Herbert West – Deanonymizer

Mihir Nanavati, Nathan Taylor, William Aiello and Andrew Warfield
University of British Columbia
{mihirn, tnathan, aiello, andy}@cs.ubc.ca

Abstract

The vast majority of scientific journal, conference, and grant selection processes withhold the names of the reviewers from the original submitters, taking a better-safe-than-sorry approach for maintaining collegiality within the small-world communities of academia. While the contents of a review may not color the long-term relationship between the submitter and the reviewer, it is best to not require us all to be saints. This paper raises the question of whether the assumption of reviewer anonymity still holds in the face of readily-available, high-quality machine learning toolkits. Our threat model focuses on how a member of a community might, over time, amass a large number of unblinded reviews by serving on a number of conference and grant selection committees. We show that with access to even a relatively small corpus of such reviews, simple classification techniques from existing toolkits successfully identify reviewers with reasonably high accuracy. We discuss the implications of the findings and describe some potential technical and policy-based countermeasures.

1 Introduction

"...I believe the paper in its present form requires substantial modifications before it is fit for publication...." [18]

How many of us, publishing in peer-reviewed forums, have been on both the giving and receiving end of such statements? Throughout our careers we have all read anonymous feedback from program committee (PC) members, perhaps blistering in its ferocity, and, against our better judgement, tried to deduce the identity of the author by contrasting the style with our recollections of past writings of members of the PC. PC members feel comfortable being frank, and occasionally, even stinging, in their evaluation of work knowing that their identity is concealed from the paper's authors. The paper review process, and indeed, academic research as a whole, depends on honest, objective, and occasionally unpleasant appraisals of each other's work.

Broadly speaking, anonymity is a mechanism that aids objectivity and honesty. It is applied when specific knowledge of actors' behaviour could have unfortunate consequences for either the individuals themselves or the environment in which they act as a whole. In an election, votes are anonymous to avoid both retribution and bribery. Corporate stewardship policies often guarantee anonymity for whistleblowers to ensure that dangerous or unethical practices may be exposed without fear of consequences. Incognito forum postings are the key to the purity – and sometimes, ill-temperedness – of the avant-garde humor of bulletin boards such as 4chan. In their own words, "Anonymity is authenticity. It allows you to share in a completely unvarnished, unfiltered, raw way." [6] For our part, the academic peer review process uses anonymity because humans are human: Despite best intentions, it is difficult not to think positively of a favorable reviewer and begrudgingly of a negative one.

But, are these assumptions about anonymity in the peer review process still well founded? The infinite memory of the Internet means that a review can persist, either buried in someone's email or home directory, or in a publicly-accessible location, like a HotCRP server, long after the conference is held. Text stylometics, the study of extracting patterns unique to a particular author may be used alongside off-the-shelf machine learning algorithms in order to classify and label documents. We examine the possibility of applying such probabilistic text analysis techniques to automate the process of deanonymizing conference reviews. To be able to reconstruct the author of a review once is a matter of good luck. To be able to do this repeatably would spell a fundamental shift in how the peer review process plays out.

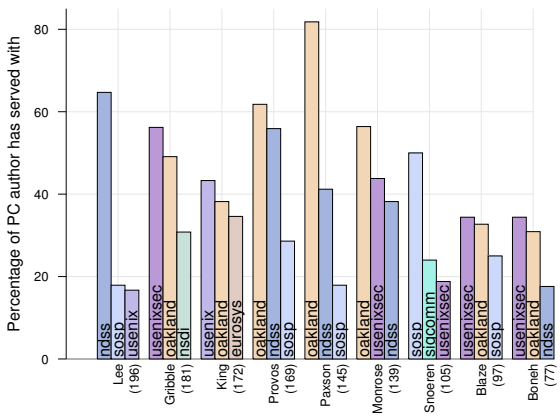

Figure 1: Percentage of 2011 conference PC membership with review samples available to HotSec'11 committee members (in parentheses, total number of individuals the PC member has served with in our dataset)

2 Threat Model

Blind reviewing allows reviewers to evaluate submissions on the basis of merit, without being unduly influenced by the political consequences of their critique. As one example, it protects young researchers who might be too intimidated to honestly appraise the work of senior, more established peers. While the value of double blinding has been questioned [16, 4], a majority of academics favor some form of blinded review over open review [12].

