Overview of PAN'16

New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation

Paolo Rosso[1], Francisco Rangel[2], Martin Potthast[3], Efstathios Stamatatos[4],
Michael Tschuggnall[5] and Benno Stein[3]

[1]PRHLT Research Center, Universitat Politècnica de València, Spain
[2]Autoritas Consulting, S.A., Spain
[3]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany
[4]Dept. of Information & Communication Systems Eng., University of the Aegean, Greece
[5]Department of Computer Science, University of Innsbruck, Austria

pan@webis.de   http://pan.webis.de

Abstract This paper presents an overview of the PAN/CLEF evaluation lab. During the last decade, PAN has been established as the main forum of digital text forensic research. PAN 2016 comprises three shared tasks: (*i*) author identification, addressing author clustering and diarization (or intrinsic plagiarism detection); (*ii*) author profiling, addressing age and gender prediction from a cross-genre perspective; and (*iii*) author obfuscation, addressing author masking and obfuscation evaluation. In total, 35 teams participated in all three shared tasks of PAN 2016 and, following the practice of previous editions, software submissions were required and evaluated within the TIRA experimentation framework.

1 Introduction

Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN) is a forum for the digital text forensics, where researchers and practitioners study technologies that analyze texts with regard to originality, authorship, and trustworthiness. The practical importance of such technologies is obvious for law enforcement and marketing, yet the general public needs to be aware of their capabilities as well to make informed decisions about them. This is particularly true since almost all of these technologies are still in their infancy, and active research is required to push them forward. Therefore, PAN focuses on the evaluation of selected tasks from digital text forensics in order to develop large-scale, standardized benchmarks, and to assess the state-of-the-art techniques. The targeted audiences in terms of research areas range from linguistics and computational linguistics to data mining and machine learning; targeted audiences in terms of users of envisioned tools are professionals, such as forensic linguists, copyright protectors, lawyers, criminal investigators, and educators, but also laymen web users.

Previous editions of PAN have been organized in the form of workshops (2007 - 2009) as well as evaluation labs (2009 - 2015), and they were held in conjunction with the conferences SIGIR, ECAI, SEPLN, and in the recent years CLEF and FIRE. Tables 1 and 2 overview key figures of PAN/CLEF and PAN/FIRE labs. At PAN'16 we

Table 1. Key figures of the PAN workshop series since 2009.

Statistics	**SEPLN**	**CLEF**						
	2009	2010	2011	2012	2013	2014	2015	2016
Follower	78	151	181	232	286	302	333	
Registrations	21	53	52	68	110	103	148	147
Runs/Software	14	27	27	48	58	57	54	35
Notebooks	11	22	22	34	47	36	52	26
Attendees	18	25	36	61	58	44	74	-

Table 2. Key figures of the FIRE workshop series since 2011.

Statistics	**FIRE**				
	2011	2012	2013	2014	2015
Follower					
Registrations	6	12	16	20	31
Runs/Software	6	8	8	17	20
Notebooks	6	2	6	4	6
Attendees	6	2	6	3	9

focused on authorship tasks from the fields of (*i*) author identification, (*ii*) author profiling, and (*iii*) author obfuscation evaluation. More specifically, the tasks will include two variants per field, namely author clustering and diarization, age and gender prediction, and author masking and obfuscation. A brief introduction to each of them follows (see Figure 1), more details are given in the corresponding sections.

- Author Clustering/Diarization. Author clustering is the task where given a document collection the participant is asked to group documents written by the same author so that each cluster corresponds to a different author. This task can also be viewed as establishing authorship linking between documents. The training corpus comprised a set of author clustering problems in 3 languages (English, Dutch, and Greek) and 2 genres (newspaper articles and reviews). In PAN'16 we focused on document-level author clustering, while a variation of author clustering was included in the PAN'12 edition [23]. However, it was focused on the paragraph-level and therefore it is more related to the author diarization task. The task of author diarization is to identify different authors within a single document. Such documents may be the result of a collaborative work (e.g., a combined master thesis written by two students, a scientific paper written by four authors, ...), or the result of plagiarism. The latter is thereby a special case, where it can be assumed that the main text is written by one author and only some fragments are by other writers (the plagiarized or intrusive sections).
- Age/Gender Prediction. Since PAN'13 we have been organizing the shared task of author profiling [61,60], focussing mainly on age and gender identification (at PAN'15 also personality recognition [59]). While the goal in previous editions was to study different genres, at PAN'16 we aimed at evaluating age and gender identification in a cross-genre setting. The training was carried out on tweets, and the

test on blogs, social media and hotel reviews, in the following languages: English, Spanish, and Dutch.

- Author Masking/Obfuscation Evaluation. While the goal of author identification and author profiling is to model author style so as to deanomyize authors, the goal of author obfuscation technology is to prevent that by disguising the authors. Corresponding approaches have never been systematically evaluated for quality, nor whether they are capable of confusing existing author identification and profiling software. The author obfuscation shared task at PAN aims at closing this gap. Concretely, author masking and author obfuscation evaluation aim respectively at perturbing an author's style in a given text to render it dissimilar to other texts from the same author, and at adjusting a given text's style so as to render it similar to that of a given author. The success of corresponding approaches has been evaluated considering readability and paraphrase quality. Our final aim is to check whether the state-of-the-art techniques of author identification and author profiling research fields (the software submissions to author identification and author profiling of previous years is available on our TIRA experimentation platform) is robust against author obfuscation technology.

