

# Rumour detection and monitoring in open source intelligence: understanding publishing behaviours as a prerequisite

## Abstract

In the context of information warfare, rumour detection has become a central issue. From classical media-related campaign, to propaganda and indoctrination that lie at the core of terrorism, rumour is a mean widely used and thus a threat that must be identified as soon as possible, and in the best-case scenario, anticipated and curbed. The emergence of a new informational environment due to the adoption of the Internet as a massive information diffusion medium has led to a situation suitable to the creation and propagation of rumours. Indeed, the Web gives everyone not only the possibility to observe information flows but also the opportunity to influence and create them.

In order to tackle the issue of rumour detection, one has to understand the mechanisms underlying their propagation. In this perspective, we believe that it is essential to identify and understand the publishing behaviours of the sources. Therefore, we focus in this paper on the identification of groups of sources with similar publishing characteristics. We propose to tackle this problematic by using clustering methods on data extracted from Web sources. The resulting clusters obtained from the clustering are then interpreted as groups of Websites behaving similarly and used to characterize publishing behaviours.

## Keywords

Publishing behaviours, rumour detection, Web sources, stacked clustering.

## Introduction

The adoption of the Internet as a massive information diffusion medium has considerably modified information dynamics. The Web gives everyone not only the possibility to observe information flows but also the opportunity to influence and create them. With very little knowledge and few means any Internet user is able to send information to selected recipients or, more important, display it publicly, sharing it to potentially anyone logged on the Web. New tools available to publish information are invented regularly; the emergence of social networks, like Twitter, is a recent example. Forums, chatrooms and above all Weblogs are rapidly growing means to create and spread information. Thus, nowadays, information Websites based classically on the model of traditional media are now mixed with autonomous and personal publishing modalities.

Every publication means to publish information has its own technical characteristics and is adapted to specific practices; nevertheless they all have in common the ability to be interconnected. Thus, an information publisher or information source can not only play the role of an initiator, but also the role of an intermediary, regardless of the tool used to publish. Therefore every source is playing its share in the way information propagates on the Web.

As a consequence of these important changes, three observations can be outlined:

- The quantity of available open source information is considerably growing.
- Flows of information are speeding up in an uncontrolled way.
- Sources of information are branching out, wearing multiple and unsorted faces.

Information is now more than ever subject to amplification, modification and distortion as the number of possible sources takes off. This media environment is suitable to the emergence and propagation of rumours that are not limited to insignificant subjects: rumours can have major consequences on political, strategic or economical decisions. Increasingly, they are triggered off on purpose for various

reasons: campaigns can be carried out in order to discredit a company, endanger strategic choices or question political decisions.

One of the specificity of the rumour is to be widely spread. An inaccurate, excessively modified, or from start to end made up information can be considered as a rumour only if it has reached a considerable amount of people. Therefore, in order to be able to detect rumours, one has to understand the mechanisms underlying their propagation. We believe that these mechanisms can be revealed by studying changes in publishing behaviours. In this context, it seems essential to identify and understand these behaviours.

This paper focuses on the identification of groups of sources with similar publishing characteristics. We propose to tackle this problematic by using clustering methods on data extracted from Web sources. The resulting clusters obtained from the clustering are interpreted as groups of Websites behaving similarly and used to characterize publishing behaviours. This study is grounded on the use of real observed data extracted from Web publications. It does not need any a priori knowledge on the data as the proposed methodology is based on the computation of raw data. Nevertheless, some choices and hypotheses were made to be able to extract and structure Web publications, notably we introduce a model to formalize simply sources citations thanks to a network structure.

This paper is organized as follows: in section 2, we first depict the problem of rumour detection as a defence and security issue, and how it is related to our process of identification of publishing behaviours. Section 3 describes the methodology we used to obtain publishing behaviours from real data. In section 4, we present the conditions of our experiments and the obtained results.

## **Rumour detection**

### ***Rumours and terrorism***

The link between rumours and terrorism is strong on many levels. A high level of rumour propagation is usually the expression and the indication of a global state of emotional tension like fear or anxiety among a population. For example, warfare or conflict situations are known to provide optimal conditions for rumour dissemination [14].

