Image-Based Hoax Detection

Giulio Angiani
Department of Engineering and Architecture
University of Parma, Italy
giulio.angiani@unipr.it

Gaudioso Junior Balba
Department of Engineering and Architecture
University of Parma, Italy

Paolo Fornacciari
Department of Engineering and Architecture
University of Parma, Italy
paolo.fornacciari@unipr.it

Gianfranco Lombardo
Department of Engineering and Architecture
University of Parma, Italy
gianfranco.lombardo@unipr.it

Monica Mordonini
Department of Engineering and Architecture
University of Parma, Italy
monica.mordonini@unipr.it

Michele Tomaiuolo
Department of Engineering and Architecture
University of Parma, Italy
michele.tomaiuolo@unipr.it

ABSTRACT

In the last few years, the impact of information spread through online social networks has continuously grown. For this reason, understanding the trustworthiness of news has become one of the most important challenges for an Internet user, especially during crisis events or in political, health and social issues. As part of a more comprehensive project for the detection of fake news, this paper proposes a machine learning method to evaluate the trustworthiness of a piece of information especially considering its associated image. In the work described in this paper, the training and test datasets have been first collected from the web, downloading more than 1000 images related to trusted and fake Facebook pages. All collected images have been processed using the Google Vision online service for extracting their specific internal details. For each image, various kinds of features have been considered, including its color composition, the recognized objects, the list of sites in which it is published, and eventually the contained text. These details have been then used for training a classifier using different algorithms which allowed us to reach an accuracy of about 85% in hoax identification. Future research will focus on social-network information related to images, to improve the system accuracy and acquire more knowledge about various types of news spread online.

Goodtechs '18, November 28–30, 2018, Bologna, Italy

ACM ISBN 978-1-4503-6581-9/18/11…$15.00
https://doi.org/10.1145/3284869.3284903

KEYWORDS

Fake news, Hoax detection, Image analysis

ACM Reference Format:
Giulio Angiani, Gaudioso Junior Balba, Paolo Fornacciari, Gianfranco Lombardo, Monica Mordonini, and Michele Tomaiuolo. 2018. Image-Based Hoax Detection. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs '18), November 28–30, 2018, Bologna, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3284869.3284903

1 INTRODUCTION

In the last years, more and more people use the Web and especially online social networks for acquiring information about current issues. This phenomenon, which is already considered by many as indispensable, has led Internet users to share a very large amount of data and news, often without verifying its reliability/credibility. For this reason, one of the most actual and important challenges for researchers, but also for all network users, is to discern the authentic information from fake news [14, 19]. In order to reach this goal, many steps often must be followed by a user, for example performing an image search on the web, looking for original information sources, checking for a user who posted particular news. So, almost all people prefer sharing or believing in a piece of information without controlling its trustworthiness. This user's behaviour is very observable on social network platforms like Twitter and Facebook, but also on instant messaging systems like Whatsapp. Regarding this issue, this research work proposes a machine learning approach, starting from the analysis of the images contained in many posts, collected from Facebook.

A more complete project, which we are still developing, has the goal to estimate the credibility of news, analyzing four aspects: *i* the *text* posted by a user with text-analysis techinques, like in [6]; *ii* the study of the *community* which the posting user belongs to; *iii* the analysis of the reliability of external related sources (e.g., cited websites, linked Facebook pages, Twitter reported profiles, etc.); and finally *iv* the pure analysis of the images, possibly posted together with a text. All these aspects are estimated separately with different machine-learning algorithms and then the final result is extracted by an ensemble-learning approach.

The research field of image analysis is widely studied, for many and varied reasons. Usually it's possible to retrieve various kinds of information related to an image, including (*ii*) the objects it contains (object-recognition [13]); (*ii*) its graphical composition; and (*iii*) the proportion of used colors [17].

In this paper, we present in particular a methodology for the analysis of information credibility using image analysis techniques, and the obtained results.

We have downloaded more than 2000 posts from two kinds of Facebook [8] accounts: (*i*) trusted, like institutional or newspaper profiles; and (*ii*) untrusted, like satirical profiles, or profiles already labeled from other Facebook users as fake or unreliable.

By using all these data we have been able to develop an automatic system which can correctly classify news as true or fake with an accuracy of about 84%, applying the Random Forest algorithm exclusively on the data extracted from images.

The rest of the manuscript is organized in the following way: Section 4 describes the data collection and methodology used for this study; results and discussions are presented in Section 5; finally, Section 6 provides some concluding remarks.