2 Author Identification

Previous editions of PAN focused on author identification tasks that could be handled as supervised classification problems. In particular, the task was to assign documents of unknown authorship to one of the candidate authors. This was based on the fact that for each candidate author samples of their texts were available. Variations of this task considered cases where the set of candidate authors is either closed or open [4,23] as well as a singleton (*author verification*) [26,72,71]. At PAN'16, we focus on unsupervised author identification tasks where there is lack of candidate authors and samples of known authorship. In more detail, we focus on two main tasks: (*i*) given a set of documents, identify groups of documents by the same author (*author clustering*) and (*ii*) given a single multi-author document, identify parts of document written by the same author (*author diarization*).

2.1 Author Clustering

Author clustering is the task of grouping documents by their author in a given document collection [31,63]. This task is useful in multiple domains where there is lack of reliable authorship information in document collections [21,1]. For example, in a collection of novels published anonymously we might be able to decide that they are written by a single person. Given some proclamations published by terrorist groups we can identify proclamations, either of the same or different groups, by the same authors. Provided a collection of online product reviews by users with different aliases we can extract the conclusion that some of the aliases actually correspond to the same person.

In this edition of PAN we study two application scenarios:

(a) **Complete author clustering**: This scenario requires a detailed analysis where, first, the number of different authors (k) found in the collection should be identified and, second, each document should be assigned to exactly one of the k authors.

(b) **Authorship-link ranking**: This scenario views the exploration of the given document collection as a retrieval task. It aims at establishing authorship links between documents and provides a list of document pairs ranked according to a confidence score (the score shows how likely it is the document pair to be by the same author).

In more detail, given a collection of (up to 100) documents, the task is to (*i*) identify groups of documents by the same author and (*ii*) provide a ranked list of authorship links (pairs of document by the same author). All documents within the collection are single-authored, in the same language, and belong to the same genre. However, the topic or text-length of documents may vary. The number of distinct authors whose documents are included in the collection is not given.

To evaluate the complete author clustering task, we use *extrinsic* clustering evaluation (i.e., true labels of data are available) and, in particular, *B-cubed Precision*, *B-cubed Recall*, and *B-cubed F-score*. These measures have been found to satisfy several formal constraints including cluster homogeneity, cluster completeness, and the *rag bag* criterion (where multiple unrelated items are merged into a single cluster) [3]. As concerns authorship-link ranking, we use *mean average precision* (MAP), a standard scalar evaluation measure for ranked retrieval results.

Corpora A new corpus was developed for this shared task comprising several instances of clustering problems in three languages (Dutch, English, and Greek) and two genres (articles and reviews) per language. A more detailed description of this corpus is following:

- English part: All documents have been published in the UK daily newspaper *The Guardian*.[1] Opinion articles about politics and UK were used in the training corpus while the evaluation corpus was based on opinion articles about society. Moreover, book reviews on the thematic area of culture were considered.
- Dutch part: It includes opinion articles from the Flemish daily newspaper *De Standaard* and weekly news magazine *Knack*. In addition, it comprises reviews taken from the CLiPS Stylometry Investigation (CSI) corpus [76]. These are both positive and negative reviews about both real and fictional products from the following categories: smartphones, fastfood restaurants, books, artists, and movies.
- Greek part: The opinion articles included in this part published in the online forum *Protagon*.[2] The training corpus was based on articles about politics and the evaluation part utilized articles about economy. In addition, this corpus comprises a collection of restaurant reviews downloaded from a relevant website.[3]

For each combination of language-genre, we produced several instances of clustering problems corresponding to different ratios $r = k/N$, where N is the number of documents in a given collection. This ratio indicates the percentage of single-document clusters as well as the number of available authorship links. For instance, if r is high

[1] http://www.theguardian.com

[2] http://www.protagon.gr

[3] https://www.ask4food.gr

Table 3. Statistics of the author clustering evaluation corpus. Corresponding statistics of the training corpus are inside parentheses.

Language	Genre	Instances	Avg. Docs	Avg. words
English	articles	3 (3)	70 (50)	583.2 (751.1)
English	reviews	3 (3)	80 (80)	1,015.1 (1,032.3)
Dutch	articles	3 (3)	57 (57)	1,098.6 (1,137.1)
Dutch	reviews	3 (3)	100 (100)	152.6 (129.5)
Greek	articles	3 (3)	70 (55)	736.1 (739.1)
Greek	reviews	3 (3)	70 (55)	466.7 (573.4)

then most documents in the collection belong to single-document clusters and the number of authorship links is low. In this evaluation campaign, we selected to examine the following three cases:

- $r \approx 0.9$: only a few documents belong to multi-document clusters and it is unlikely to find authorship links.
- $r \approx 0.7$: the majority of documents belong to single-document clusters but it is likely to find authorship links.
- $r \approx 0.5$: less than half of the documents belong to single-document clusters and there are plenty of authorship links.

Table 3 shows statistics of the corpus used in this evaluation campaign. As concerns the length of documents, reviews in Dutch and Greek are shorter than the corresponding articles while English book reviews are longer than English articles. The number of documents per clustering instance ranges between 50 and 100.

Results We received 8 submissions in the author clustering subtask. Following the practice of previous editions of PAN, the participants submitted their software in TIRA experimentation framework where they were able to apply their approach in both training and final evaluation corpora. The task of PAN organizers was reduced to review the output of submitted systems and publish evaluation results. A summary of the evaluation results is presented in Table 4 (average values for all instances of the evaluation corpus). The baseline corresponds to a naive approach where the provided documents are randomly grouped in clusters. Average performance of 50 repetitions of this baseline approach is shown.