Similarly, besides the loss of human lives and infrastructures, a terrorist attack induces psychological trauma at both individual and collective level that leads to a general situation of group anxiety and fear. Thus, a terrorist attack usually goes with multiple rumours that can be purposely created in order to affect the national morale and the trust in the government in power.

Terrorist organizations are becoming real expert in communication strategies. Terrorist attacks are planned in conjunction with a communication strategy: they usually start with a warning and are obviously claimed afterwards. Moreover, the core of terrorism is based on indoctrination and propaganda that are conducted by disinformation campaigns and manipulation of information. The efficiency and impact of these campaigns on the public opinion is closely related to the critical sensibility of the population and its level of anxiety and fear. The use of obnoxious rumours by terrorists is a very efficient method to maintain and worsen such a situation.

In this context, rumour detection has become a central issue in anti-terrorism activities for two main reasons: firstly because they are indicators of a situation where terrorism propaganda is the most efficient and secondly because they are created purposely by terrorist organization to manipulate public opinion and legitimate their actions.

### ***Related work***

The study of rumours is a classical theme in social sciences. Case studies on specific rumours have been conducted in order to reveal their context of formation and the common characteristics between

rumours [6]. Some models inspired from physics have been introduced in order to grasp the social context of a rumour formation [3].

Work on rumours is related to information spreading. Many studies on the problem of information propagation are inspired from the more common issue of contagion and generally use models based on the standard model for viral epidemics in populations: the susceptible-infected-recovered (SIR) model [11]. Studies on information propagation are usually based on a network where a node is an information source and an edge an information exchange between two nodes. On this subject, research has focused on the effects of the topological properties of the network on the propagation of infection [12] or on inferring the source of a rumour in a network [13].

We distinguish two methods to detect rumours. The first one is based on the analysis of the informational content. Statistical and linguistic analysis like the method presented in [4] to track meme could be use to detect rumours. The second method focuses on the structure of the publication network through which rumours might propagate. A study of the variations of such a network has been conducted in [5].

Most of these works consider the network and its reaction to the propagation model. However, nodes in the network usually have the same behaviour defined by the propagation model. Sources are not differentiated and there has been no work on how to include multiple publishing behaviours. In the perspective of detecting rumours, it seems central to understand the publishing behaviours of the information sources. Our contribution to this subject is the identification from real data of typical behaviours of sources publishing and propagating information on a network.

## **Identifying publishing behaviours**

The objective is to identify groups of homogeneous sources, by considering publishing characteristics. We propose to tackle this problematic by using a clustering method. This section describes the different steps of the methodology we propose to apply. It starts with the choice of a series of descriptors that could differentiate behaviours of publication. These descriptors derive directly from real data extracted from Website publications and therefore are linked to what we are able to extract. We present in this section the choices of the extraction process based on a simple model of information propagation. Then we describe the stacked clustering method that starts with the choice of the number of clusters and eventually identifies a partition of sources.

### ***Defining descriptors from extracted data***

In order to define some descriptors to characterize publishing behaviours, we first need to choose a model to represent the mechanisms of publishing behaviours. We based our strategies to extract the data and to choose the final definition of the descriptors on this model.

### **A model for information propagation**

We propose in [5] a simple model of information representation that can be implemented and used with real data and from which characteristics of publishing behaviours can be derived. Indeed, for practical reasons, we focus on the visible and extractable data that is to say Websites publications.

The model is based on a graph of Websites. Each node of the graph represents a Website. The nodes are linked to each other by directed edges meaning "*is a source of information for*". This way, information propagation can be easily monitored following the directed edges on the network. We base our representation of the source network on the following hypothesis: when a publisher explicitly refers to another Web page using a hyperlink, the Website pointed by the cited link is considered as one of the information sources for the publisher.

## Extraction strategies

We developed a specific Web crawler to extract the data needed for our analysis. We named our crawler ONICS (Outils de Navigation, d'Indexation et de Classement des Sources), which stands for browsing, indexing and sorting tool for sources. The specificities of ONICS are to extract every publication of a list of sources chosen by the user and to identify the hyperlinks cited in the published articles. A more detailed description of the crawling choices and extraction process is available in [5]. Basically, this tool is able to store in a database the text of articles, their publication date and the hyperlinks cited for all the publications of a list of sources. From this data we are able to build the source network described above.

