguise examples of the former as examples of the latter.

This semantic requirement for watermarking can in turn be considered a special case of the *nontriviality* requirement of a zero-knowledge proof: to be useful, it should demonstrate that one possesses *valuable* knowledge, knowledge not available to everybody—or in more concrete terms, there should be people who can not pass the protocol, making the ability to pass the protocol valuable in some way[1]. This nontriviality requirement is not actually part of the definition of a zero-knowledge proof as defined in textbooks on the subject [10,11,8].

3 An Implementation of Public-Key Watermarking Using Ambiguity Attacks

As described in [14], we can use ambiguity attacks constructively, as components in a watermarking system. The idea is simple: if there is no way for an adversary to distinguish a valid watermark from an invalid one, as is commonly the case, we can conceal a real watermark in a collection of false ones using zero-knowledge protocols to demonstrate that at least one is real.

The fake watermarks are components of the original multimedia signal. They are not added, but already reside "within" the multimedia, in the sense that they set off a watermark detector. In a sense, we decompose the multimedia signal into a sum of false watermarks plus residual data; thus removing a large number of false watermarks is the same as removing a significant part of the image or audio clip.

Note that we use zero-knowledge protocols not to show that a watermark is detectable; the detection process has no asymmetric properties. Rather, we focus our zero-knowledge efforts on verification, showing in zero knowledge that at least one of a collection of watermarks is legal.

If we embed M_r watermarks and find M_f counterfeit watermarks, and an attacker can selectively damage K watermarks, the probability of a successful attack is

$$P_{\text{attack}} = \binom{M_f}{K - M_r} \Big/ \binom{M_f + M_r}{K}$$

... a value roughly equal to p^{M_r}, where p is the fraction $K/(M_r + M_f)$ of watermarks the attacker can damage. We expect the collection of real watermarks to fall within a power budget, so that increasing M_r is not a matter of adding more power, but dividing that same power among more watermarks. Thus, attack probability drops exponentially with M_r, although we will see a penalty in detector probability unless we make the watermarks longer.

Note that by adding multiple marks rather than one, we avoid the *oracle attack* of removing each watermark one at a time, in M separate challenges, until the protocol cannot succeed. We also note that this attack probability will in practice be small but macroscopic: a probability of 2^{-64}, for instance,

[1] Or the dual *efficiency* requirement: a zero-knowledge proof of something everyone knows should take zero steps.

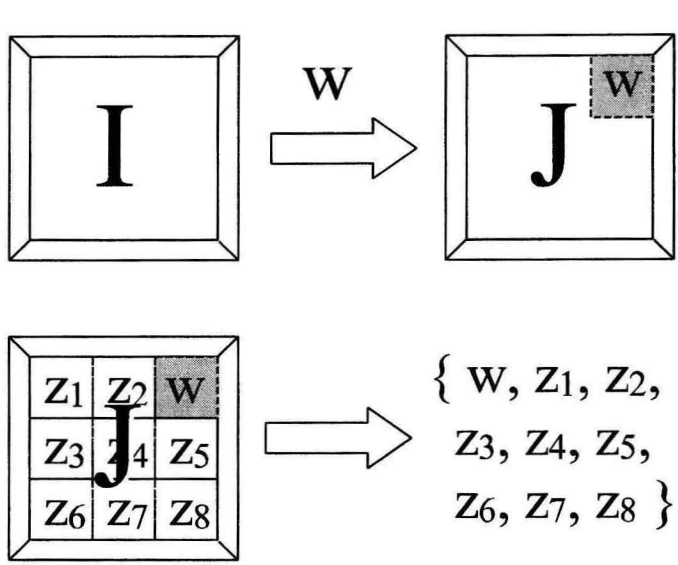

Fig. 1. Basic watermarking using invertibility attacks. Left, the original image I is watermarked with w, then a sufficiently large number of fake watermarks are found by inspection. Then, the set of watermarks is published, without disclosing which watermark is legitimate.

is unrealistic. However, this is the probability of the attacker defeating a legal challenge by the watermark owner, a time-consuming process not amenable to brute-force. If P_{attack} is as large as $1/10000$, an attacker need "only" lose 5000 lawsuits before defeating the watermarking system.

3.1 Implentation: The Gaussian Case

Consider first that we have a data signal to watermark which is a vector s of Gaussian random variables, independent and identically distributed $\sim N(0, \sigma_s^2)$. For practical purposes, we construct a signal vector s from an image or audio clip by extracting and processing a collection of features, usually in the frequency domain. Ideally, the extracted features are both well-behaved and significant, in the sense that damaging the features will greatly damage the multimedia.

In any case, we wish to generate M_r real watermark vectors $\{W_k\}$, and also decompose our N-dimensional vector s into into M_f false watermarks $\{Z_k\}$. To accomplish this, we require that a watermark also be a vector of iid Gaussian random variables $\sim N(0, \sigma_w^2)$, $\sigma_w < \sigma_s$.

Since a sum of Gaussian vectors is itself a Gaussian vector, we can easily generate a decomposition of s into $\{Z_k\}$ as follows:

Algorithm 1 *A decomposition of s into false watermarks*

1. *Compute K random watermarks $Y_1 \cdots Y_k$, such that $Y = \sum_i Y_i$ is approximately the same length as s.*
2. *Generate an arbitrary (random) orthonormal matrix A such that $AY = (1+\lambda)s$.*
3. *Set $Z_i = AY_i$, and $M_f = K$.*

Since these vectors will be very large, of dimension $N \sim 100000$, we need an efficient method of generating and applying an orthonormal transformation. Fortunately, there is indeed a linear-time algorithm for both generating such an A, and computing Ax for a vector x, without constructing huge matrices. This takes advantage of the fact that all coefficients of s and $\sum Y_i$ are nonzero with high probability, facilitating Gaussian elimination.

3.2 Error Probabilities

Note that the number of false watermarks will be approximately $M_f = \sigma_s^2/\sigma_w^2$, and so our attack probability is wholly dependent upon the chosen watermark strength. If Alice has a power budget σ_A^2, then she can choose a value for σ_w^2 to make M_f large, and then add $M_r = \sigma_A^2/\sigma_w^2$ watermarks.

Meanwhile, consider the detector false alarm and miss probabilities. For a watermark detected by correlation, we have for the maximum-likelihood detector,

$$p_f = p_m = 1 - \Phi(\frac{\sqrt{(N)}\sigma_w}{2\sigma_s})$$

...where N is the dimensionality of the vectors. This means that a weaker σ_w (and thus a large M) must be compensated by increasing N. For a fixed p_f, N and M are directly proportional.

A note about so-called false-alarm "probabilities": in security applications, false alarms do not occur at random, but are often engineereed. Hence p_f should be intepreted not as a probability, but as an indicator of the feasibility of a brute-force attack. If an attacker randomly generates legal watermarks in hopes of finding one which sets of the detector (an attack which can not be prevented), he will succeed after approximately $1/2p_f$ attempts. This is the attacker's *worst-case* scenario, and so p_f should be far smaller than is needed in signal processing applications. We tentatively choose $p_f = 2^{-56}$, with a plan to choose smaller values in the future.

Of course, if we can choose N as large as we want, we can make all direct attack probabilities as small as we wish, while making added watermarks as weak, and thus imperceptible, as we wish. However, images only contain so much information, and extracting hundreds of thousands of useful feature coefficients can be impractical.