In both complete author clustering and authorship-link ranking, the submissions of Bagnall and Kocher achieved the best results. A high B-cubed recall indicates that an approach tends to produce large clusters while a high B-cubed precision usually corresponds to many single-item clusters. For the authorship-link ranking task, the approaches by Bagnall and Gobeill are significantly better than the rest of participants. A more detailed presentation of evaluation results is provided in [70].

2.2 Author Diarization

The author diarization task of the PAN'16 lab continues and extends the previous tasks from 2009-2011 on intrinsic plagiarism detection [46]. The original problem is related

Table 4. Evaluation results for the author clustering task (submissions are ranked according to BCubed F-score).

Participant	**B3 F-score**	**B3 Recall**	**B3 Precision**	**MAP**
Bagnall	0.8223	0.7263	0.9765	0.1689
Kocher	0.8218	0.7215	0.9816	0.0540
Sari & Stevenson	0.7952	0.7330	0.8927	0.0399
Zmiycharov et al.	0.7684	0.7161	0.8521	0.0033
Gobeill	0.7058	0.7669	0.7373	0.1146
Baseline	0.6666	0.7140	0.6412	0.0015
Kuttichira	0.5881	0.7202	0.5122	0.0014
Mansoorizadeh et al.	0.4008	0.8218	0.2804	0.0085
Vartapetiance & Gillam	0.2336	0.9352	0.1947	0.0120

to the question, whether an author has misused text from others without proper references, and if yes, which text fragments are affected. Thereby the key word *intrinsic* indicates that potential plagiarized sections have to be found by inspecting solely the respective document, i.e., any comparisons with external sources are disallowed [74]. Consequently, authors have to be identified by analyzing the writing style in some way. This is not an artificial restriction, but has practical relevance in plagiarism detection systems, e.g., to limit or pre-order the search space, or to investigate older documents where potential sources are not digitally available.

Tasks and Corpora Based on that, the shared task at PAN'16 focuses on identifying authorships within a single document. Thereby it is not only searched for plagiarism, but also for the contributions of different writers in a multi-author document. Among examples for the latter are collaboratively written student theses or scientific papers composed by a known number of cooperating researchers. As an overall keyword for the task, the title *author diarization* has been chosen[4], consisting of three related subtasks:

(a) **Traditional intrinsic plagiarism detection**: Assuming a major author who wrote at least 70% of a document, the task is to find the remaining text portions written by one or several others.
(b) **Diarization with a given number of authors**: The basis for this subtask is a document which has been composed by a known number of authors. Participants should then attempt to group the individual text fragments by authors.
(c) **Unrestricted diarization**: As a tightening variant of the previous scenario, the number of collaborating authors is not given as an input variable for the last subtask. Thus, before/during analyzing and attributing the text, also the correct number of clusters, i.e., writers, has to be found.

For all subtasks, distinct training and test datasets have been provided, which are based on the Webis-TRC-12 dataset [54]. The original corpus contains documents on 150 topics used at the TREC Web Tracks from 2009-2011 (e.g., [12]), whereby professional

[4] The term "diarization" originates from the research field *speaker diarization*, where approaches try to automatically identify, cluster or extract different (parallel) speakers of an audio speech signal like a telephone conversation or a political debate [39].

writers were hired and asked to search for a given topic and to compose a single document from the search results. From these documents, the respective datasets for all subtasks have been generated by varying several configurations like the number and proportions of authors in a document, the decision, if they are uniformly distributed or if switches in authorships are allowed to occur within a single sentence, at the end of a sentence or only between paragraphs. As the original corpus has already been partly used and published, the test documents are created from previously unpublished documents only. Overall, the number of training/test documents for the respective subtasks are: (a) 71/29, (b) 55/31 and (c) 54/29.

Table 5. Intrinsic Plagiarism Detection Results (Problem a).

		Micro			Macro		
Rank	Team	Recall	Precision	F	Recall	Precision	F
1	Kuznetsov *et al.*	**0.19**	**0.29**	**0.22**	**0.15**	**0.28**	**0.17**
2	Sittar *et al.*	0.07	0.14	0.08	0.10	0.14	0.10

Table 6. Diarization Results (Problems b and c).

			BCubed		
#authors	Rank	Team	Recall	Precision	F
known (Problem b)	1	Kuznetsov *et al.*	0.46	**0.64**	**0.52**
	2	Sittar *et al.*	**0.47**	0.28	0.32
unknown (Problem c)	1	Kuznetsov *et al.*	0.42	**0.64**	**0.48**
	2	Sittar *et al.*	**0.47**	0.31	0.35

Results The performance of the submitted algorithms have been measured with two different metrics. For the intrinsic plagiarism detection subtask, the micro-/macro-metrics proposed in [55] have been used, whereby the ranking is based on the macro calculation[5]. On the other hand, the diarization subtasks have been measured with the BCubed clustering metrics [3], as they reflect the inside-document clustering nature of those tasks very well. The final results of the 2 participating teams are presented in Tables 5 and 6. Fine-grained sub results depending on the dataset configuration, e.g., the number of authors in a document and their contribution rate, are presented in the respective overview paper of this task [70].