## Extracted descriptors

We choose to derive the descriptors from our model of information propagation. Therefore, they do not take into account the content and themes of the articles published but only the structure of the source network and its dynamics. We use the following descriptors for each source:

- The average and standard deviation of the *publication intervals*
- The *number of published articles and links*
- The *diversity* of the cited sources defined as the ratio between the number of different sources and the overall number of cited links
- The average and standard deviation of the *recentness* of the cited articles. The measure of recentness is defined as the interval between the publication date of the studied article and the publication date of the cited articles.

We believe that these seven descriptors characterise the publication habits of a source including the criteria each Website relies on to select its own sources of information. Thus, we will be able to extract publishing behaviours by using a clustering method on these data.

## Using a stacked clustering method

The choice of the clustering method is central in the process. We suppose that we do not have a priori knowledge on the awaited results, in particular, the precise number or size of the clusters are unknown. Therefore, we choose to use a clustering method robust enough not to question the validity of the resulting clustering. Preliminary tests with k-means and hierarchical clustering algorithms gave useless results: the k-means algorithms gave very unstable partitions due to random initialization whereas hierarchical clustering gave very different results with different linkage strategies making the choice of selecting a final partition quite arbitrary. To tackle this issue, we choose to apply a stacked clustering method [7] using multiple k-means iteration and a hierarchical clustering.

This section describes how the results of the k-means iterations are used, firstly to choose the number of clusters and secondly as an input of a hierarchical clustering. We decided not to take into consideration the result for k equals 2, because we believe that identifying only two kinds of behaviours presents very little interest. The k-means algorithm has some restrictions as it takes the number of clusters k as an input parameter and its result heavily depends on the initial clusters. To circumvent these problems we chose to run it multiple times (1000 times for each k, with k defined from 3 to 10) with random initial clusters. The results presented in the following sections are based on the analysis of the 8000 obtained partitions.

## Choosing the number of clusters

From the results of the k-means clustering, we use a measure introduced in [7] to choose the number of clusters. The idea is to define the best number of clusters as the integer k for which the partition resulting of the k-means is the most stable.

Let  $I_k$  be a set of partitions of k clusters. Let  $P_i \in I_k$  a partition. We then look for the number k for which the stability of  $I_k$  is maximum. For that we need a similarity measure to evaluate the degree of match

between two partitions, and a stability measure for  $I_k$  that acts as an aggregator of the similarity measure for all the partitions of  $I_k$ . We used the *adjusted Rand index* [8, 9] for the similarity measure and the pairwise individual stability [7] for the stability measure.

### The adjusted Rand index

The adjusted Rand index is used to compute the similarity between two data partitions. This index takes values between 0 (for partitions completely independent of one another) and 1 (for identical partitions). For two partitions A and B, let  $n$  be the total number of points in the dataset;  $a$  the number of point pairs in the same cluster under both A and B;  $b$  the number of point pairs in the same cluster under A but not B;  $c$  the number of point pairs in the same cluster under B but not A;  $d$  the number of point pairs in different clusters under both A and B. The measure is defined as follows:

$$AR(A, B) = \frac{\left(\frac{n(n-1)}{2}\right)(a+d) - ((a+b)(a+c) + (c+d)(b+d))}{\left(\frac{n(n-1)}{2}\right)^2 - ((a+b)(a+c) + (c+d)(b+d))}$$

### Pairwise individual stability

The pairwise individual stability  $S(k)$  uses the adjusted Rand index of each pair of partitions of  $I_k$  (*i.e.*  $|I_k|(|I_k|-1)/2$  pairs where  $|I_k|$  is the number of partitions of  $k$  clusters). It computes the sum of the adjusted Rand index for all the pairs of partitions

$$S(k) = \sum_{i, j \in I_k, i < j} AR(P_i(k), P_j(k))$$

The chosen number of clusters is then the value of  $k$  where  $S(k)$  is maximum.

