Validating viral marketing strategies in Twitter via Agent-based Social Simulation, extended material

Emilio Serrano [a]

[a] *Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros Informáticos, Departamento de Inteligencia Artificial, Campus de Montegancedo, Boadilla del Monte, 28660, Madrid, Spain*

Abstract

Extended material for the paper "Validating viral marketing strategies in Twitter via Agent-based Social Simulation" presented in the Expert Systems with Applications journal (ESWA) in 2015.

Key words: Agent-based Social Simulation, Viral marketing, Social Network Analysis, Rumor Spreading Model, Twitter, Big Data

Contents

Email address: `emilioserra@fi.upm.es` (Emilio Serrano).

Preprint submitted to Expert Systems with ApplicationsReceived: date / Accepted: date

1 Related works

In the spirit of the systematic review methods (Nassirtoussi et al., 2014), several review questions were formulated before locating and selecting relevant studies. These questions are the following:

- Q1. Does the work deals with rumors spread?
- Q2. Does it include the Twitter case?
- Q3. Real data is employed in the study?
- Q4. Does the paper simulate the information diffusion?
- Q5. Is there agent-based social simulation?
- Q6. Are there what-if scenarios?
- Q7. A general methodology is presented to validate and use simulations?
- Q8. Is the data provided?
- Q9. Is the implementation given?
- Q10. Is it free and open source software?

Note that these questions fall in three main categories: (1) type of target studied (Q1-Q3); (2) method employed (Q4-Q7); and, (3) reproducibility of the research (Q8-Q10). Moreover, the questions are not disjoint, e.g. if no real data is employed (Q3), data cannot be provided (Q8). Table 1 summarizes the works revised and answers for these review questions.

Valecha et al. (Valecha et al., 2013) analyze Twitter data of the Haiti earthquake in 2010 [1]. The authors categorize seven different communication modes for four time stages at this occurrence. The paper concludes that information with credible sources contributes to suppress the level of anxiety in Twitter community, which leads to gossip controlling and high information quality. In this vein, Mendoza et al. (Mendoza et al., 2010) explore the behavior of Twitter users in the 2010 earthquake in Chile. The authors classify the tweets manually in affirms, denies, or unknown. They also conclude that hearsay tend to be questioned more than news by the Twitter community. Starbird et al. (Starbird et al., 2014) present another exploratory work which deals with the 2013 Boston Marathon Bombing [2] and conclude that corrections to the misinformation emerge but are muted compared with the propagation of the misinformation. Cha et al. (Cha et al., 2010) use Twitter data to gain insights into viral marketing and, more specifically, to compare three measures of influence: indegree, retweets, and mentions. These authors conclude that popular users who have high indegree are not necessarily influential in terms

[1] On January 12, 2010, a devastating earthquake with a magnitude of 7.3 struck Haiti. More than 220,000 people were killed and over 300,000 injured.

[2] The Boston Marathon bombings were a series of attacks and incidents which began on April 15, 2013, when two pressure cooker bombs exploded during the Boston Marathon, killing 3 people and injuring an estimated 264 others.

	Target system			Method				Reproducibility		
Ref.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Valecha et al.	✓	✓	✓					UR		
Mendoza et al.	✓	✓	✓							
Starbird et al.	✓	✓	✓							
Cha et al.	✓	✓	✓							
Weng et al.		✓	✓	✓	✓					
Gupta et al.	✓	✓	✓							
Kwon et al.	✓	✓	✓					UR		
Qazvinian et al.	✓	✓	✓					UR		
Nekovee et al.	✓			✓						
Zhao et al.	✓			✓						
Shah and Zaman	✓			✓						
Domenico et al.	✓	✓	✓	✓						
Jin et al.	✓	✓	✓	✓						
Tripathy et al.	✓	✓	✓	✓	✓	✓				
Liu and Chen	✓	✓		✓	✓					
Seo et al.	✓	✓	✓	✓	✓	✓				
Yang et al.	✓	✓	✓	✓	✓	✓				
Gatti et al.		✓	✓	✓	✓	✓				

Table 1
Review questions for survey. Check mark: yes, empty space: No, UR: under request.

of retweets or mentions, while influence is gained limiting tweets to a single and specialized topic. These works hint at the potential of understanding hearsay diffusion and having strategies to control them. Nevertheless, they do not cope with these strategies or their evaluation by simulation techniques.