Attacks against anonymity represent a significant challenge to the trust-based underpinnings of the peer review process. While such attacks have typically sought to reveal the identity of authors of submissions [8, 3], this paper attempts to undermine the anonymity of the reviewers. In contrast to attacks relying on malicious Postscript submissions with identity-leaking covert channels [2], our threat model assumes no such side-channels, nor the existence of any file format vulnerabilities.

Our attacker is an author of a paper, who, having served on at least one PC, has submitted to a conference using blind reviewing and is attempting to identify the author of a particular review. This identification uses stylometric analysis and text-based classifiers trained on non-blinded reviews acquired as a member of a PC. We assume that any metadata that might leak information about the identity of the reviewer has been stripped by the management system. Initially, we assume that the attacker receives the entire "for the authors" review text, as written by the reviewer. When we consider countermeasures, in Section 5, the text may be transformed before being sent to the attacker.

Obtaining Review Samples PC members, typically, have non-blinded access to all the reviews of a particular conference to aid discussion, representing a rich source of sample reviews of their peers-cum-victims to save for future use. The problem is exacerbated by the relatively small size of academic communities, which results in a significant overlap between different PCs and a small degree of separation between PC members and submitters to conferences [23], thus providing attackers access to earlier reviews of several PC members.

To illustrate this, we considered twelve major systems and security conferences[1] for analysis. For each HotSec'11 PC member, we tabulated the other academics with whom they have shared PC service over the five-year period 2006-2010, inclusive. Using this, and assuming a HotSec'11 PC member submits a paper to one of these conferences in 2011, we enumerated the PC members with whom they have previously shared service. In the worst-case attack scenario, the HotSec attacker has a corpus of reviews from each PC member to use in an attempt to deanonymize his or her reviews.

For each HotSec'11 PC member, we determined the three conferences with the greatest fraction of overlap. A subset of this evaluation is shown in Figure 1.[2] For example, Vern Paxson has conceivably had access to reviews of over 80% of the members of the Oakland'11 PC. Cumulatively, the authors of multi-authored papers may have samples of the entire PC. Conferences for which a HotSec'11 PC member is also a member of are omitted, since it is intuitive that they will have access to samples of the entire PC. Broadening this to more conferences or longer histories would obviously provide a larger training set for the attacker.

3 Text Classification

Classification is a form of supervised learning. A classification algorithm trains by consuming a set of input data and related output label, and a mapping between the two. We say that the algorithm has learned if, given an input not found in the initial dataset, it is able to choose a reasonable output label with a high degree of confidence. In the case of conference reviews, the mapping we wish to learn is that between a review text and its author.

Text classification is used in a broad range of fields, ranging from technically-driven problems such as spam filtering [22] and email and online postings forensics [10, 5, 24] to attributing authorship in the humanities [20, 9, 11, 14]. Key differences between determining authorship of paper reviews versus creative and informal writing are the shorter length and more limited, technically-

[1] SOSP, OSDI, NSDI, EuroSys, SIGCOMM, Usenix, FAST, Oakland (IEEE S&P), Usenix Security, HotSec, NDSS and CCS

[2] The subset was selected randomly to ensure an unbiased sample.

oriented vocabulary of the former. With lyricism, flowery metaphors, and elegant prose taking a back seat to the rather pragmatic and often hastily thrown together-approach taken to technical writing, the opportunity for unique characteristics of the author to present themselves are restricted. As a result, we are unable to apply authorship attribution techniques that only perform well with a large training corpus or varied vocabulary.

Text Features We perform a linguistic profiling of the authors, obtaining an identifiable signature or *writeprint* [1], later used to determine the identity of the author of a given text sample. In our work, this profiling involves scoring all unigrams, bigrams, and trigrams – that is, word sequences of length one, two, and three, respectively – by their frequency in a particular author's text but weighted inversely by their frequency in the entire corpus, as in *tf-idf* [17]. Unlike pure word and bigram frequencies, this metric causes reviews to get sorted on an authorial, rather than topical, basis. For clarity, consider the set of reviews written about this paper: one would expect the frequency of words such as *reviews* and *anonymity* to be high, but *n-grams* containing these words would not help identify the author as they are likely to be often used in nearly every review.