### Choosing a final partition

After having chosen the number of clusters  $k$ , we use the overall results of the  $k$ -means clustering in order to select one final clustering. To do that, we first derive the co-association matrix for each of the 8000 partitions. The *co-association matrix*  $M^{(P)}$  for partition  $P$  is defined as:

$$M^{(P)} = \{m_{i,j}^{(P)}\} \text{ where}$$

$$m_{i,j}^{(P)} = 1 \text{ if element } i \text{ and element } j \text{ are in the same cluster in partition } P$$

$$m_{i,j}^{(P)} = 0 \text{ if element } i \text{ and element } j \text{ are in different clusters in partition } P$$

From the 8000 co association matrices, we then derive the *consensus matrix*  $M$  defined as :

$$M = M^{(P1)} + M^{(P2)} + \dots + M^{(P8000)}$$

The consensus matrix contains the number of times where points are in the same cluster. From this observation, we note that if the number is high for two points, it means that they belong usually to the same cluster and on the opposite, if it is low, they do not probably belong together. Therefore, it seems natural to consider the consensus matrix  $M$  as a similarity matrix between the points [7]. Finally, after having converted the similarity measures into distance measures (by subtracting the maximum value to each similarity measure), we are able to apply a hierarchical algorithm.

Finally, the chosen partition is the result of this hierarchical clustering obtained by cutting the dendrogram at the best value of  $k$  computed in the last section.

## Experiments

### Data

To conduct our experiments, we use a database containing 190000 articles and 140000 links for a total of 110 Websites crawled daily between February and November 2009 using the ONICS tool. For this database was created for testing purposes, we chose to crawl a series of generalist information Websites. The database forms our main corpus that we used to establish the publishing behaviours. For further tests we used a secondary corpus that has been originally built in the context of a competitive intelligence analysis within the defence area.

### Results

In this section, the result of the overall methodology of the identification of publishing behaviours is presented.

Figure 1 shows the results of the stability measure  $S(k)$  for the main corpus (in blue with squares) and the secondary corpus (in red with circles).

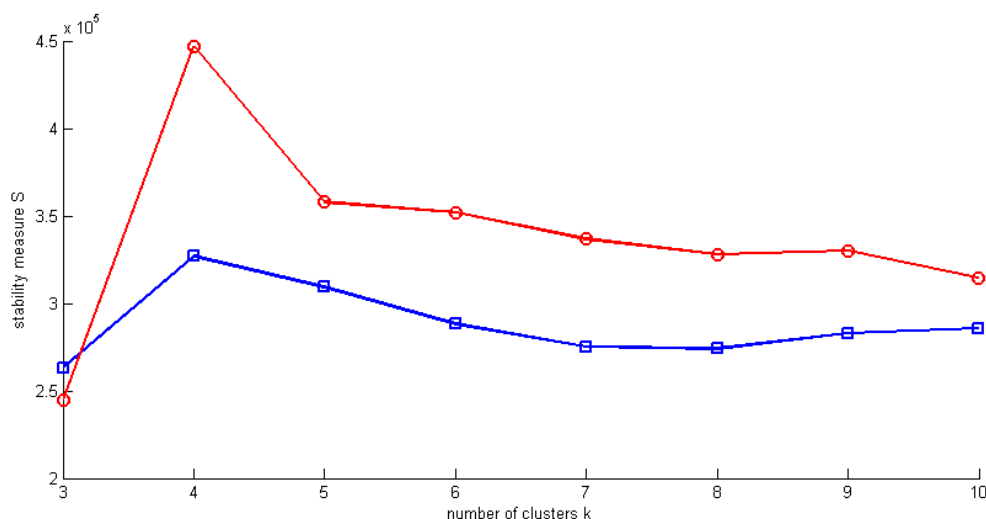


Figure 1: Results of the stability measure for two different corpora

We obtain an optimal number of clusters of 4. The test for the secondary corpus seems to confirm that the probable number of different publishing behaviours is 4.

Figure 2 shows the final result of the hierarchical clustering. It is the dendrogram cut for a number of clusters of 4. We can note, as a first observation that the sizes of the clusters are really unbalanced.