Weng et al. (Weng et al., 2013), without dealing with gossips specifically, address meme propagations in Twitter. Memes are parts of cultural tradition, e.g. thoughts, cultural techniques, behaviors, etcetera (Flentge et al., 2001). In Weng et al.'s work, memes are identified with a Twitter hashtag, i.e. a metadata tag used in Twitter and which consists of a word or an unspaced phrase prefixed with "#". The authors, based on real data, compare memes propagation with four simple simulated models: random, cascade, social reinforcement, and homophily. Finally, the authors present a method to detect if a meme will go viral depending on the meme first 50 tweets and machine learning techniques. Although this is a very significant work which gives sound results to support the hypothesis presented, it does not intend to give realistic simulated models or use them for designing and testing any strategy. Moreover, as displayed in table 1, data and implementations are not given.

Other works also propose machine learning models after an exploratory data analysis of Twitter. Gupta et al. (Gupta, Lamba, & Kumaraguru, 2013; Gupta, Lamba, Kumaraguru, & Joshi, 2013) study tweets of the Boston marathon blasts and propose a regression prediction model. This model allows calculating the number of nodes which will be infected in a network assuming that fake content is published by a specific user. In this vein, Kwon et al. (Kwon et al., 2013) identify a large number of characteristics in rumors under three main categories: temporal, structural, and linguistic. Then these features are used in several machine learning algorithms to classify a Tweet as rumor or non-rumor. Qazvinian et al. (Qazvinian et al., 2011) also deal with misinformation detection and explore the effectiveness of three categories of features: content-based, network-based, and specific memes. These machine learning models are important contributions for viral marketing, but they do not allow researchers to test marketing strategies with them. Moreover, as pointed out in some works (Qazvinian et al., 2011), identifying new emergent rumors directly from the Twitter data is more challenging than the classification of a dataset previously retrieved. In a sense, the research line presented in these works is complementary of the presented here. On the one hand, machine learning approaches may employ features taken from simulated models (Kwon et al., 2013). On the other hand, the strategies tested with simulation can be undertaken when detected gossips by these machine learning approaches.

The epidemiological modeling is popularly employed to model rumor diffusion. In this line, the population is divided into several classes such as susceptible (S), infected (I), and recovered (R) individuals. These analytical models are usually formulated using differential equations since the transition rates from one class to another are mathematically expressed as derivatives. The standard model in this line is the *SIR* model (Hethcote, 2000) (susceptible, infected, recovered). Moreover, the *SI* (susceptible, infected) and *SIS* (susceptible, infected, susceptible) models are also very used. Nekovee et al. study the SIR model applied to gossip spread in complex social network (Nekovee et al., 2007). In this vein, Zhao et al. (Zhao et al., 2013) extends the SIR model with forgetting mechanisms. Shah and Zaman (Shah & Zaman, 2011) use a SI model to study algorithms to find a misinformation source in a network. Domenico et al. (De Domenico et al., 2013) study Twitter hearsay about the Higgs boson discovery and reproduce the global behavior using the SI model and extending it. Jin et al. (Jin et al., 2013) employ the *SEIZ* model (which considers exposed individuals, E, and skeptics, Z) for capturing diffusion of gossips and news in Twitter. The main appealing of these works is the accuracy they achieve by adjusting automatically the model parameters, e.g. population size, with fourth generation programming languages such as MATLAB. On the other hand, comparing these model to real-world data is difficult and they often require overly simplistic assumptions (Rand & Rust, 2011). These works employ social simulation (a society is modeled), but they are not ABSS works (equations describe the society instead of agents). Furthermore, unlike ABSS, they do not allow the exploration of individual-level theories of behavior which can be used to examine larger scale phenomena (Rand & Rust, 2011). For example, if a single Twitter user gives extensive information for an event while the remaining users post just one tweet (as in Mendoza et al.'s (Mendoza et al., 2010) work); ABSS allows this special user to be modeled.