[5] conforming to previous PAN events

3 Author Profiling

Author profiling distinguishes between classes of authors studying their sociolect aspect, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, native language, or personality type. Author profiling is a problem of growing importance in applications in forensics, security, and marketing. E.g., from a forensic linguistics perspective one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, the demographics of people that like or dislike their products. Pennebaker's [43] investigated how the style of writing is associated with personal attributes such as age, gender and personality traits, among others. In [5] the authors approached the task of gender identification on the British National Corpus and achieved approximately 80% accuracy. Similarly in [20] and [8] the authors investigated age and gender identification on formal texts. Recently most investigations focus on social media. For example, in [28] and [66] the authors investigated the style of writing in blogs. On the other hand, Zhang and Zhang [79] experimented with short segments of blog post and obtained 72.1% accuracy for gender prediction. Similarly, Nguyen et al. [41] studied the use of language and age among Dutch Twitter users. Since 2013 a shared task on author profiling has been organised at PAN [61,60,59]. It is worth mentioning the second order representation based on relationships between documents and profiles used by the best performing team of all editions [33,32,2]. Recently, the EmoGraph graph-based approach [57] tried to capture how users convey verbal emotions in the morphosyntactic structure of the discourse, obtaining competitive results with the best performing systems at PAN 2013 and demonstrating its robustness against genres and languages at PAN 2014 [58]. Moreover, the authors in [78] investigated on PAN-AP-2013 dataset a high variety of different features and showed the contribution of information retrieval based features in age and gender identification and in [35] the authors approached the task with 3 million features in a MapReduce configuration, obtaining high accuracies with fractions of processing time.

Tasks and Corpora In the Author Profiling task at PAN'16 participants approached the task of identifying age and gender from a cross-genre perspective in three different languages: English, Spanish and Dutch. English and Spanish partitions were labelled with age and gender. For labelling age, the following classes were considered: 18-24; 25-34; 35-49; 50+. Dutch partition was labelled only with gender. The dataset was split into training, early birds and test, as in previous editions. Training partition was collected from Twitter for the three languages. For English and Spanish, early birds partition was collected from social media and test partition from blogs. Both were compiled from PAN'14's dataset. In case of Dutch, both early birds and test partitions were collected from reviews. The number of authors per language and age class can be seen in Table 7. The corpus is balanced per gender but imbalanced per age.

For evaluation, the accuracy for age, gender and joint identification per language is calculated. Then, we average the results obtained per language (Eq. 1).

Table 7. Distribution of authors with respect to age classes per language. Dutch partition is labelled only with gender information. The corpus is balanced per gender.

	Training			Early birds			Test		
	EN	ES	NL	EN	ES	NL	EN	ES	NL
18-24	13	16		35	8		5	2	
25-34	68	64		46	10		12	6	
35-49	91	126		51	8		16	13	
50+	39	38		40	4		5	5	
Σ	211	244	192	172	30	25	38	26	250

$$\begin{aligned}\overline{gender} &= \frac{gender\_en + gender\_es + gender\_nl}{3}\\ \overline{age} &= \frac{age\_en + age\_es}{2}\\ \overline{joint} &= \frac{joint\_en + joint\_es}{2}\end{aligned} \tag{1}$$

The final ranking is calculated as the average of the previous values (Eq. 2):

$$ranking = \frac{\overline{gender} + \overline{age} + \overline{joint}}{3} \tag{2}$$

In summary, the Author Profiling shared task at PAN'16 focuses on the following aspects:

- **Age and gender identification:** As in previous editions, the task is to predict age and gender, and also the joint identification of age and gender for the same author.
- **Cross-genre evaluation:** The aim is at evaluating the performance of author profiling systems in a cross-genre setting. The training is provided in one genre (Twitter) and the evaluation is carried on another genre (social media, blogs or reviews).
- **Multilingual:** Participants are provided with data in English, Spanish and Dutch.

Results This year 22[6] have been the teams who participated in the shared task. In this section we show a summary of the obtained results. In Table 8 the overall performance per language and users' ranking are shown[7]. We can observe that in general accuracies in both English and Spanish datasets are similar, although the highest results were achieved in Spanish (42.86%). With respect to Dutch, were only the gender accuracy is shown, results are not much better than the random baseline (the highest value is equal

[6] In the four editions of the author profiling shared task we have had respectively 21 (2013: age and gender identification), 10 (2014: age and gender identification in different genre social media), 22 (2015: age and gender identification and personality recognition in Twitter) and 22 (2016: cross-genre age and gender identification) participating teams.

[7] The authors of waser16 team found an error in their implementation when performing cross validation

Table 8. Global ranking as average of each language joint accuracy. (*) Authors withdrew their participation due to a software error.

Ranking	Team	Global	English	Spanish	Dutch
1	Busger *et al.*	**0.5263**	**0.3846**	**0.4286**	0.5000
2	Modaresi *et al.*	0.4934	0.3205	**0.4286**	0.5040
3	Bilan *et al.*	0.4834	0.3333	0.3750	0.5500
4	Modaresi(a)	0.4602	0.3205	0.3036	0.5000
5	Markov *et al.*	0.4593	0.2949	0.3750	0.5100
6	Bougiatiotis & Krithara	0.4519	0.3974	0.2500	0.4160
7	Dichiu & Rancea	0.4425	0.2692	0.3214	0.5260
8	Devalkeneer	0.4387	0.3205	0.2968	0.5060
9	Waser *et al.**	0.4293	0.3205	0.2679	0.5320
10	Bayot & Gonçalves	0.4255	0.2179	0.3036	**0.5680**
11	Gencheva *et al.*	0.4015	0.2564	0.2500	0.5100
12	Agrawal & Gonçalves	0.3971	0.1923	0.2857	0.5080
13	Deneva	0.3880	0.2051	0.2679	0.4980
14	Kocher & Savoy	0.3800	0.2564	0.1964	0.5040
15	Roman-Gomez	0.3664	0.2821	0.1250	0.5620
16	Garciarena *et al.*	0.3660	0.1538	0.2500	0.5260
17	Zahid	0.3154	0.1923	0.2143	-
18	Aceituno	0.2949	0.1667	0.0893	0.5040
19	Poonguran	0.1793	-	-	0.5140
20	Ashraf *et al.*	0.1688	0.2564	-	-
21	Bakkar *et al.*	0.1560	0.2051	-	-
22	Pimas *et al.*	0.1410	0.1410	-	-