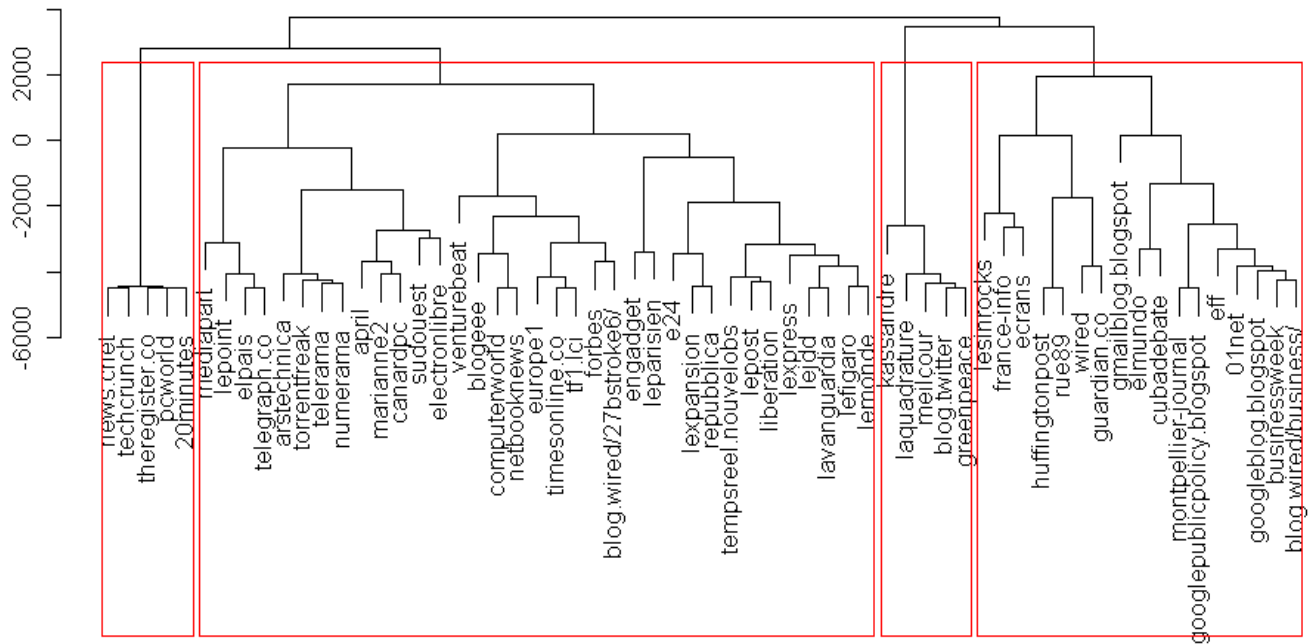


Figure 2: Dendrogram of the hierarchical clustering with a cut for 4 clusters

## Interpretation of the clustering results

From the result of the final partition, we try to extract the main characteristics of each cluster in order to identify the main aspects of the different behaviours. Figure 3 represents for each cluster the average value (scaled for each descriptor to have mean 0 and standard deviation 1) of each of the initial descriptors. The descriptors, from the left to the right are: the standard deviation and the average of the publication intervals (resp. `pub_sd` and `pub_avg`), the number of article and links published (resp. `nb_a` and `nb_l`), the diversity of the cited sources and the average and standard deviation of the recentness of the cited articles (resp. `rec_avg` and `rec_ds`).