Works studied above do not use ABSSs except for Weng et al. paper (Weng et al., 2013), i.e. question five has "no" as an answer in table 1. However, there are a few works in this line as the one by Tripathy et al. (Tripathy et al., 2010). These authors present a study and an evaluation of rumor-like methods for combating the propagation of rumors on social networks. They use variants of the independent cascade model (Weng et al., 2013) for misinformation spread. Besides, the authors criticize epidemic diffusion models such as SIS and SIR because, among others, anti-rumors can be disseminated from person to person unlike vaccines for viruses which can only be administered to individuals. Tripathy et al. also propose an anti-rumor strategy which consists of embedding agents called *beacons* in the network which detect gossips and propagate anti-rumors. In this paper, the spread model and anti-rumor strategy baselines are reproductions of Tripathy et al.'s work. Liu and Chen (Liu & Chen, 2011) build an agent-based rumor propagation model using SIR as baseline and implemented in NetLogo (Tisue & Wilensky, 2004), a popular ABSS framework. This model is not founded on real data although the authors find out interesting conclusions with regard to the Twitter case using the simulation model. Seo et al. (Seo et al., 2012) present a simple ABSS based on gathering retweets (not necessarily rumors), getting the largest connected component in the network, and calculating the retweet probability of each edge $x \rightarrow y$ with the number of retweets given in that edge. More than the simulation, the contribution rests on the use of this model to evaluate a method to identify hearsay and their sources by injecting special nodes called *monitors* (which are very similar to the beacon nodes of our baseline approach (Tripathy et al., 2010)). Yang et al. (Yang et al., 2003) employ ABSS to analyze the 2013 Associated Press hoax incident [3]. The authors give three profiles for Twitter users (broadcaster, acquaintances, and odd users); probability density functions for each profile; and a study of the effects of removing relevant network nodes in the information spread. The authors conclude that removing the node of the highest *betweenness centrality* (Rodriguez et al., 2015) has the optimal effect

[3] On April 23 2013, the Associated Press Twitter account was hacked and a malicious message was sent stating that the White house had been attacked and President Obama was injured.

in reducing the spread of the malicious messages. Gatti et al. (Gatti et al., 2013) address the general information diffusion modeling instead of the gossip diffusion. These authors explore President Obama's Twitter network as an egocentric network and present an ABSS approach where each agent behavior is determined by the Markov Chain Monte Carlo simulation method. As in other works revised (Yang et al., 2003), simulation is employed to find users with more impact on the information flow.

The last works revised present significant contributions in the use of ABSS to study information diffusion in Twitter and have been studied in depth for the current contribution. Nonetheless, as shown in table 1, the efforts in reproducibility are quite questionable. None of them give: the data the results are based on, the simulation implementation, or the source code (three last questions in the table). This hinders researchers from verifying the results or reusing these works in their research or developments. Furthermore, the works also lack general methods to conduct ABSS researches in this scope.

2 Agent-based model and marketing strategies design

This section details the agent-based model design and marketing strategies design tasks of the method presented. These tasks are given for the general problem of viral marketing strategies in Twitter and the specific case study of rumor spread and control.