2 LITERATURE REVIEW

The web and social media has changed the dynamics of information transmission and the role of the trust of an information in the detection of online anti-social behaviors and various techniques based on artificial intelligence have been proposed for their automatic detection [5]. The spread of misinformation in such a context might be particularly difficult to detect and correct because of the social reinforcement, that is people are more likely to trust an information someway consistent with their system of belief [4]. Automatic hoax recognition has always been a challenge in the Internet. For example, in [21] the authors present an intelligent automatic hoax detection system based on neural networks and advanced text processing. In [10] the authors define a ranking function based on the analysis of the text content to measure the blog credibility: they estimate the blog credibility by exploiting the quantity structure and the content similarity in reference to a German news corpus.

Various agencies in the USA have wondered if and how media technology can influence democratic life in the country and, most recently, the focus of concern has shifted to social media which have a different structure than previous media technologies: content can be relayed among users with no editorial judgment [1]. Some automatic framework have been studied; for example in [16] a machine learning framework to automatic detect the early stages of viral spreading of political misinformation is presented.

In [7], the authors examine the role that social media played during the Gezi protests, with an emphasis on how trust was built and maintained among the protestors. Technological affordances worked as an interface facilitating the social identification process. Trust in the person behind the information was a major criterion for trusting the information.

Recently, an automatic detection of fake news is presented in [15] based on the identification of linguistic differences in fake and legitimate news content. In [12]the authors try to elicit typical characteristics of Wikipedia hoaxes in order to build an automatic classification system to determine whether a given article is a hoax. In particular, they aim to gain a better understanding of how hoaxes differ from legitimate articles and if there are any features that make a legitimate article be mistaken for a hoax. An interesting work related to detect hoaxs on social networks is showed in [20], in which the authors demonstrated that Facebook posts can be classified with high accuracy as hoaxes or non-hoaxes on the basis of the users who "liked" them. Their classification techniques are based on logistic regression and boolean crowdsourcing algorithms.

Finally, in [11] the authors describe a survey about false information on web and social media and they classify algorithms to identify fake news (that is, opinion-based false information) and hoaxes (that is, fact-based false information) into three categories: featured-based, graph-based, and modeling-based. In fact, many approaches to recognize hoax or fake news are based on text features, while image features are often ignored. According to the study described in [9] however, images are very popular and have a great influence on microblogs news propagation. Moreover, it is possible to find distinctive distribution patterns and different visual content for images in the real and fake news. Often the most recent techniques of machine learning are used to include visual content in the detection of a fake news. In [18] a deep multimodal learning model is employed for assessing the integrity of an image by combining its visual content with related information from a knowledge base. In [22] the Adversarial Neural Networks (ANN) are at the basis of a multi-modal (text and image) system for fake news detection. In particular, in the last article, the visual features and the text features are simply concatenated to create a multimodal features set that are the input of the ANN-based fake news detector.

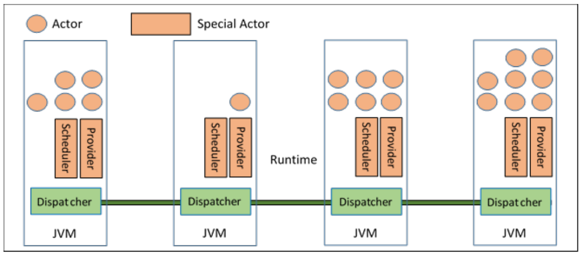

Figure 1: Distributed ActoDeS application architecture.

3 APPLICATION CONTEXT

This research work is part of a more comprehensive project for the detection of fake news, characterized by a multi-faceted approach. The project is based on ActoDES, which is a software framework which adopts the actor model for simplifying the development of complex distributed systems [3]. In particular, it includes two types of actors: a scheduler and a service provider. The former has the duty to handle the concurrent execution of actors, while the latter provides runtime services, needed by actors to complete their tasks. A subscription service is also available, to facilitate the development of collaborative applications with actors. This service has the task to receive incoming messages into a specific mailbox, and forward them to subscriber actors. The actors can eventually handle the messages differently, according to their own behaviors.

The whole system design is composed of four main components, based on ActoDES, related to different kinds of analysis that are performed for identifying fake news. Namely, these subsystems are:

- **Text**, for the anaysis of texts posted by a user, with text-analysis techinques, like in [6];
- **Community**, to perform social network analysis on the ego-network of a given user;
- **Source**, for the analysis of the reliability of external related sources (e.g., cited websites, linked Facebook pages, Twitter reported profiles, etc.);
- **Image**, to analyze the images possibly posted together with a text.