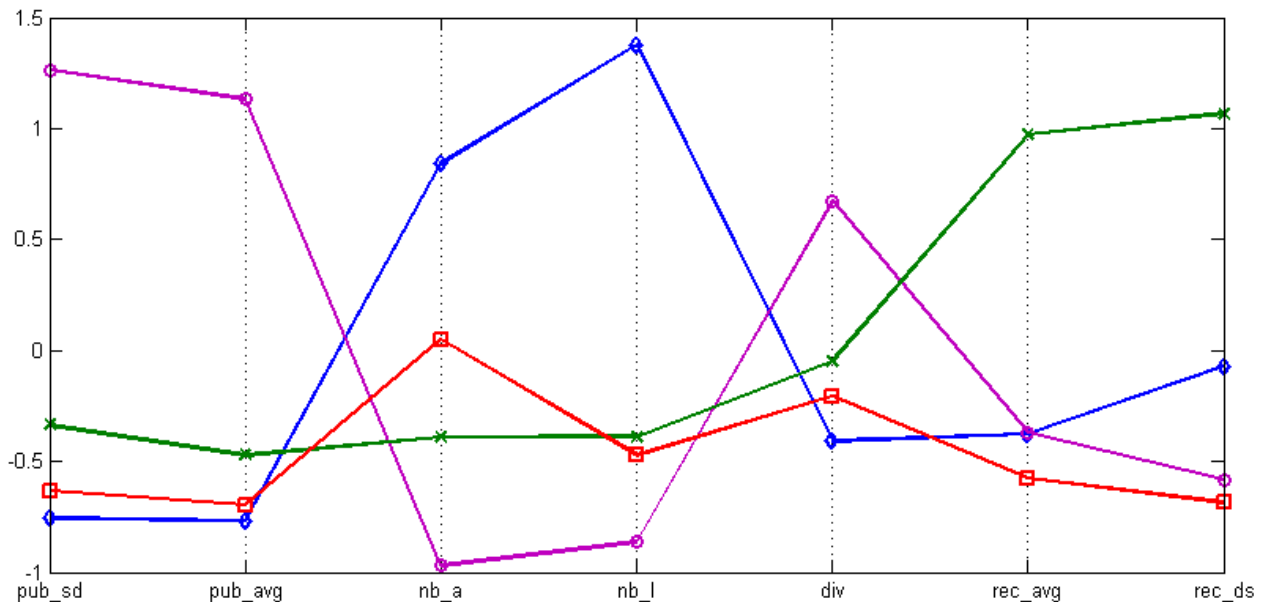


Figure 3: Centroids visualization on parallel coordinates

From these results, the following observations can be derived:

- The first cluster (in blue and diamonds) is small (it contains 8% of all the Websites) and is characterized by a low publication interval and a high number of articles; these measure are

the expression of a very substantial publishing behaviour. A high number of links and low diversity measure is consistent with the fact that this cluster is mainly composed of specialised blog (techcrunch, pword).

- The second cluster (in violet and circles) is small as well (8% of the Websites) and on the opposite is characterized by a very high publication interval and a low number of links. It is worth pointing that this is the cluster with the highest diversity measure that is however not very representative because the Websites within this cluster publish rarely and uses few links. They seems to be even more specialised than the first cluster (kassandre, laquadrature) which may explain their very low publishing activity.
- The third cluster (in green and crosses) and fourth (in red and squares) have respectively an average size (27%) and big size (56%) and have quite the same characteristics (low publication interval and medium number of links). The difference lies in the descriptor of recentness of the cited articles. At first sight they seem composed of quite similar type of Websites. When exploring in more details the composition of these two clusters, we note that the third cluster contains Websites probably more familiar with Web specific tools for publication like blog platform or collaborative publishing (huffingtonpost, googleblog) whereas the fourth cluster contains mainly Web version of existing newspapers (timesonline, lemonde, elpais...). This interpretation is confirmed by the very high reactivity measure of the third cluster and the fact that even with a smaller frequency of publication, the third cluster uses more hyperlinks per article than the fourth cluster.

The centroids visualization shows that the standard deviation and the average (for both measures: publication intervals and recentness) are two descriptors with the same impact on the final clustering. It should be considered not to use both in future implementations.

The sizes of the clusters give a precise idea of the composition of the network observed in our extracted data. This is an important knowledge when studying rumours within the perimeter of this precise network. However, it is worth noting that there is no reason to conclude that these proportions are representative of any network extracted from Web publications or even less the general composition of the Web.

## Conclusions and future works

Initiating rumours is a powerful mean for terrorists to affect public opinion and legitimate their actions. Moreover, the emergence of all kinds of rumours is an indicator of a situation favouring the impact of terrorist communication campaigns. Thus the detection, monitoring and curbing of rumours is an important challenge for counter-terrorism organizations.