The text is tokenized into words and stemmed to their root form, so that different inflections such as plurality or tense suffixes do not affect their frequency. Tokens include word contractions and punctuation marks to mark the prevalence of words like *they're* or *I'm* as opposed to *they are* or *I am*. An author using the Harvard comma will have a preponderance of *", and"* and *", or"* bigrams.

In practice, observed features of a particular author included persistently misspelt words, em-versus-en-dashes, and demarcations of different regions of the review. These features are further explored in Section 4.

Classification Our classifier is built using the open-source Natural Language Toolkit [15] for Python. We use the provided classes to perform multi-label naïve Bayes classification, which makes the common, simplifying assumption that all the text features are conditionally independent of one another, given an output label. While features like trigrams and bigrams may be closely correlated, in practice naïve Bayesian classification has been used successfully for text classification [7, 19, 21].

During the training stage, features from each of the authors are extracted and scored, and the top features are considered. Specified threshold values prevent features with low scores from being added, ensuring that a lack of prominent, distinctive features does not lead to insignificant features being selected simply to make the numbers. When classifying a new review, frequencies for each of the features are computed using the rules of Bayesian inference; the chosen author label is the one for which the probability of authorship of the text is greatest.

Data Set	Reviewers	Corpus Size	Avg. Review Length
class1	9	125,138	424
class2	16	225,067	403
conf1	17	45,619	217
conf2	14	43,922	488

Table 1: Size of Data Sets (in words)

Despite using a simple classification algorithm and a very basic feature set, our classifier yields satisfactory results on the sample data. Advanced classification techniques, coupled with more exhaustive feature lists discussed in other literature [11, 14, 13, 24] could be implemented to better classification accuracy.

4 Evaluation

Evaluating our classifier requires data sets representative of actual conference reviews. While obtaining such a data set may not be hard for a PC member, using a *real* data set for research purposes entails several ethical considerations not applicable to an adversary. Reviews are clearly not intended as *ex post facto* research data, and it is assumed that PC members will not use them in an inappropriate manner.

The evaluation uses four independent data sets, comprised of reviews from two seminar-style graduate classes, and two actual conferences where some of the authors of this submission were members of the PC. In the graduate courses, each student wrote reviews into a HotCRP system for at least 20 papers. All the students gave their permission to use their reviews for this study. For the target conferences, we followed the Feynmanian principle of "You Just Ask Them". With the permission of the program chairs, we mailed other members of the PC explaining the proposed usage of the reviews and asking their permission. Around 66% of the PC members across both conferences agreed to participate, resulting in the data sets listed in Table 1.

For the conference data sets the review names were anonymized using a preprocessing script, and all but the "for the author" text was excised. While statistical data like review length and corpus size was used to tune the classifier, the content of the reviews remained unread. Since the reviews in the graduate class data sets were already available through class participation, the contents were analyzed to determine linguistic tells, used as text features by the classifier.

The top five features of four different authors from the *class2* data set are shown in Table 2. Frequent misspellings such as *particularily* in the case of the first au-

Author 1	Author 2	Author 3	Author 4
“particularily”	“your system”	“looks at”	“Cons :”
“Lastly,”	“visualisation”	“sounds like”	“Pros :”
“Additionally”	“For example”	“what extent”	“awesome”
“so-called”	“is generally well-written”	“paper looks”	“problems existing”
“Indeed,”	“easy to follow”	“anomalous”	“slowdown factor”

Table 2: Distinctive Author Features

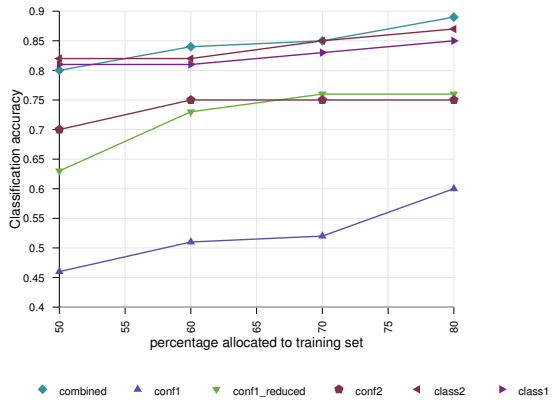

Figure 2: Classifier Accuracy

thor score highly. Review structure, such as giving feedback in enumerated lists, or section headings in reviews are usually good indicators of author identity, as are both the use and choice of connective adverbs, such as *lastly*, *additionally*, *etc.* Amusingly, the phrases reviewers use to signify approval or soften their criticism about papers tend to be both rather unique and quite distinguishable.