To realize such a system, other services, developed for this project, provide additional functionalities to actors, for the continuous analysis of various social streams. In particular, a Facebook service allows other actors to send various kinds of requests:

- User diary, to obtain the recent posts of a specified user and save them in a local storage system.
- Content query, to similarly obtain recent posts published on Facebook and selected according to some constraints specified by the actor.
- Stream, to continuously receive messages published on the platform during the execution; posts are obtained in the form of JSON objects, which are then stored in a NoSQL repository (namely a MongoDB database).

4 METHODOLOGY

The focus of this section is on how the extracted information from Google Cloud Vision API has been processed for the creation of the dataset. The features, as briefly anticipated in the introduction, are: (*i*) color composition in RGB; (*ii*) presence of faces and their relative emotions; (*iii*) the contained text; (*iv*) the objects present inside the image; (*v*) categorization labeling by Google; (*vi*) the pages and links where the image is found; and (*vii*) safe search analysis.

Distribution on the Web (DW)

Google Image Search retrieves all the links (each one could be a web page or an image source) where the searched image is used or stored. Specifically, the API divides all these links in three categories: (*i*) Fully equal image sources; (*ii*) Partially equal image sources; and (*iii*) Web pages where those fully/partially equal images are used. The features selected in this case are the quantities of the above-mentioned categories.

Safe Search (SS)

The SafeSearch analysis gives 4 evaluations according to the content of an image: Adult, Violence, Medical and Spoof. For each of them, a score that ranges from 1 to 5 is assigned. It is to be emphasized that the Spoof evaluation is profoundly based on the likelihood of the image being modified. An example of the GCV response for Safe search is reported below.


```
{
  "responses": [
    {
      "safeSearchAnnotation": {
        "adult": "VERY_UNLIKELY",
        "spoof": "UNLIKELY",
        "medical": "VERY_UNLIKELY",
        "violence": "VERY_UNLIKELY",
        "racy": "VERY_UNLIKELY"
      }
    }
  ]
}
```


Color Composition (CF)

The API gives the 10 most dominant colors of an image, providing for each of them 2 types of value: score and pixel fraction. Pixel fraction, just as the name suggests, is the fraction of pixels that the color occupies in the analyzed image, while the score's value is based on how much visual impact the color has, not taking account of how much space it occupies. The processing selected is the calculation of the weighted mean of the R, G and B values of the dominant colors, having their relative score and pixel fraction as weights. At the end of this process, these are the 6 features to be included in the dataset: Impacts and Pixel Fractions of Red, Green and Blue in each image. A snippet of the GCV response for Color Composition service request is reported below.


```
{
  "responses": [
    {
      "imagePropertiesAnnotation": {
        "dominantColors": {
          "colors": [
            {
              "color": {
                "red": 123,
                "green": 113,
                "blue": 88
              },
              "score": 0.25912306,
              "pixelFraction": 0.05346841
            },
            {
              "color": {
                "red": 110,
                "green": 123,
                "blue": 64
              },
              "score": 0.20531657,
              "pixelFraction": 0.08815253
            },
            [...]
          ]
        }
      },
[...]
  ]
}
```


Faces and Emotions (FA)

Face recognition is one of the first interesting application of machine-learing on images [23]. IN GCV there are 7 kinds of analysis for each detected face: joy, anger, sorrow, surprised, headwear, blurred, and under exposed. Each of them is given a degree of likelihood that varies from 0 to 6. In order to get more significant information from these features, aside from calculating the average degree for each emotion, the number of occurrences of strong emotions (with at least 4 degree likelihood) is also taken into consideration. An example is reported below.


```
{
  "joyLikelihood": "VERY_LIKELY",
  "sorrowLikelihood": "VERY_UNLIKELY",
  "angerLikelihood": "VERY_UNLIKELY",
  "surpriseLikelihood": "VERY_UNLIKELY",
  "underExposedLikelihood": "VERY_UNLIKELY",
  "blurredLikelihood": "VERY_UNLIKELY",
  "headwearLikelihood": "VERY_UNLIKELY"
}
```


Detected Objects (DO)

Just as its name suggests, this feature tries to analyze what are the objects that can be found inside an image, and a score varying from 0 to 1 is assigned for every object detected. Obviously the images do not contain the same objects so in order to have a well-aligned dataset, all the detected objects from all the images are collected and chosen as features: assigning its relative score if it is present in a given image, otherwise 0. Here an example of score obtained posting an image with some refugees on a boat.


```
{
 "labelAnnotations": [
    {
      "description": "water",
      "score": 0.94952434,
    },
    {
      "description": "men on boat",
      "score": 0.94455415,
    },
    [...]
  ]
}
```