to 56.80%). It should be highlighted that this occurs even when the Dutch test set is the largest one. In Table 9 the best results per language and task are shown. A more in-depth analysis of the results and the different approaches can be found in [62].

Table 9. Best results per language and task.

	Age and Gender		
Language	*Joint*	Gender	Age
English	0.3974	0.7436	0.5513
Spanish	0.4286	0.7321	0.5179
Dutch	-	0.5680	-

4 Author Obfuscation

The development of author identification technology has reached a point at which it can be carefully applied in practice to resolve cases of unknown or disputed authorship. For a recent example, a state-of-the-art forensic software played a role in breaking the anonymity of J.K. Rowling who published her book "The Cuckoo's Calling" under the pseudonym Robert Gailbraith in order to "liberate" herself from the pressure of stardom, caused by her success with the Harry Potter series.[8] Moreover, forensic author identification software is part of the toolbox of forensic linguists, who employ it on a regular basis to support their testimony in court as expert witnesses in cases where the authenticity of a piece of writing is important. Despite their successful application, none of the existing approaches has been shown to work flawless, yet. All of them have a likelihood of returning false decisions under certain circumstances—but the circumstances under which they do are barely understood. It is particularly interesting if and how these circumstances can be controlled, since any form of control over the outcome of an author identification software bears the risk of misuse.

In fiction, a number of examples can be found where authors tried to remain anonymous, and where they, overtly or covertly, tried to imitate the writing style of others. In fact, style imitation is even a well-known learning technique in writing courses. But the question of whether humans are ultimately capable of controlling their own writing style so as to fool experts into believing they have not written a given piece of text, or even that someone else has, is difficult to answer based on observation alone: are the known cases more or less all there is, or are they just the tip of the iceberg (i.e., examples of unskilled attempts)? However, when the "expert" to be fooled is not a human but an author identification software, the rules are changed entirely. The fact that software is used to assist with author identification increases the attack surface of investigations to include any flaw in the decision-making process of the software. This is troublesome since the human operator of such a software may be ignorant of its flaws, and biased toward taking the software's output at face value instead of treating it with caution. After all, being convinced of the quality of a software is a necessary precondition to employing it to solve a problem.

At PAN 2016, we organize for the first time a pilot task on author obfuscation to begin exploring the potential vulnerabilities of author identification technology. A number of interesting sub-tasks related to author obfuscation can be identified, from which we have selected that of author masking. This task complements, and is built on top of the task of authorship verification, a sub-task of author identification, which was organized at PAN 2013 through PAN 2015 [26,71,72]:

Authorship verification:
Given two documents,
decide whether they have been
written by the same author.

vs.

Author masking:
Given two documents by the same author,
paraphrase the designated one so that the
author cannot be verified anymore.

The two tasks are diametrically opposed to each other: the success of a certain approach for one of these tasks depends on its "immunity" against the most effective approach for the other. The two tasks are also entangled, since the development of a new approach for one of them should build upon the capabilities of existing approaches for the other. However, compared to authorship verification, author obfuscation in general, and author masking in particular received little attention to date.[9] A reason for this may be rooted in the fact that author masking requires (automatic) paraphrasing as a subtask, which poses a high barrier of entry to newcomers.

[8] http://languagelog.ldc.upenn.edu/nll/?p=5315

[9] An overview of related work can be found in the full task overview paper [51].

Table 10. Average performance drops in terms of "final scores" of the authorship verifiers submitted at PAN 2013 to PAN 2015 when run on obfuscated versions of the corresponding test datasets as per the submitted obfuscators.

Participant	PAN 2013	PAN 2014 EE	PAN 2014 EN	PAN 2015
Mihaylova *et al.*	-0.10	-0.13	-0.16	-0.11
Keswani *et al.*	-0.09	-0.11	-0.12	-0.06
Mansoorizadeh *et al.*	-0.05	-0.04	-0.03	-0.04

Notwithstanding the task's inherent challenges, 3 teams successfully submitted an approach. Keswani *et al.* [27] employ circular translation as a means of obfuscation, where the to-be-obfuscated text is translated to another language, and the resulting translation again, and so on, until, as a final step, the last translation goes back to the initial language. The presumption is that the original text will be sufficiently changed to obfuscate its author. Mansoorizadeh *et al.* [36] attack the feature of (stop) word frequencies on which many verification approaches are based and exchange the most frequent words in the to-be-obfuscated text with synonyms, carefully chosen not to distort the originals meaning. Mihaylova *et al.* [38] take a more "writing engineering"-based approach: it targets a number of style-indicating features that are frequently used within author identification approaches and tries to attack them by transforming the to-be-obfuscated text with certain rule-based and random text operations.