The monitoring of communication campaigns of terrorist organizations in open source is a case study very promising in the context of research on rumour detection.

The method presented in this paper can be directly applicable on a predefined network of sources in close contact or managed by terrorist activists. By giving a categorization of the sources according to their communication habits, it improves the knowledge that counter-terrorism organizations have on the sources and may help to assess their level of toughening or relation to the terrorist organizations.

Nevertheless, the monitoring of open sources on which this method is based is only one possible perspective to tackle the issue of information manipulation by terrorist communication campaigns.

This paper presented a methodology to identify groups of sources with similar publishing characteristics. We chose not to take choices based on *a priori* knowledge concerning the awaited results and based our choices on the use of real data extracted from Web sources publications.

We used a robust stack clustering methodology in order to select a possible partition. It is based on multiple results of k-means clustering, at first, to choose a number of clusters and then to feed a hierarchical clustering. The methodology is very robust but cannot be applied to very large sets of data

because it is very demanding in computational resources. This disadvantage that must always be kept in mind, does not worry us unduly because in the case of strategic intelligence, the number of interesting sources rarely exceeds a thousand.

This study provides interesting results in the establishment of publishing behaviours. Notably, we identified four different groups and more important the aspects on which they differentiate. Moreover, this work reveals the proportion of each group in a real network of sources. Because it is extracted from observed data, these results should be very useful for future works on information propagation and particularly studies on rumour detection.

As an extension of this work, we suggest to validate the choice of our descriptors with experts in intelligence analysts. At the same time, other interesting descriptors may be defined to confirm our results or even identify other underlying behaviours among the four groups we presented. These may be based for example on the article itself (size or structure) or on more sophisticated topological measures of the network. The difficulty of the task remains in the ability to extract the values of these descriptors from real Web publication data. Moreover, it would be interesting to experiment the presented methodology on other corpora in order to see if new behaviours can be identified.

In a perspective of research on rumour detection, we plan to use these results to calibrate a model simulating dynamics of citation between Web sources. Simulation gives numerous possibilities to understand the evolution of rumours in a network by controlling the parameters that could initiate or on the contrary terminate a rumour. By adding distinct publishing behaviours derived from observed data in such a model, we hope to achieve better understanding of the reality of this complex phenomenon.

## References

- [1] D. H. Zanette. Dynamics of rumor propagation on small-world networks. *Physical Review E*, Vol. 65, No. 4, 2002.
- [2] Nekovee, Y. Moreno, G. Bianconi, and M. Marsili. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, Vol. 374, No. 1, pp. 457-470, 2007.
- [3] S. Galam. Modeling Rumors: The No Plane Pentagon French Hoax Case. In *Physica A: Statistical Mechanics and its Applications*, Vol. 320, pp. 571-580, 2003.
- [4] J. Leskovec, L. Backstrom, J. Kleinberg. Meme-tracking and the Dynamics of the News Cycle. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.
- [5] F. Nel, A. Carré, P. Capet, and T. Delavallade. Detecting Anomalies in Open Source Information Diffusion. In *IST087 NATO Symposium on Information management and Exploitation*, 2009.
- [6] P. Froissart. Rumor. In *The International Encyclopedia of Communication*. Blackwell Publishing, 2008.
- [7] L. I. Kuncheva and D. P. Vetrov. Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, pp 1798-1808, 2006.
- [8] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, Vol. 2, No. 1, pp. 193-218, 1985.

- [9] J. M. Santos and M. Embrechts. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. ICANN, Vol. 5769 of *Lecture Notes in Computer Science*, pp 175-184, 2009.
- [10] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: Wiley, 2004.
- [11] P. S. Dodds and D. J. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, Vol. 92, No. 21, 2004.
- [12] M. E. J. Newman. The spread of epidemic disease on networks, *Physical Review E*, Vol. 66, No. 1, 2002.
- [13] D. Shah and T. R. Zaman. Rumors in a Network: Who's the Culprit ? In *Neural Information Processing Systems (NIPS)*, 2009.
- [14] G. W. Allport and L. Postman. *The Psychology of Rumor*. New York, NY:Russell & Russell, 1947.