Using Profiling Techniques to Protect the User's Privacy in Twitter

Alexandre Viejo, David Sánchez, and Jordi Castellà-Roca

Departament d'Enginyeria Informàtica i Matemàtiques,
UNESCO Chair in Data Privacy, Universitat Rovira i Virgili
Av. Països Catalans 26, E-43007 Tarragona, Spain
{alexandre.viejo,david.sanchez,jordi.castella}@urv.cat

Abstract. The emergence of microblogging-based social networks shows how important it is for common people to share information worldwide. In this environment, Twitter has set it apart from the rest of competitors. Users publish text messages containing opinions and information about a wide range of topics, including personal ones. Previous works have shown that these publications can be analyzed to extract useful information for the society but also to characterize the users who generate them and, hence, to build personal profiles. This latter situation poses a serious threat to users' privacy. In this paper, we present a new privacy-preserving scheme that distorts the real user profile in front of automatic profiling systems applied to Twitter. This is done while keeping user publications intact in order to interfere the least with her followers. The method has been tested using Twitter publications gathered from renowned users, showing that it effectively obfuscates users' profiles.

Keywords: Microblogging, Noise addition, Privacy, Profiling, Twitter.

1 Introduction

Twitter is a very popular online social network and microblogging system. Basically, this social tool allows registered users to share short text-based posts (named *tweets*) up to 140 characters with anybody else on the Internet. Nowadays, the company behind this social network claims to have 100 million active users who generate 230 million tweets on average per day [1]. This fact proves two points: (i) how sharing information has become a real necessity for common people; and (ii) the relevance of Twitter in the present day.

Nevertheless, tweets generally contain personal information [2] and this fact motivates the existence of systems that analyze those publications and build *user profiles*. This implies that profiling data such as user preferences can be linked with her identity. This may invite malicious attacks from the cyberspace (*e.g.*; personalized spamming, phishing, etc) and even from the real world (*e.g.,* stalking) [4].

Several profiling mechanisms that gather information from Twitter (and other Internet services) can be found in the literature [5–10]. In general, all these tools

V. Torra et al. (Eds.): MDAI 2012, LNAI 7647, pp. 161–172, 2012.

generate *user profiles* that contain the interests of the users. In order to achieve that, they generally use a knowledge base (*e.g.,* Wikipedia [11], news repositories, the Web) to semantically interpret message contents and the number of term occurrences/co-occurrences to calculate the weight of each topic of interest (*e.g.,* sports, technology, health, etc), which is found in the tweets of a certain user. As an example of this process, the authors in [8] present a knowledge-based framework that builds user profiles from text messages shared in social platforms. Researchers [9, 10] have also noted the difficulty of analyzing textual contents in Twitter due to the use of short text-based messages. Tweets are, quite commonly, ungrammatical and noisy (due to hashtags, re-tweets, URLs). Hence, due to the difficulty of applying exhaustive syntactical analyses, Twitter profiling systems [7–10] usually analyse the whole user tweets as a bag-of-words.

In order to prevent privacy problems that are inherent to the existence of *profilers* and *user profiles*, it is convenient to design privacy preserving tools that allow users to protect themselves from profilers. Nevertheless, it is worth to mention that privacy-preserving methods based on restricting the visibility of the user-generated content compromise as well the capability of users to gain attention from others. Indeed, since this is one of the main motivations of using Twitter, this straightforward approach may not be widely adopted. Therefore, a successful privacy-preserving scheme should protect the privacy without limiting the visibility of user-generated data.

1.1 Contribution and Plan of This Paper

In this paper we propose a new scheme designed to prevent automatic text-based data extraction techniques from profiling users (*i.e.,* to discover dominant profile categories) of microblogging services. The basic idea is to introduce a set of fake messages within the user account while maintaining users' messages intact. We have used Twitter to test our proposal but the results achieved can be applied to any other social platform which works with text-based messages.