The performance of an author identification approach rests with its capability to achieve its goal of fooling a given expert, be it a software or a human. In this regard, we call an obfuscation software

- **safe**, if a forensic analysis does not reveal the original author of its obfuscated texts,
- **sound**, if its obfuscated texts are textually entailed by their originals, and
- **sensible**, if its obfuscated texts are well-formed and inconspicuous.

These dimensions are orthogonal; an obfuscation software may meet all of them to various degrees of perfection. However, achieving perfection in all three dimensions may not be necessary for practical applicability: for instance, if the fact that a text has been obfuscated is obvious, there may not be a problem as long as the changes made cannot be easily undone. To operationalize the three dimensions, we employ state-of-the-art automatic authorship verifiers to measure safety, and manual peer-review to assess soundness and sensibleness. Regarding safety, we measure the impact of author obfuscation on classification accuracy, whereas soundness and sensibleness are manually assessed on a Likert scale by multiple reviewers. In this connection, we also invite participants to conduct their own evaluation with regard to the aforementioned dimensions, giving them access to each other's obfuscations, thus crowdsourcing further ideas at evaluating author obfuscation approaches. As an evaluation dataset we employ the joint training datasets and the joint test datasets that were used for the authorship verification tasks of PAN 2013 to PAN 2015. This ensures compatibility between tasks and allows us to study the impact of the 3 author obfuscation approaches on the authorship verifiers submitted to the authorship verification tasks.

Regarding safety, Table 10 shows averaged performance drops when running the authorship verifiers submitted to PAN 2013 to PAN 2015 on obfuscated versions of the corresponding test datasets when compared to their performance on the original test datasets: the average performances drop significantly for each pair of obfuscator and year. The top average performance drop of -0.16 “final score” (i.e., the combination of AUC and C1) on the PAN 2014 test dataset comprising English novels has been achieved by the obfuscator of Mihaylova *et al.* [38]. The order of obfuscators by average performance drop remain stable across years, whereas the achieved drops differ based on the different test datasets. This result shows that the authorship verifiers are to some extent vulnerable to obfuscation. Regarding soundness and sensibleness of the texts, however, the texts produced by the safest obfuscator are less than ideal (i.e., while the original text’s message can be partly inferred from the obfuscated text, many grammar mistakes are introduced). It is encouraging, though, that the obfuscated texts of Mihaylova *et al.*’s obfuscator achieve better soundness and sensibleness compared to the cyclic translations produced by Keswani *et al.*’s obfuscator. An in-depth analysis of the performances can be found in the full-length task overview paper [51]. This also includes a review of the results of peer-evaluation, where participants evaluated the runs of all obfuscators in anonymized form (knowing of course which of the runs was produced by their own obfuscator), and whether they are in line with our own evaluation results. Two of the submitted peer-evaluations were submitted by external reviewers who did not submit an obfuscator of their own.

5 Conclusions

PAN 2016 evaluation lab at CLEF attracted a high number of teams from all around the world. This demonstrates that the shared tasks on author identification, profile and obfuscation are of particular interest for researchers. New corpora have been developed covering multiple languages (English, Spanish, Greek, Dutch). These new resources will help fostering research in digital text forensics and future techniques will be able to be compared with the evaluation results obtained by the participating teams in the three shared tasks. The author obfuscation shared task will allow to shed light on the robusteness of state-of-the-art author identification and author profiling techniques against author obfuscation technology.

For the first time since 2009 a shared task on external plagiarism detection has not been organized at PAN/CLEF. A shared tasks on plagiarism detection will be organized at PAN/FIRE instead: after addressing previously text reuse in source code, at monolingual [13] and cross-language [14] levels, and plagiarirms in Arabic texts [7], this year the focus of the plagiarism detection task will be on texts written in Farsi[10]. Moreover, with respect to author profiling, a PAN/FIRE task on personality recognition in source code will be organized[11].

Acknowledgements

We thank the organizing committees of PAN’s shared tasks Ben Verhoeven, Walter Daelemans, Patrick Juola. Our special thanks go to all of PAN’s participants, to Adobe[12]

[10] http://ictrc.ac.ir/plagdet/

[11] http://www.autoritas.es/prsoco

[12] http://www.adobe.com

and to MeaningCloud[13] for sponsoring the author profiling shared task award. The work of the first author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030). The work of the second author was partially supported by Autoritas Consulting and by Ministerio de Economía y Competitividad de España under grant ECOPORTUNITY IPT-2012-1220-430000.