2.1 Agent-based model design

As explained, the common decisions in agent-based model design are (Rand & Rust, 2011): scope of the model, agents definition, agents' properties, agent's behaviors, environment, time step, and input and output. These are detailed below:

(1) *Scope of the model.* This is the part of the target system the model is focused on and what aspects can be ignored. The idea of developing a model is that it should be as simple as possible so as to study it easily, but at the same time, the model must describe reality. Therefore, there is a trade-off between the KISS approach ("Keep it Simple Stupid", or "Keep it Short and Simple" in a more polite manner) and the KIDS approach ("Keep it Descriptive Stupid") (Serrano & Botia, 2013). In the authors' experience, a major mistake when using ABSS research methods is trying to model the world instead of focusing on just the research relevant aspects. Regarding the Twitter case, there are a number of phenomena that

can be interesting for marketing purposes: retweets propagation, number of mentions, number of tweets with a specific hashtag, activity per time zone, etcetera. Nonetheless, concerning the misinformation propagation, the mainstream approach is not modeling the messages (or tweets) evolution, but users' state evolution regarding a specific rumor or contra-rumor as shown in section 1. These states usually follow the epidemiological terminology: infected, cured, etcetera.

(2) *Agents.* Another important decision is what the agents represent in the ABSS. Note that agents do not necessarily mean smart agents (Nwana, 1996) capable of, among others, learning; or deliberative agents (Woolridge, 2001) which make decisions using symbolic reasoning. The typical ABSS agent is a reactive agent which interact with others autonomously based on a behavior model that can vary from production rules to machine learning models such as artificial neural networks (Campuzano et al., 2015). In the Twitter case, the straightforward decision is to have agents per each Twitter user. Moreover, in the gossip diffusion case, there usually are special users capable of executing marketing strategies such as releasing counter rumors.

(3) *Properties.* These are the fields that describe each agent. Again, these completely depend on the scope of the model. For the Twitter rumor spread case, as explained, typical properties include: an identifier; a position in the environment (explained below); the agent's state with respect to the rumor (infected, cured, vaccinated, etcetera); and, if needed, and agent type field which can determinate other properties scope or the agent behavior.

(4) *Behaviors.* Agents exhibit a behavior which involves interacting with the environment and other agents each time step. Commonly, these behaviors are stochastic processes which depend on given probabilities. There are countless manners of defining behaviors in ABSS: production rules (Serrano et al., 2014), machine learning models (Serrano et al., 2013), probability density functions (Garcia-Valverde et al., 2012), etcetera. For the Twitter misinformation propagation case, the main approach observed in the specialized literature (see section 1) is to define textually the agent behavior. This always leads to gaps when programming the specified model. Thus, the authors recommend the use of pseudo-code or any other general software modeling technique for general software such as UML diagrams. The use of flow diagrams, which roughly correspond to UML activity diagrams, are very popular in the ABSS literature (Gilbert & Troitzsch, 2005).

(5) *Environment.* The environment defines the agents' interaction topology. For Twitter works, the environment is widely described as a network or graph[4] where nodes represent users. Other kinds of networks are also

[4] The terms graph (composed of vertices and edges or arcs) and network (composed of nodes and links) are used interchangeably in this paper. In social sciences

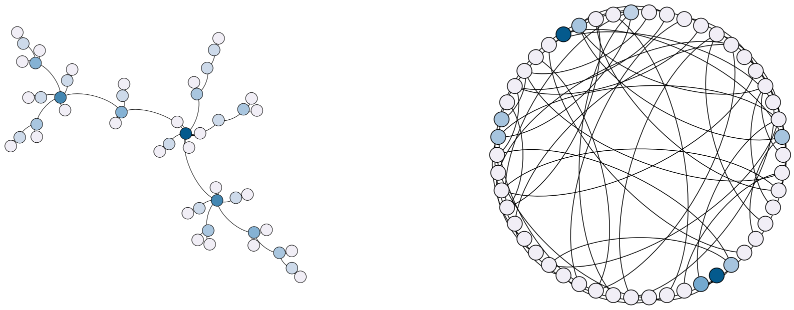

Fig. 1. On the left, *Barabási Albert* (BA) scale-free synthetic network example. Darker colour for higher node degree. On the right, *Watts Strogatz* (WS) small–world network. Darker colour for higher clustering coefficient.

possible depending on the scope, such as networks of retweets. The links do not have the same meaning in all works either. While some authors represent the asymmetry in Twitter (user u_1 follows u_2 does not mean that u_2 follows u_1), others consider undirected links since Twitter has mechanisms to make information flow from the follower to the followed (such as responses, mentions, retweets, and private messages).