Section 2 discusses previous work aiming at providing privacy to microblogging services and introduces the basis of our approach. Section 3 presents our privacy-preserving method, detailing how it distorts the user profile (in order to hide it) as new messages are published. Section 4 evaluates the proposal, applying it to several well-differentiated Twitter users. The final section presents the conclusions and some lines of future research.

2 Previous Work

In the literature there is an important lack of mechanisms that try to preserve the privacy of the users of Twitter and similar platforms.

In [12], the authors present a Firefox extension that allows users to specify which data or activity need to be kept private. Sensitive data is substituted with fake one, while the real data is stored in a third party server that can

be only accessed by the allowed users. Note that this solution requires a centralized infrastructure which must be honest and always available. Clearly, this requirement is a very strong shortcoming of this proposal.

This behaviour avoids profiling but it also jeopardizes the capability of the users to gain attention from others. As explained previously getting attention (which in turn provides publicity, vanity and ego gratification) is the main motivation for Twitter users (and other similar platforms) [3].

In order to avoid this major problem, we propose a method which aims to distort user profiles while keeping their original messages intact. In order to do so, we studied techniques proposed in the research field of *unstructured text anonymization* which aim to hide original information while not completely removing/transforming it. In [13], authors explain three different methods to deal with the anonymization of textual documents: (i) *Named entity generalization*: sensitive entities can be generalized (*e.g.,* iPhone → cell phone) to achieve some degree of privacy while preserving some of their semantic meaning in the document [14]; (ii) *Entity swapping*: this method is based on swapping relatively similar entities [15] between documents of the same set, or within the same document depending on the concrete case; and (iii) *Entity noise addition*: this method introduces new entities to user documents that can help to hide the original information.

The main problem of the first method is that it may introduce significant loss of information in user messages derived from the degree of generalization introduced. The second technique is designed to work with documents of an adequate length and which are properly structured. Clearly, tweets do not hold these requirements. The last method is the more suitable for short-length documents and, hence, it might be applied to Twitter and similar microblogging services. The main shortcoming is that, if noise is added within user posts, it may generate uncomfortable publications like the *Entity swapping* technique. Nevertheless, it might be successful if it is implemented allowing readers to distinguish between legitimate and distorted publications, which is precisely the goal of our method.

3 Our Proposal

The proposed method follows two main steps: *user profiling* and *semantic noise addition*. In the first one, our method uses similar techniques as those proposed by profiler systems to characterize the user profile according to her published tweets. In the second one, it assesses which are the dominant and dominated categories in her profile and which should be the contents of the fake tweets to be added in the user account (*i.e.,* distortion) to achieve a balanced profile.

It is important to note that fake tweets are constructed as a concatenation of terms which lack the semantically-coherent discourse that a human reader would expect. In this manner, human readers could easily discern between user tweets and fake ones while general approaches for profiling users would be unable to do it. Obviously, it is always possible to design an ad-hoc solution trained to discern between user messages and fake ones. Nevertheless, developing an ad-hoc system

is costly and, in turn, it might be defeated by modifying certain aspects of the way in which fake messages are generated.

In the following two sections, our method is formalized and described in detail.

3.1 User Profiling

Let us consider a user u that starts posting a first tweet t_1 in her Twitter account. Let us also consider that the user profile P is defined (like in related works [5] and [6]) as a set of well-defined categories $C = \{c_1, \ldots, c_k\}$ (*e.g.,* science, health, society, sports, etc), for which their relative weight (v_i) should be computed according to the amount of evidences found, for each category, in user tweets. Hence, after the i-th tweet t_i, the user profile (P_i) will be characterized according to the set of weighted profile categories obtained from analyzing all tweets from t_1 to t_i: $P_i = \{< c_1, v_1 >, \ldots, < c_k, v_k >\}$.