Classifier Accuracy The *accuracy* of of a multi-label classifier is simply the fraction of the test set that the classifier labels correctly. To measure the accuracy of our classifier, we measured the median accuracy over 15 cross-validated runs. Cross-validation randomly divides the data set into training and testing data, to prevent overfitting for features present in a specific subset of the data. Each run considered the 500 most significant text features for every author, pruned by a lower bound to ensure insignificant features were not adding noise to the classifier.

Figure 2 displays the accuracy of the classifier for different partition sizes for each of the data sets. In addition to the four data sets in Table 1, we use two data sets explained below. We believe that the results from three of the four original data sets demonstrate that such an approach is viable, even with a small amount of data, and improves significantly with larger corpus sizes.

The *conf1* data set shows surprisingly low accuracy even when accounting for the short average length of reviews. Further examination revealed high variance in the

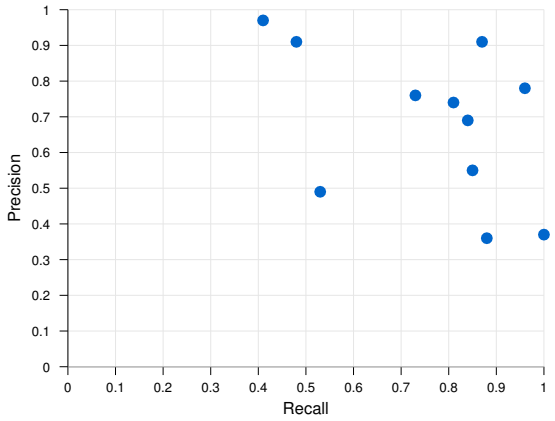

Figure 3: Identifiability of Authors in *conf1*

total corpus size for individual authors in the data set. Not only are the authors with extremely small corpuses hard to classify themselves, but they also added a significant amount of noise to the classifier disturbing the results of all other authors. *conf1_reduced* uses the same conference reviews as *conf1*, but discards any authors with a corpus size of less than 2500 words. This eliminates seven of the reviewers, resulting in a data set with ten reviewers. The improvement in the results shown in the graph is more than would be expected purely because of the decrease in number of classification labels.

combined simulates a real world attack, where an attacker collects review samples of program committee from several different conferences. We create a data set of twenty-three reviewers by combining *class1* and *class2* and merging the data of overlapping authors. While we do not combine *conf1* and *conf2* because combining overlapping reviewers would reveal their identities to us, we expect to have similar results.

Author Identifiability We quantify the classifiability of individual reviewers using two common metrics: *precision* for a given author is the fraction of the reviews attributed to an author that have been labelled correctly, whereas *recall* is the fraction of all reviews written by that author that have been identified correctly. If no incorrect classifications occur, both precision and recall would be equal to 1.

Figure 3 plots the precision and recall of all the reviewers in *conf1*. While both metrics are important, for review deanonymization we focus on precision because increasing author coverage at the expense of incorrect attributions is not a worthwhile trade-off. The classifiability of individual reviewers is not uniformly distributed across the entire set, with some reviewers having easily classifiable styles while other are more prone to misclassification.

5 Discussion

Preliminary evaluation of the classifier shows a reasonable degree of success in deanonymizing conference reviews using only a small number of training samples. Given the simplicity of the implementation, and the inevitable progress in technology we believe that building deanonymization tools using off-the-shelf toolkits is entirely feasible, making examining possible technological and social responses more important than ever.

Technological Responses: Text Reanonymization Technological countermeasures include transforming the text of the review to eliminate particular features that reveal the identity of the author. A reviewer can, individually, attempt to obfuscate the text to randomly deviate from these features, or may mimic the features of another member of the PC. Collectively, a review system can normalize the entire corpus to the most common features. Since a transformed review must largely preserve the semantics of the original to be useful, the space of possible transformations is limited.