Another decision is whether to use a real network or a synthetic one. Regarding the use of real networks (the ones with nodes and links extracted from Twitter), there are two main approaches. The first one is using *egocentric networks* which examine only immediate neighbors from a seed node and their associated interconnections. Gatti et al.'s (Gatti et al., 2013) work, revised in section 1, follows this approach using President Obama's Twitter account as seed. The second approach for using real networks is: gathering tweets independently of their authors, gathering these authors' followers and friends [5], and using the biggest connected component in the resulting graph as final network. As explained in section 1, this approach is used in Seo et al.'s work (Seo et al., 2012).

Concerning the use of synthetic networks, *Barabási Albert* (BA) scale-free networks are the most popular option when modeling social networks (Liu & Chen, 2011). Although the scale-free nature of a large number of networks is still debated by the scientific community, social networks such as Twitter are widely claimed to be scale-free. In a nutshell, the creation of these networks is undertaken under the assumption that the

literature, the terms actors and ties or relations are also broadly used for the network/graph elements.

[5] The Twitter API terminology is followers and friends where friends mean followed users.

probability a user u_1 connects to another u_2 depends on the number of connections that u_2 already has. This makes *hubs* appear, i.e. nodes with a degree that greatly exceeds the average degree. See a BA graph example in figure 1 on the left. Another option for the rumor case is the use of *Watts Strogatz* (WS) small-world networks, as in the work presented by Tripathy et al. (Tripathy et al., 2010), where: if u_1 is connected to u_2 and u_2 is connected to u_3; u_1 and u_3 are likely to be also linked. This makes these networks have a high *clustering coefficient*, i.e. a measure of the degree to which nodes in a graph tend to cluster together. See a WS graph example in figure 1 on the right. As in agents' behavior description, a common drawback in the literature is not giving enough information to reproduce these synthetic networks. The general model, the algorithm to generate it, and the algorithm parameters; are required to ensure reproducibility.

Although using realistic networks is always desirable, there is a clear reason for the hegemonic line of using synthetic networks: realistic networks restrict users studied; and this restricts the number of tweets to those sent by these users. Therefore, if the research depends on tweets semantic as in rumor spread, the use of realistic networks leads to having fewer messages to work with. As a result, the realism of the simulated users is considerably inferior. An extra reason to use the synthetic networks explained for the gossip propagation case is that information spreads very fast in them (Abraham et al., 2010). More specifically, WS networks have a very low *average path length*, i.e. the average number of steps along the shortest paths for all possible pairs of network nodes; and, BA networks average path length grows very little with the number of nodes because of the hubs. As a consequence, these networks are stronger adversaries for gossip control strategies and provide researchers with a baseline by assuming that the strategies performance is better than in real networks.

(6) *Time step.* ABSSs typically evolve in time using a time step. Two phases are distinguished: initialization, when the agents and the environment are created; and, iteration, where agents act according to their behavior model. Moreover, depending on the scope and the real data available, the time step will represent a different physical time unit. For the Twitter case, the data scraping and exploratory data analysis tasks (see section 3) can give insights into this decision. If the data is very scattered (e.g. days pass between relevant tweets), there is not enough information for a short time step (e.g. simulating hours).

(7) *Input and output.* The parameters and the values observed in the simulation execution are other decisions for the model. One of the most important and commonly used input in ABSS is the random seed. As seen, an ABSS involves a number of stochastic processes: selecting the order of agents execution at a time step, creating a network model, deciding among possible actions in the behavior model, etcetera. A major flaw in

ABSS research is not ensuring that all these processes depend on a single random seed, losing the simulation repeatability and reproducibility. Concerning the output in the Twitter misinformation propagation case, the common output is the number of agents per possible rumor state (infected, cured, etcetera) and per time step.