At the beginning of its execution, our system has no idea about the initial profile of the user P_0. Therefore, the profile is initialized as $P_0 = \{< c_1, 0 > , \ldots, < c_k, 0 >\}$. For each new user tweet t_i published by u, the system progressively computes and updates P_i, mimicking the common way to build user profiles of related works.

First, since tweets are slightly-grammatical short texts which are difficult to analyze [8, 10] syntactically and semantically, we opted, as done in some related works [8, 9], to implement a shallow linguistic parsing which, instead of trying to analyze well-formed sentences, focuses on extracting pieces of text with semantic content that can contribute to characterize the user profile: *noun-phrases (NPs).*

NPs are built around a noun whose semantics can be refined by adding new nouns or adjectives (*e.g., iPhone* $\rightarrow$ *new iPhone*). Each NP either refers to a generic concept (*e.g., water sports*) or it can be considered a proper noun that is an instance of a concept (*e.g., iPhone* is an instance of a *cell-phone*).

Accordingly, **the first step** of the profiling process consist in extracting NPs from user tweets. To achieve that, we rely on several commonly-used natural language processing tools[1]: *sentence detection*, *tokenization*, *part-of-speech tagging* and *syntactic parsing* (*i.e.,* chunking). As a result of this analysis, prepositional or nominal phrases are detected. From these, only NPs (which are those with semantic content) are extracted [16]. It is worth to mention that this process also removes superfluous elements like misspelled words, abbreviations, emoticons, etc.

As output of a tweet t, the set $M = \{< NP_1, w_1 >, \ldots, < NP_p, w_p >\}$ is obtained, in which NP_i is each Noun Phrase extracted from t and w_i is its number of appearances.

The **second step** of the profiling process consists, as done in related works [8, 9], in semantically analyzing extracted NPs in order to classify them, if possible, according to the defined categories C. By doing this, the number of NPs corresponding to each category can be evaluated to characterize the user profile in a later step.

[1] OpenNLP Maxent Package: `http://maxent.sourceforge.net/about.html`

To enable this classification from a semantic point of view (*i.e.,* to associate each NP to its conceptual abstraction), we rely on a predefined knowledge base [17]. This knowledge base can be a taxonomy, folksonomy or a more formal ontology [20] as long as it offers a structured conceptualization of one or several knowledge domains expressed by, at least taxonomic relationships. In order to improve the recall of the semantic analysis and due to the proliferation of proper nouns in user tweets, a large base that potentially includes up-to-date Named Entities (NEs) [18, 19] is desired.

In our work, we rely on the *Open Directory Project (ODP)* hierarchy of categories because it is the largest, most comprehensive human-edited directory of the Web, constructed and maintained by a vast community of volunteer editors [21]. The purpose of ODP is to list and categorize web sites. Manually created categories are taxonomically structured and populated with related web resources. Nowadays, it classifies almost 5 million web sites in more than 1 million categories (considering also up-to-date Named Entities).

In order to semantically classify NPs in M, the system queries each NP_i to ODP. If found, ODP returns the most likely hierarchy H_i of categories ($H_i = h_{i,1} \rightarrow \ldots \rightarrow h_{i,l}$) to which NP_i belongs. For example, if the system queries the NP *"iPhone"*, ODP returns: *iPhone* $\rightarrow$ *Smartphones* $\rightarrow$ *Handhelds* $\rightarrow$ *Systems* $\rightarrow$ *Computers*.

If NP_i is not found in ODP, the system tries with simpler forms of the NP by removing adjectives/nouns starting from the one most on the left (*e.g., "new iPhone"* $\rightarrow$ *"iPhone"*) to improve the recall while maintaining the core semantics. The fact that NPs incorporate qualifiers is quite common in texts, but these are hardly covered in knowledge structures which try to model them in a generic way.