Web Entity (WE)

This feature is similar to the previous one, but it goes deeper in details; it tries to get the identity or even the characteristics of the object or person found in the image; it can even try to extract its main source (news source in this case). Of course all of this information on Web entities is completely dependent on the knowledge of Google across the web, so somehow this could be considered as a biased feature. A Web entity tag has also got its score, which can vary from 0 to an even larger number than 1. The procedure applied for aligning scores in object detection was also applied for this feature.

OCR

This operation has also been applied to the images in order to extract possible texts from them. Aside from the presence of text, a simple sentiment analysis has also been applied to the extracted texts, which gives two types of scoring: the negativity or positivity of the text that goes from -1 to 1, and the magnitude of the said sentiment that goes from 0 to any positive number. In our set of images, only 312 (177 hoax, 135 non hoax) have readable texts, which are relatively small numbers and also almost equally partitioned between the 2 classes. In fact, in the validation phase, the presence of OCR features would not affect the quality of the models. Of course, these data could still be used to fine tune the part of the system that takes care of different kinds of text analysis.

All of these features are then used in different kinds of classification models, in particular: Gaussian and Bernoulli Naive Bayes, Logistic Regression, Decision Trees, Random Forest and SVM. Classifications have also been applied to different subsets of the above mentioned features in order to know which of them could have affected the quality of the models at most.

5 RESULTS

In this section we show the results we achieved in classification using different algorithms and different sets of features. Below, we present such results, starting with those obtained using all the features identified in the project (see Tab. 1). Then, other results obtained using different features subsets are shown.

Table 1: Classification results using all the features

	All Features		
Model	Precision	Recall	F1-Score
Gaussian	0.74	0.81	0.77
Bernoulli NB	0.83	0.83	0.83
Logistic Regression	0.78	0.62	0.69
Decision Trees	0.71	0.71	0.71
Random Forest	0.77	0.86	0.81
SVM	0.54	0.62	0.57

The models using all the features exceeded the expectations, especially with Bernoulli Naive Bayes and Random Forest algorithms. In order to see which subset could have contributed to this performance, just as anticipated in the last section, we have repeated the training phase with only some subsets of the features. The first important subset is Web Entities; as shown in the table, it has an outstanding precision in all models, especially in the Random Forest one. But these results were to be expected for the fact that this feature is somehow biased and completely dependent on the kind of information Google has collected about the images. Just as stated in the previous section, Google could also trace the original source of the image (so for example, all the ones that come from Clickhole or any other well-known hoax/satire sources would be classified automatically as Hoax, the same thing goes for the Non-Hoax ones) or even the characteristic for which it is famous or notorious. In many cases Google even tags an image as hoax (exactly how it is pre-classified).

Even if this could be read as a weakness in the whole process, it can be very useful in general for the main goal to identify a fake information, especially if this result would be supported by the outcomes of the other parts of the project (as we said in section 1).

The following subsets are instead not biased, for the fact that they provide only objective information about the images: Detected Objects, Distribution on Web, Color Factors, Visual Impact. The subsets names are self-explanatory, apart from "Visual Impact". This last subset is composed of: SafeSearch (Violence, Spoof, Medicine, Adult), Face sentiment analysis and Color Composition.

Table 2: Classification results using only the web entities

	Web Entities (WE)		
Model	Precision	Recall	F1-Score
Gaussian	0.77	0.82	0.79
Bernoulli NB	0.82	0.87	0.84
Logistic Regression	0.88	0.58	0.70
Decision Trees	0.80	0.69	0.74
Random Forest	0.81	0.92	0.86
SVM	0.77	0.65	0.71

Table 3: Classification results using only the detected objects

	Detected Objects(DO)		
Model	Precision	Recall	F1-Score
Gaussian	0.51	0.75	0.61
Bernoulli NB	0.67	0.68	0.68
Logistic Regression	0.66	0.63	0.65
Decision Trees	0.59	0.66	0.62
Random Forest	0.66	0.73	0.70
SVM	0.73	0.56	0.64

Table 4: Classification results using Distribution on Web

	Distribution on Web (DW)		
Model	Precision	Recall	F1-Score
Gaussian	0.56	0.51	0.54
Bernoulli NB	0.53	0.93	0.68
Logistic Regression	0.56	0.47	0.51
Decision Trees	0.58	0.58	0.58
Random Forest	0.60	0.62	0.61
SVM	0.57	0.53	0.55