2.2 Marketing strategies design

In relation to the possible marketing strategies in Twitter, most of them rest on having one or several Twitter users representing your brand (meaning your product/service/company), its collaborators, or people after receiving some incentive. These strategic or seed users have to initiate the process of awareness diffusion by propagating the information to their friends via their social relationships (Long & Wong, 2014). In consequence, the classic target is to maximize (advertisement case) or minimize (malicious rumors case) the information spread while the minimum of these strategic users are created or selected. Some examples in the rumor propagation literature are the *beacons* agents proposed by Tripathy et al. (Tripathy et al., 2010) or the *monitors* introduced by Seo et al. (Seo et al., 2012).

The explained problem usually leads to the question of what are the most important users in the network. This can be studied with Twitter specific metrics such as the number of mentions and retweets as proposed by Cha et al. (Cha et al., 2010); or *centrality* metrics (Abraham et al., 2010), indicators which identify the most important vertices within a graph such as the outdegree and the indegree (which would be followers or friends in Twitter). In both cases, accurate "importance" metrics may not be retrievable and some approximations might be used. For example, the Twitter API does not allow recovering the number of retweets or mentions for a user. Furthermore, centrality metrics such as the *closeness* requires the whole (and static) network. In the examples given above of strategic agents, Tripathy et al. (Tripathy et al., 2010) only consider random positions for the *beacons* while Seo et al. (Seo et al., 2012) study the effects of having *monitors* in positions with different centrality indicators such as the *betweenness*.

As explained above, a marketing strategy deals with the creation of special agents in this context; and, therefore, the guidelines to model general Twitter users agents with ABSS given in section 2.1 are also valid for modeling of these strategies. This work contemplates marketing strategies as an extra task in the method; i.e. those what-if scenarios that ABSS allows to understand, evaluate, and predict. The reason is that a single model of the market should be valid for evaluating a number of strategies over it (and vice versa). Nonetheless, another common mistake in ABSS research is assuming that first the reality is modeled and then experts can decide any kind of what-if scenario to be evaluated over the model. As stated in the model scope decision of section 2.1, modeling only relevant aspects of the market is the key in ABSS and, therefore, there is a strong coupling between marketing strategies and the simulated market.

3 Data scraping and exploratory data analysis

This section deals with the data scraping and exploratory data analysis tasks of the method presented.

3.1 Data scraping and pre-processing

As stated in the introduction, although there are extensive works in Twitter data analysis such as Russell's books (Russell, 2011a, 2011b), to the best of the authors' knowledge, this is the first research work where guidelines are given to use Twitter data in an ABSS research and discuss the Twitter API (*Twitter REST API documentation website*, 2015) limitations in this regard. The Twitter REST APIs provide programmatic access to read and write Twitter data. A very convenient manner of getting familiar with this API is using its console (*Twitter REST API console website*, 2015). This console allows a web browser to send requests to the Twitter API with one of several API methods such as "/search/tweets.json". A large number of these methods require users to choose an authentication method. Once the correct URLs of the requests are validated in the console, using a general programming language to program a script which requests the wanted data is straightforward. The authors have experimented with the *Python* programming language and, more specifically, with its "*requests*"[6] and "*requests-OAuthlib*"[7] libraries. The former library, requests, provides developers with very simplified operations for http requests; while the latter provides requests with support for *OAuth*, i.e. the authorization standard used in the Twitter API.

The most relevant API methods for data scraping in Twitter and their limitations as discussed below:

- *search/tweets.json*. It searches and returns tweets and their associated information in json format. This information includes: tweet id, tweet text, hashtags, number of retweets, geographical information, the author's public information in json format, etcetera.