The **third and final step** of the profiling process applied to the first tweet t_1 consists in updating the user profile P_1 according to the categories to which extracted NPs belong. Concretely, for each NP_i, ODP has retrieved a hierarchy H_i of categories, hence, the system checks if any of the profile categories (c_j in C) is included in H_i. In the affirmative case, the system states that NP_i is a taxonomical specialization of c_j (*i.e.,* NP_i *is-a* c_j) and it adds the contribution of NP_i to c_j by adding the amount of occurrences of NP_i (w_i) to the category weight (*i.e.,* v_j of c_j). As more NPs found to be a taxonomical specialization of c_j are considered, the weight v_j of c_j is incremented accordingly, as follows:

$$v_j = v_j + \sum_{\forall NP_i \ is-a \ c_j} w_i \tag{1}$$

Once all NPs are considered, the user profile P_1 corresponding to the first tweet t_1 is defined by a ranked list of categories, according to their weights: $P_1 = \{< c_1, v_1 >, \ldots, < c_k, v_k >\}$, where v_j states the sum of contributions according to the number of term occurrences/co-occurrences related to each particular profile category c_j.

3.2 Semantic Noise Addition

After the publication of t_1 and the characterization of the user profile P_1, the objective of our system is to introduce additional terms in the user account as a new fake tweet ft that will balance the user profile towards a uniformly distributed one (according to the considered categories), while maintaining the original tweets unaltered. In this manner, dominant profile categories will become indistinguishable for a profiling system.

In the **first step** of the semantic noise addition process, the system uses the user profile P_1 constructed after the publication of t_1 to analyze the set of weighted categories and selects the one with the maximum weight (*i.e.,* $< c_{MAX}, v_{MAX} >= argmax_{\forall < c_i, v_i > \in P_1}(v_i)$). Then, for the rest of the categories c_j in P_1, it computes the difference with respect to the maximum one ($\Delta(< c_j, v_j >; P_1)$), as follows:

$$\Delta(< c_j, v_j >; P_1) = v_{MAX} - v_j \tag{2}$$

This difference quantifies the number of term occurrences/co-occurrences that are needed to balance each non-maximal category c_j with respect to the dominant one c_{MAX}.

In the **second step** of the semantic noise addition process, for all non-maximal categories, and starting from the c_j for which its Δ is the largest (*i.e.,* the one with the least dominance in the user profile), the system randomly retrieves $\Delta(< c_j, v_j >; P_1)$ terms from the ODP hierarchy under the corresponding category c_j.

In the **third and final step** of this process, retrieved subcategories for all non-maximal categories are then put together in the form of a new fake tweet ft_1 to be published after the user tweet t_1. This represents the semantically correlated noise added to balance the user profile.

Note that, due to limitations imposed by Twitter regarding message lengths (a maximum of 140 characters), the number of terms to be added in order to fake tweets ft should fulfill this restriction. Hence, even though a certain number of terms should be added to obtain a -theoretic- perfectly balance user profile, in practice, that number could be lower to fulfill Twitter restriction. The fact that a lower amount of fake terms are allowed to be added will cause a slower balancing of the user profile, as it will be shown in the evaluation section. As a general rule, considering that the average length of terms in ODP is 8 characters, up to 15 terms (counting separator whitespaces) could be fitted, in average, in each ft.

Also note that, since fake tweets are raw lists of terms of different domains put together without a narrative thread, human readers would easily distinguish them from those created by the original user. On the contrary, an automatic profiling based on term distribution is assumed to fail when characterizing the user profile, due to the added *semantic noise*.

The **whole compound process** (profiling+noise addition) is iteratively executed as new tweets are posted by the user. Concretely, for the i-th legitimate

tweet, t_i, the profile characterization P_i will reflect the aggregation of all previous ones (*i.e.,* both legitimate, from t_1 to t_i, and fake ones, from ft_1 to $ft_i - 1$). Note that each new tweet (both legitimate or fake) contributes to update category weights, increasing previous values according to new extracted category terms. As a result of P_i profiling, the system will create a new fake tweet ft_i that balances the characterization of P_i. As new fake tweets are added, the user profile will tend to balance, while the system adapts its behavior (*i.e.,* semantic noise addition) to the new user messages. Reaction time for profile balancing will depend on the amount of noise required to be added (that would depend on the homogeneity of user messages according to the computed profile) and on the maximum number of terms allowed to be published (according to the Twitter length limitation of messages). The fact that the system dynamically re-computes the user profile after each new tweet, enables our proposal to adapt to changes in user preferences or topics of interest, considering also the past history.