Language Translation Despite advances in machine translation, the translation services provided by BabelFish, Bing and Google add artifacts that obfuscate the text. Translation artifacts are magnified with the distance between language families. Simple obfuscation is achieved by translating text from a starting language to an intermediate one, and back again. Greater amounts of obfuscation can be obtained by iterating this procedure or cycling through more than one intermediary language.

We use Bing Translator to obfuscate the *class2* data set, by performing a single translation cycle using German as an intermediate language. Obfuscating the reviews of only a single author makes them trivially identifiable, both to the classifier and to human beings. Obfuscating the entire corpus, however, greatly reduces classification accuracy: from 82% to 62% with half the corpus used as training data, and from 85% to 66% with 80% used for training. A randomly selected paragraph from a review is shown before and after translation in Table 3. However, a casual examination of the translated text reveals that improved anonymity comes at the price of massively-reduced understandability.

HotCRP anonymity assistance plug-in We envisage an "Office Assistant"-style HotCRP plug-in to help normalize reviews via analysis of an author's review as it is being written. If the author were to write a highly ranked feature, the plug-in would flag it in real-time as being unduly characteristic of that author and therefore leaking of his or her identity. In addition to obfuscating authorship, given the corpus of already-written reviews in the HotCRP system, the plug-in could also help the author mimic the style of *another* PC member.

The capabilities of the plug-in are conjecture at this point. However, it is clear that, as in other areas of security and privacy, the discussion points to an arms race between deanonymization and reanonymization techniques, generating interesting research questions in machine learning and natural language processing.

Social Responses: Collaboration and Peer Pressure Not all approaches to dealing with such deanonymization need be technological. Short reviews tend to be the hardest to classify, but have the unfortunate side effect of failing to provide useful feedback to the original authors. Faculty can conscript graduate students to rewrite their reviews to avoid detection. While the churn in many labs may be sufficient to confound any classifier, this could turn into a race between advances in classification techniques, and the speed at which students graduate. Other methods to outsource this rewriting, while maintaining the semantic integrity, bear further examination.

Peer pressure can be an excellent tool in curbing the use of such deanonymizers. While we believe that the majority of PC members feel strongly about the ethics involved and do their utmost to avoid breaking the spirit of anonymity, these values could be reinforced in several ways. The PC chair could, for instance, emphasize the need to avoid attempting to deanonymize submissions or subsequent reviews to maintain the fidelity of the review process. Furthermore, invitations to join a PC could be accompanied by a "terms and conditions", detailing a fair usage policy for reviews. While this would not prevent a determined adversary from building a corpus of reviews, they would be aware that it was a frowned-upon practice.

While we have focused on attacks from within the community, with members of the clique exploiting privileged information, a complete outsider can use publicly available text samples for such an attack. Many researchers contribute to technical blogs or have professional web pages detailing different aspects of their work. Doctoral theses are substantial bodies of single-author technical text. Major conferences, such as SIGCOMM in 2006, have flirted with an open review system with non-blinded reviews available to everyone. Some academics [18] publicly post anonymous reviews they have received. A clustered collection of such reviews could hint at the identity of the reviewer.

Preventing the usage of public data as training samples is more difficult. Tacit agreements or explicitly stated terms may deter the submitter from using deanonymization tools; attempting to enforce it, however, is a quagmire the community may elect to avoid altogether by embracing an open review system. Regardless of the end result, we believe that these issues need to be engaged by the community at large.

Original Text	Translated Text
Sensitive data, present in the memory of the system, is liable to exposure. Applications may have sensitive information, such as passwords, on their heap which is not explicitly cleared.	It is exposure to sensitive data in the memory of the system. Applications can have on their heap sensitive information, such as passwords, that is not explicitly disabled.

Table 3: Sample of Translated Text

6 Conclusion

Anonymous speech breaks the chain between the expression of an idea and its consequences, facilitating truthful reporting in journalism and governance, as well as the scientific peer review process. The shielding of identity is, however, a double-edged sword. Critics of anonymity point out that the resulting loss of accountability allows individuals to be unfair, or even downright cruel – witness the incidence of cyber-bullying on social network sites. In the case of conference submissions, proponents of open review systems contend that reviewers would be incentivized to write more constructive reviews if their comments were visible to peers.