References

1. Almishari, M., Tsudik, G.: Exploring Linkability of User Reviews. In: Computer Security, ESORICS pp. 307–324 (2012)
2. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-Y-Gómez, M., Villaseñor-Pineda, L., Jair-Escalante, H.: INAOE's Participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol. 1391 (2015)
3. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A Comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints. In: Information Retrieval 12 (4) pp.461–486 (2009)
4. Argamon, S., Juola, P.: Overview of the International Authorship Identification Competition at PAN-2011, In: Working Notes Papers of the CLEF 2011 Evaluation Labs (2011)
5. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, Genre, and Writing Style in Formal Written Texts. In: TEXT 23, 321–346 (2003)
6. Bagnall, D.: Author Identification Using Multi-headed Recurrent Neural Networks. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol.1391 (2015)
7. Bensalem I., Boukhalfa I., Rosso P., Abouenour L., Darwish K., Chikhi S.: Overview of the AraPlagDet PAN@ FIRE2015 Shared Task on Arabic Plagiarism Detection. In: Notebook Papers of FIRE 2015. CEUR-WS.org, vol. 1587 (2015)
8. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proceedings of EMNLP '11. (2011)
9. Burrows, S., Potthast, M., Stein, B.: Paraphrase Acquisition via Crowdsourcing and Machine Learning. In: ACM TIST 4(3), 43:1–43:21 (2013)
10. Castillo, E., Cervantes, O., Vilariño, D., Pinto, D., León, S.: Unsupervised Method for the Authorship Identification Task. In: CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org vol. 1180 (2014)
11. Chaski, C.E.: Who's at the Keyboard: Authorship Attribution in Digital Evidence Invesigations. In: International Journal of Digital Evidence 4 (2005)
12. Clarke, C. L., Craswell, N., Soboroff, I., Voorhees, E. M. Overview of the TREC 2009 Web Track. In: DTIC Document (2009)
13. Flores E., Rosso P., Moreno L., Villatoro E.: On the Detection of SOurce COde Re-use. In: ACM FIRE'14 Post Proceedings of the Forum for Information Retrieval Evaluation, pp 21-30 (2015)
14. Flores E., Rosso P., Villatoro E., Moreno L., Alcover R., Chirivella V.: PAN@FIRE: Overview of CL-SOCO Track on the Detection of Cross-Language SOurce COde Re-use. In: Notebook Papers of FIRE 2015. CEUR-WS.org, vol. 1587 (2015)
15. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: UJM at Clef in Author Identification. In: CLEF 2014 Labs and Workshops, Notebook Papers. CEUR-WS.org vol.1180 (2014)

[13] http://www.meaningcloud.com/

16. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Recent Trends in Digital Text Forensics and its Evaluation. In: Proceedings of CLEF 2013. Springer-Verlag, LNCS(8138), pp. 53–58 (2013)
17. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Proceedings of SIGIR 12. ACM (2012)
18. Hagen, M., Potthast, M., Stein, B.: Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol. 1391 (2015)
19. van Halteren, H.: Linguistic Profiling for Author Recognition and Verification. In: Proceedings of ACL 04. (2004)
20. Holmes, J., Meyerhoff, M.: The Handbook of Language and Gender. In: Blackwell Handbooks in Linguistics, Wiley (2003)
21. Iqbal, F., Binsalleeh, H., Fung, B. C. M., Debbabi, M.: Mining Writeprints from Anonymous e-Mails for Forensic Investigation. In: Digital Investigation 7(1-2) pp. 56–64 (2010)
22. Jankowska, M., Keselj, V., Milios, E.: CNG Text Classification for Authorship Profiling Task—Notebook for PAN at CLEF 2013. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CEUR-WS.org vol. 1179 (2013)
23. Juola, P.: An Overview of the Traditional Authorship Attribution Subtask. In: Working Notes Papers of the CLEF 2012 Evaluation Labs (2012)
24. Juola, P.: Authorship Attribution. In: Foundations and Trends in Information Retrieval 1, 234–334 (2008)
25. Juola, P.: How a Computer Program Helped Reveal J.K. Rowling as Author of A Cuckoo's Calling. In: Scientific American (2013)
26. Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN-2013. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CEUR-WS.org vol. 1179 (2013)
27. Keswani, Y., Trivedi, H., Mehta, P., Majumder, P.: Author Masking through Translation—Notebook for PAN at CLEF 2016. In: Conference and Labs of the Evaluation Forum, CLEF (2016)
28. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. In: Literary and Linguistic Computing 17(4) (2002)
29. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring Differentiability: Unmasking Pseudonymous Authors. In: J. Mach. Learn. Res. 8, 1261–1276 (2007)
30. Koppel, M., Winter, Y.: Determining if Two Documents are Written by the same Author. In: Journal of the American Society for Information Science and Technology 65(1), 178–187 (2014)
31. Layton, R., Watters, P., Dazeley, R.: Automated Unsupervised Authorship Analysis Using Evidence Accumulation Clustering. In: Natural Language Engineering 19(1) pp. 95–120 (2013)
32. López-Monroy, A.P., Montes-y Gómez, M., Jair-Escalante, H., Villasenor-Pineda, L.V.: Using Intra-Profile Information for Author Profiling—Notebook for PAN at CLEF 2014. In: Working Notes Papers of the CLEF 2014 Evaluation Labs. CEUR-WS.org vol. 1180 (2014)
33. López-Monroy, A.P., Montes-y Gómez, M., Jair-Escalante, H., Villasenor-Pineda, L., Villatoro-Tello, E.: INAOE's Participation at PAN'13: Author Profiling Task—Notebook for PAN at CLEF 2013. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CEUR-WS.org vol. 1179 (2013)
34. Luyckx, K., Daelemans, W.: Authorship Attribution and Verification with Many Authors and Limited Data. In: Proceedings of COLING (2008)
35. Maharjan, S., Shrestha, P., Solorio, T., Hasan, R.: A Straightforward Author Profiling Approach in MapReduce. In: Advances in Artificial Intelligence. Iberamia. (2014)