4 Evaluation

In this section, we evaluate the performance of the proposed system in balancing and, hence, hiding Twitter user profiles.

As evaluation data, we took eight well-differentiated Twitter users, whose profiles should correspond to eight root categories in the ODP hierarchy, as shown in Table 1. To select individual users, we used the *WhoToFollow* [22] search engine provided by Twitter. It provides a list of the most relevant Twitter users according to a specific topic. For each profile category (taken from ODP), we took the most relevant Twitter user as indicated by *WhoToFollow* for the corresponding topic. These are also shown in Table 1. For each user, we took the 100 most recently published tweets as evaluation data.

To numerically quantify the degree of balancing, θ, of a user profile P after each published tweet (both legitimate or fake), the proposed system sums the differences Δ in the number of occurrence v_j for each category c_j in P with respect to the maximum one, c_{MAX} (see Section 3.2). Then, the result is normalized by the total number of occurrences needed to balance a profile with respect to c_{MAX} in the worst case (*i.e.,* when the contribution of the other non-maximal categories is zero). The normalizing factor corresponds to the product of the number of non-maximal categories in P, this is $|P| - 1$, by the number of occurrences of c_{MAX}, that is v_{MAX}.

$$\theta(P) = \frac{\sum_{\forall < c_j, v_j > \in P} (\Delta(< c_j, v_j >; P))}{v_{MAX} \times (|P| - 1)} \tag{3}$$

The numerical interval of θ goes from 0 to 1, where 0 means a perfectly balanced profile (*i.e.,* zero difference between all non-maximal categories with respect to the most dominant one) and 1 means maximal difference (*i.e.,* the contribution of all categories except the maximum one is zero).

Table 1. ODP Categories, corresponding WhoToFollow topics, most relevant Twitter users for each one and user description (last accessed: January 22th, 2012)

ODP category	WhoToFollow topic	Twitter user	User description
Arts	Arts and Design	@johnmaeda	President of the Rhode Island School of Design
Business	Business	@businessinsider	The latest business news and analysis
Computers	Technology	@guardiantech	News from the Guardian tech team
Health	Health	@CDC_eHealth	Center for Disease Control, USA
Science	Science	@ReutersScience	Science by Reuters.com
Shopping	Fashion	@glamour_fashion	Glamour magazine's fashion team
Society	News	@nytimes	The New York Times
Sports	Sports	@espn	Sports news

Due to the lack of related works which propose method of profile distortion methods for Twitter (and microblogging services in general), we compare our method with the original data (*i.e.,* no fake tweets are added) and with a naive distortion method which adds a fixed amount of *random noise* (*i.e.,* a number of random terms taken from ODP) per each fake tweet. In this last case, the semantics associated to user tweets are not considered in the construction of fake tweets.

Finally, in order to quantify the influence of the amount of noise added per fake tweet on our method, we have fixed several upper bounds. In the most favorable setting (*high semantic noise*), we allowed up to 15 terms to be added per fake tweet, which corresponds, in average, to the maximum amount allowed by Twitter (see Section 3.2). In the intermediate situation (*medium semantic noise*), we allowed up to 7 terms per fake tweet. The most constrained scenario (*low semantic noise*) only allowed up to 3 terms per fake tweet. Note that, for the *random* approach, the amount of added noise is constant and fixed to *high* (*i.e.,* 15 terms per fake tweet).

Figure 1 shows profile balancing results according to the number of user tweets analyzed (up to 100) for the different methods and scenarios. Note that the horizontal scale quantifies the number of analyzed *user tweets* including also, in the case of the *semantic* and *random noise* addition methods, the corresponding fake ones.