We believe that, in the absence of adequate technological and policy responses, the ubiquity and long shelf-life of data along with the steady march of machine learning will threaten the underlying assumptions of anonymity. While this may not, in itself, be an undesirable outcome, the violation of anonymity that has been taken for granted may expose people to retaliation. We hope to encourage the community into examining the ramifications of this loss of anonymity, and possible countermeasures with enough vigour to prevent such collateral damage.

7 Acknowledgements

This work would not have been possible without the participation of both members of the PC of the conferences and the graduate students who graciously allowed their reviews to be used as data for classification. We would also like to thank the anonymous HotSec reviewers for their feedback and comments about the paper.

References

[1] ABBASI, A., AND CHEN, H. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. In *ACM TOIS* (2008).

[2] BACKES, M., DRMUTH, M., AND UNRUH, D. Information flow in the peer-reviewing process. In *IEEE S&P* (May 2007).

[3] BLANK, R. M. How blind is blind review? *American Psychologist 39* (1984), 1491–94.

[4] BLANK, R. M. The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *AER 81* (December 1991).

[5] DE VEL, O. Mining e-mail authorship. In *KDD* (2000).

[6] EWALT, D. Why anonymity rules. *Forbes* (March 12 2011).

[7] HAND, D., AND YU, K. Idiot's Bayes - Not so stupid after all? *International Statistical Review* (2001).

[8] HILL, S., AND PROVOST, F. The myth of the double-blind review? Author identification using only citations. *SIGKDD Explor. Newsl. 5* (December 2003), 179–184.

[9] HOLMES, D., AND FORSYTH, R. The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing* (1995).

[10] IQBAL, F., BINSALLEEH, H., FUNG, B. C. M., AND DEBBABI, M. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation 7*, 1-2 (October 2010), 56–64.

[11] JUOLA, P. Authorship attribution. *Found. Trends Inf. Retr. 1* (December 2006).

[12] KMIETOWICZ, Z. Double blind peer reviews are fairer and more objective, say academics. *British Medical Journal* (2008).

[13] KOPPEL, M., AKIVA, N., AND DAGAN, I. Feature instability as a criterion for selecting potential style markers. *J. Am. Soc. Inf. Sci. Technol. 57* (September 2006).

[14] KOPPEL, M., AND SCHLER, J. Exploiting stylistic idiosyncrasies for authorship attribution. In *IJCAI03 Workshop on Computational Approaches to Style Analysis and Synthesis* (2003).

[15] LOPER, E., AND BIRD, S. NLTK: The Natural Language Toolkit. In *ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* (2002).

[16] MADDEN, S., AND DEWITT, D. Impact of double-blind reviewing on SIGMOD publication rates. *SIGMOD Record 35* (2006).

[17] MANNING, C. D., RAGHAVAN, P., AND SCHTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[18] MEMBER OF THE 18TH INTERNATIONAL COLLOQUIUM ON STRUCTURAL INFORMATION, A. P., AND (SIROCCO), C. C. Review of *Multiparty Equality Function Computation in Networks with Point-to-Point Links*, 2011.

[19] MOSTELLER, F., AND WALLACE, D. L. *Applied Bayesian and Classical Inference - The Case of The Federalist Papers*. 1964.

[20] MOSTELLER, F., AND WALLACE, D. L. *Inference and Disputed Authorship: The Federalist*. 1964.

[21] RENNIE, J. D. Improving multi-class text classification with Naive Bayes. *Master's Thesis, Massachusetts Institute of Technology Tech Report AITR-2001-004* (2001).

[22] SAHAMI, M., DUMAIS, S., HECKERMAN, D., AND HORVITZ, E. A Bayesian approach to filtering junk E-mail. In *Learning for Text Categorization* (1998).

[23] SAVAGE, S. On the caching and prefetching of program committees. *ACM SIGCOMM, Outrageous Opinions Session* (1999).

[24] ZHENG, R., LI, J., CHEN, H., AND HUANG, Z. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol. 57* (February 2006).