36. Mansoorizadeh, M.: Submission to the Author Obfuscation Task at PAN 2016. In: Conference and Labs of the Evaluation Forum, CLEF (2016)
37. Meyer zu Eißen, S., Stein, B.: Intrinsic Plagiarism Detection. In: Proceedings of ECIR 06. Springer-Verlag, LNCS(3936), pp. 565–569 (2006)
38. Mihaylova, T., Karadjov, G., Nakov, P., Kiprov, Y., Georgiev, G., Koychev, I.: SU@PAN'2016: Author Obfuscation—Notebook for PAN at CLEF 2016. In: Conference and Labs of the Evaluation Forum, CLEF (2016)
39. Miro, X. A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., Speaker Diarization: A Review of Recent Research. In: Audio, Speech, and Language Processing, IEEE Transactions on 20 (2) pp.356–370 (2012)
40. Moreau, E., Jayapal, A., Lynch, G., Vogel, C.: Author Verification: Basic Stacked Generalization Applied to Predictions from a set of Heterogeneous Learners. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol. 1391 (2015)
41. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How old do you think I am?"; A Study of Language and Age in Twitter. In: Proceedings of ICWSM 13. AAAI. (2013)
42. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In: Proceedings of HLT '11 (2011)
43. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological Aspects of Natural Language Use: Our Words, Our Selves. In: Annual Review of Psychology 54(1), 547–577 (2003)
44. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Working Notes Papers of the CLEF 2010 Evaluation Labs (2010)
45. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-Language Plagiarism Detection. In: Language Resources and Evaluation (LREC) 45, 45–62 (2011)
46. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Working Notes Papers of the CLEF 2011 Evaluation Labs (2011)
47. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: Working Notes Papers of the CLEF 2012 Evaluation Labs (2012)
48. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CEUR-WS.org vol. 1179 (2013)
49. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Proceedings of CLEF 14. Springer-Verlag, LNCS(8685), pp. 268–299 (2014)
50. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection. In: Working Notes Papers of the CLEF 2014 Evaluation Labs. CEUR-WS.org vol. 1180 (2014)
51. Potthast, M., Hagen, M., Stein, B.: Author Obfuscation: Attacking the State of the Art in Authorship Verification. In: CLEF 2016 Working Notes. CEUR-WS.org (2016)
52. Potthast, M., Göring, S., Rosso, P., Stein, B.: Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol. 1391 (2015)
53. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Proceedings of SIGIR 12. ACM. (2012)

54. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Proceedings of ACL 13. ACL. (2013)
55. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Proceedings of COLING 10. ACL. (2010)
56. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Proceedings of PAN at SEPLN 09. CEUR-WS.org 502 (2009)
57. Rangel, F., Rosso, P.: On the Impact of Emotions on Author Profiling. In: Information Processing & Management, Special Issue on Emotion and Sentiment in Social and Expressive Media 52(1) pp. 73–92 (2016)
58. Rangel, F., Rosso, P.: On the Multilingual and Genre Robustness of EmoGraphs for Author Profiling in Social Media. In: Experimental IR Meets Multilinguality, Multimodality and Interaction, CLEF pp. 274–280 (2015)
59. Rangel, F., Rosso, P., Celli, F., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol. 1391 (2015)
60. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: Working Notes Papers of the CLEF 2014 Evaluation Labs. CEUR-WS.org vol. 1180 (2014)
61. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013—Notebook for PAN at CLEF 2013. In: Working Notes Papers of the CLEF 2013 Evaluation Labs. CEUR-WS.org vol. 1179 (2013)
62. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations. In: CLEF 2016 Working Notes. CEUR-WS.org (2016)
63. Samdani, R., Chang, K., Roth, D.: A Discriminative Latent Variable Model for Online Clustering. In: Proceedings of The 31st International Conference on Machine Learning pp.1–9 (2014)
64. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all Character N-grams are Created Equal: A Study in Authorship Attribution. In: Proceedings of NAACL 15. ACL. (2015)
65. Sapkota, U., Solorio, T., Montes-y-Gómez, M., Bethard, S., Rosso, P.: Cross-topic Authorship Attribution: Will Out-of-topic Data Help? In: Proceedings of COLING 14. (2014)
66. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. AAAI (2006)
67. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. In: PloS one 8(9), 773–791 (2013)
68. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. In: Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
69. Stamatatos, E.: On the Robustness of Authorship Attribution Based on Character N-gram Features. In: Journal of Law and Policy 21, 421–439 (2013)
70. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by Authorship Within and Across Documents. In: CLEF 2016 Working Notes. CEUR-WS.org (2016)
71. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the Author Identification Task at PAN-2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR-WS.org vol. 1391 (2015)

72. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the Author Identification Task at PAN 2014. In: Working Notes Papers of the CLEF 2014 Evaluation Labs. CEUR-WS.org vol. 1180 (2014)
73. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author. In: Comput. Linguist. 26(4), 471–495 (2000)
74. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic Plagiarism Analysis. In: Language Resources and Evaluation (LRE) 45, 63–82 (2011)
75. Stein, B., Meyer zu Eißen, S.: Near Similarity Search and Plagiarism Analysis. In: Proceedings of GFKL 05. Springer, pp. 430–437 (2006)
76. Verhoeven, B., Daelemans, W.: Clips Stylometry Investigation (CSI) Corpus: A Dutch Corpus for the Detection of Age, Gender, Personality, Sentiment and Deception in Text. In: Proceedings of LREC 2014 (2014)
77. Verhoeven, B., Daelemans, W.: CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC (2014)
78. Weren, E., Kauer, A., Mizusaki, L., Moreira, V., de Oliveira, P., Wives, L.: Examining Multiple Features for Author Profiling In: Journal of Information and Data Management 5(3) pp. 266–280 (2014)
79. Zhang, C., Zhang, P.: Predicting Gender from Blog Posts. Technical Report. University of Massachusetts Amherst, USA (2010)