Several conclusions can be drawn from the analysis of the graphs. First, as expected, the addition of semantic noise results in the best profile balancing (*i.e.,* it is closer to zero) because the distribution of category terms at the i-th tweet tends to be uniformly distributed.

The amount of semantic noise added per tweet directly influences the results (even though in some case more than in others). When the maximum amount of noise (*i.e.* up to 15 terms per fake tweet) is allowed, user profiles are rapidly

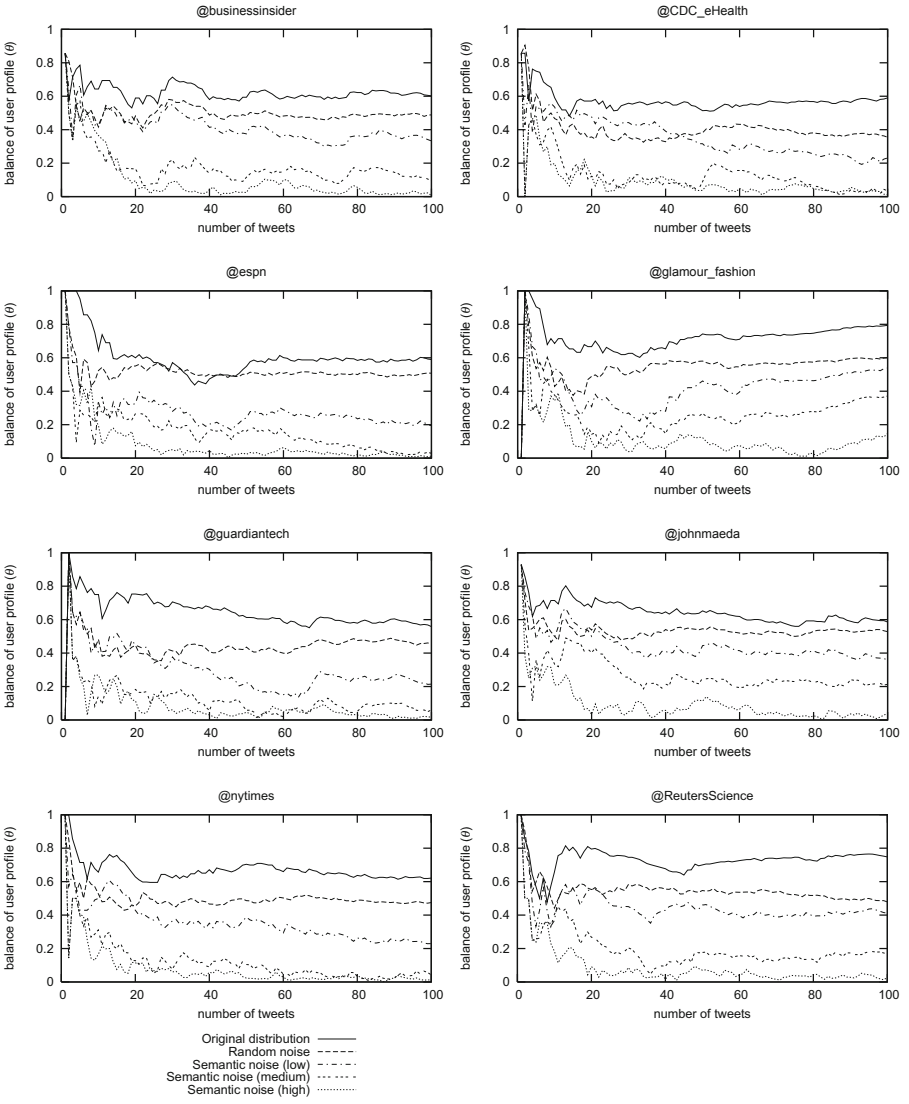

Fig. 1. Profile balancing results for the different methods and scenarios for the eight Twitter users

balanced with θ values below 0.1. Only around 20 user tweets (and, hence, 20 additional fake ones) are needed to achieve that figure. On the contrary, when restricting the amount of semantic noise to the minimum (*i.e.,* up to 3 terms per tweet), an ideal profile balancing (near to zero) can be hardly achieved. Even though more than 100 tweets could be considered, it seems unlikely to achieve a balance due to the curve shape tend to stabilize horizontally around