Table 5: Classification results using Color factors

	Color Factors (CF)		
Model	Precision	Recall	F1-Score
Gaussian	0.51	0.69	0.58
Bernoulli NB	0.48	0.99	0.65
Logistic Regression	0.54	0.64	0.59
Decision Trees	0.49	0.52	0.51
Random Forest	0.48	0.53	0.50
SVM	0.50	0.67	0.57

The one that performed well among these subsets, but not better than "Web Entities", is "Detected Objects"; showing that the objects that are present inside an image could be a good indicating

Table 6: Classification results using all visual features

	Visual Impact (CF, FA,SS)		
Model	Precision	Recall	F1-Score
Gaussian	0.57	0.60	0.59
Bernoulli NB	0.47	0.55	0.51
Logistic Regression	0.54	0.64	0.59
Decision Trees	0.53	0.56	0.54
Random Forest	0.58	0.65	0.62
SVM	0.56	0.68	0.61

factor. Instead, the one that performed poorly is the "Color Factors". Surprisingly the subset "Distribution on Web" has got a decent precision, it means that even the way the image is spread on the Web could be a factor in understanding whether it is hoax or not.

6 CONCLUSIONS AND FUTURE WORKS

The problem of hoax detection is very actual for users of online social networks and it is still an open research topic. This work has explored the possibility to identify fake news starting by the study of an image. In particular, the analysis is based on the extraction of various kinds of features from images, including: the contained objects, the proportion of different colors, the presence or absence of faces and their detected emotions, and some other information gathered using Google Cloud Vision service. As a result, we have shown that is possible to understand with a quite good accuracy if the studied image is used for disseminating some true information or fake news.

This work is part of a more comprehensive project for hoax detection. The whole system is designed for also using also other kinds of information and analysis about the source which posts some news, including its social ego-network, its Web-site structure, a sentiment analysis of the posted text and a sentence structure analysis. In future research works, we will explore also those other approaches and we will look for other objective features to reduce the impact of external knowledge in classification.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[2] ABC Australia. 2018. ABC web site. http://www.abc.net.au/news/science/2018-02-11/fake-news-hoax-images-digitally-altered-photos-photoshop/9405776

[3] Federico Bergenti, Agostino Poggi, and Michele Tomaiuolo. 2015. An Actor Based Software Framework for Scalable Applications. *Lecture Notes in Computer Science (LNCS)* 8729 (2015), 26–35. https://doi.org/10.1007/978-3-319-11692-1_3 Proc. 7th International Conference on Internet and Distributed Computing Systems (IDCS 2014); Calabria; Italy; 2014-09-22/24 [MT].

[4] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one* 10, 2 (2015), e0118093.

[5] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo. 2018. A holistic system for troll detection on Twitter. *Computers in Human Behavior* 89 (2018), 258–268. https://doi.org/10.1016/j.chb.2018.08.008

[6] Paolo Fornacciari, Monica Mordonini, and Michele Tomaiuolo. 2015. Social Network and Sentiment Analysis on Twitter: Towards a Combined Approach.. In *KDWeb*. 53–64.

[7] Gulizar Haciyakupoglu and Weiyu Zhang. 2015. Social media and trust during the Gezi protests in Turkey. *Journal of Computer-Mediated Communication* 20, 4 (2015), 450–466.

[8] Facebook Inc. 2018. Facebook web site. https://www,facebook.com/

[9] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19, 3 (2017), 598–608.

[10] Andreas Juffinger, Michael Granitzer, and Elisabeth Lex. 2009. Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd workshop on Information credibility on the web*. ACM, 51–58.

[11] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* (2018).

[12] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 591–602.

[13] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Vol. 2. Ieee, 1150–1157.

[14] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.

[15] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. *arXiv preprint arXiv:1708.07104* (2017).

[16] Jacob Ratkiewicz, Michael Conover, Mark R Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. *ICWSM* 11 (2011), 297–304.

[17] John C Russ. 2016. *The image processing handbook*. CRC press.

[18] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep Multimodal Image-Repurposing Detection. *arXiv preprint arXiv:1808.06686* (2018).

[19] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600

[20] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[21] Marin Vuković, Krešimir Pripužić, and Hrvoje Belani. 2009. An intelligent automatic hoax detection system. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 318–325.

[22] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.

[23] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM computing surveys (CSUR)* 35, 4 (2003), 399–458.