θ values between 0.2 and 0.4. This indicates that, in this case, the amount of allowed noise is not enough to achieve a good balance, especially considering that there are *eight* profile categories that should be balanced. The case in which a medium amount of semantic noise is allowed (*i.e.,* up to 7 terms, which is closer to the number of profile categories), presents a more variable behavior. For some users (like *@espn*, covering topics of Arts, Sports, Society and Shopping, *@CDC_eHealth* covering topics of Health, Society and Business, or *@ny_times*, posting about Society, Business, Arts, Computers and Shopping), the system was able to catch the lower figures obtained by the noisiest setting, despite a smoother decreasing shape. These cases correspond to users who post more heterogeneous tweets (covering several profile categories) and, hence, requiring from less noise (terms from other categories) to achieve an equilibrium. For other users (*e.g., @glamour_fashion*, *@johnmaeda*) with profiles more focused towards a dominant category (*e.g.,* Arts, in these cases), profile balancing requires more noise to achieve the balance.

In the graphs, it can also be observed how spikes present in the original distribution (*e.g.,* around the 30th tweet for *@businessinsider*, 50th for *@espen* or 65th for *@guardiantech*), which represent a notorious change in the accumulated user profile, are also reflected in the balanced profiles even more clearly. However, the fact that user profiles are re-computed after each new user tweet allows the system to dynamically adapt its behavior (*i.e.,* categories of added noise) to eventual or long-term changes in user preferences or interests.

In any case, when comparing the semantic setting with the random scenario, it can be observed that, even though the random approach always introduces the maximum amount of noise (*i.e.,* 15 terms per fake tweet), it poorly balances the user profile. In fact, assuming that random noise is uniformly distributed according to profile categories, it would hardly lower the provided figures if the user maintains her preferences throughout time.

Finally, more spikes can be observed in the 1-20 tweets zone when analyzing curves of semantic scenarios. This corresponds to the zone in which the user profile is being characterized, and the dominant category/ies may change from one tweet to the next. In consequence, noise categories may also vary from one fake tweet to the next, producing more pronounced spikes in the profile balancing. As stated in related works [9, 10], individual tweets are too short and ungrammatical to enable an accurate profile characterization. From the analysis of the results, it can be concluded that, at least, 10 tweets (and preferably 20) are required to obtain a stable profile (and hence, a more accurate and coherent noise addition), even though the concrete number may vary from one user to another, according to the topical coherency of her tweets and the homogeneity of her profile.

5 Conclusion

In this paper, we have proposed a new system that prevents text-based profilers from characterizing the dominant profile categories of the users of microblogging

services like Twitter. Our scheme generates and publishes fake tweets together with legitimate ones. These fake publications are constructed according to two basic principles: (i) They contain specially tailored terms, introducing a semantic distortion in the user profile in order to hide user characteristics (*i.e.,* dominant profile categories) in front of automatic profiling methods; and (ii) They are formed by a concatenation of terms, which leads to a lack of semantically-coherent discourse that allows human readers to easily discern between user tweets and fake ones, while preventing automatic profilers who analyze tweets according to term distribution to discover dominant topics/categories.

The evaluation results obtained show that: (i) the proposed system effectively balances user profiles in front of profilers based on term distribution; (ii) it achieves that balance with a quite limited number of publications (between 10 and 20 tweets are enough to obtain a stable profile); and (iii) it dynamically adapts its behaviour to eventual or long-term changes in user preferences or interests.

Regarding future work, it would be interesting to evaluate the use of the presented approach by regular users in their daily duties in order to confirm the qualities showed by the simulations and also to rate the usability of the system and its level of intrusiveness in a real situation.

Disclaimer and Acknowledgments. Authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004 and Audit Transparency Voting Process IPT-430000-2010-31), by the Spanish Ministry of Industry, Commerce and Tourism (through projects eVerification2 TSI-020100-2011-39 and SeCloud TSI-020302-2010-153) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

1. McMillan, G.: Twitter Reveals Active User Number, How Many Actually Say Something. Time - Techland (September 2011)
2. Consumer Reports National Research Center, Annual State of the Net Survey, Consumer Reports 75(6) (2010)
3. Rui, H., Whinston, A.: Information or attention? An empirical study of user contribution on Twitter. Information Systems and E-Business Management, 1–16 (2011)
4. Zhang, C., Sun, J., Zhu, X., Fang, Y.: Privacy and Security for Online Social Networks: Challenges and Opportunities. IEEE Network 24(4), 13–18 (2010)
5. TweetPsych (2011), `http://tweetpsych.com`
6. Peerindex (2011), `http://www.peerindex.com`
7. Ebner, M., Mühlburger, H., Schaffert, S., Schiefner, M., Reinhardt, W., Wheeler, S.: Getting Granular on Twitter: Tweets from a Conference and Their Limited Usefulness for Non-participants. Key Competencies in the Knowledge Society 324, 102–113 (2010)

8. Zoltan, K., Johann, S.: Semantic Analysis of Microposts for Efficient People to People Interactions. In: Proc. of the Roedunet International Conference – RoEduNet 2011, pp. 1–4 (2011)
9. Michelson, M., Macskassy, S.A.: Discovering Users' Topics of Interest on Twitter: a First Look. In: Proc. of the Fourth Workshop on Analytics for Noisy Unstructured Text Data (2010)
10. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) ESWC 2011, Part II. LNCS, vol. 6644, pp. 375–389. Springer, Heidelberg (2011)
11. DBpedia (2011), http://dbpedia.org/
12. Luo, W., Xie, Q., Hengartner, U.: Facecloak: an Architecture for User Privacy on Social Networking Sites. In: Proc. of the 2009 International Conference on Computational Science and Engineering, pp. 26–33 (2009)
13. Abril, D., Navarro-Arribas, G., Torra, V.: On the Declassification of Confidential Documents. In: Torra, V., Narakawa, Y., Yin, J., Long, J. (eds.) MDAI 2011. LNCS, vol. 6820, pp. 235–246. Springer, Heidelberg (2011)
14. Martínez, S., Sánchez, D., Valls, A.: Semantic adaptive microaggregation of categorical microdata. Computers & Security 31(5), 653–672 (2012)
15. Sánchez, D., Batet, M.: Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics 44(5), 749–759 (2011)
16. Sánchez, D.: A methodology to learn ontological attributes from the Web. Data and Knowledge Engineering 69(6), 573–597 (2010)
17. Martínez, S., Sánchez, D., Valls, A., Batet, M.: Privacy protection of textual attributes through a semantic-based masking method. Information Fusion 13(4), 304–314 (2012)
18. Sánchez, D., Moreno, A.: Pattern-based automatic taxonomy learning from the Web. AI Communications 21(1), 27–48 (2008)
19. Sánchez, D., Isern, D., Millan, M.: Content Annotation for the Semantic Web: an Automatic Web-based Approach. Knowledge and Information Systems 27(3), 393–418 (2011)
20. Guarino, N.: Formal Ontology in Information Systems. In: Proc. of the 1st International Conference on Formal Ontology in Information Systems, FOIS 1998, pp. 3–15 (1998)
21. Open Directory Project (2012), http://www.dmoz.org/docs/en/about.html
22. Twitter - WhoToFollow (2012), http://twitter.com/#!/who_to_follow