[6] Requests website: `http://docs.python-requests.org/en/latest/`
[7] Requests-OAuthlib website: `https://requests-oauthlib.readthedocs.org/en/latest/`

The main parameter is a query phrase which may or not include hashtags. However, recovering tweets in a temporal window is not straightforward because, although there is an *until* parameter for getting tweets before a given date, there is no *since* parameter. This problem can be relieved by using the *since_id* parameter, which gets tweets before a given tweet id.

A possible scraping strategy is employing the *result_type* parameter to obtain only popular tweets and then extend the dataset with retweets of these seed tweets. Nonetheless, the popular results are much reduced: the API currently returns only 100 popular tweets. Moreover, retrievable retweets from a specific tweet are also limited to 100 or less.

Other important constraints are: the API Rate Limits, 180 queries per 15 minute in API 1.1.; and the days during which tweets are available, the last 6 to 9 days of tweets at the moment (*Twitter REST API documentation website*, 2015). Several users accounts with their authentication parameters can be employed to parallelize the Twitter scraping process.

In a nutshell, the search API is focused on relevance and not completeness (*Twitter REST API documentation website*, 2015). As a result, when obtaining tweets of a very specific topic such as a gossip, it is not possible to know if these tweets temporal distribution is significant. The Twitter streaming API[8] is recommended for completeness. Nonetheless, this stream API returns a great deal of not relevant data when working with tweets about a specific topic and, more importantly, the historical tweets cannot be retrieved.

- *statuses/retweet/{id}.json*. This operation gets up to 100 retweets of a given tweet id. Therefore, it can be used to extend any Twitter dataset if tweets ids are included. Besides, the retwitting user's data is also given in the response. The main limitation is that it only retrieves a little percentage of the total retweets (up to 100). Moreover, this method only allows 15 requests each 15 minutes (*Twitter REST API documentation website*, 2015). Consequently, getting retweets is considerably more time consuming than getting tweets with *search*. An advantage of getting Twitter data from tweets or users ids with this or the following operations is that this data can be recovered after the maximum 9 days that the *search* operation gives.
- *statuses/show/{id}.json*. The current Twitter terms of use (*Twitter terms of use website*, 2015) state that "If you provide Content to third parties, including downloadable datasets... you will only distribute or allow download of Tweet IDs and/or User IDs". Although this is further discussed in the terms, a fairly common practice in the research community to only openly distribute lists of tweet ids instead of the raw data. This operation allows researchers to retrieve the tweet and users data from an id. As with the *retweet* operation, this method only allows 15 requests each 15 minutes (*Twitter REST API documentation website*, 2015).
- *followers/ids.json* and *friends/ids.json*. These operations allow researchers

[8] Twitter streaming API website `https://dev.twitter.com/streaming/overview`

to obtain followers and friends from a user's id (which is in the tweets data given with the operations explained above). Therefore, as explained in the environment bullet point of section 2.1, the real network topology may be obtained although this usually means to reduce the number of tweets available to build the ABSS model. As in the *retweet* operation, gathering this data is more restrictive than the *search* method: only 15 request each 15 minutes. Besides, even when the network topology can be retrieved from tweets or users ids, this topology is very dynamic and it is not recommended to gather it after a substantial time lag; e.g. a research is conducted over datasets with users' ids retrieved several months ago.

Preprocessing the large number of json files gathered in the scraping process to build datasets valid for the exploratory data analysis is also required. Again, as in the data scraping phase, Python is the programming language recommended. Python is extensively used for natural language processing tasks because of its simple syntax and its rich text processing tools. The typical fields obtained from each tweet for Twitter ABSS research include: date and time of writing, author, text, and a label with the tweet meaning according to the model scope. As explained above, there are limitations in the Twitter API which hinder researchers from obtaining some possible fields for these datasets. Some examples are: the retweets or mentions for some user, how many tweets a user may have read at some moment, or if the user has been "cured" of a rumor.

Among these datasets fields, the label is typically non-automatically added, making it the most time consuming part of data preprocessing. Since tweets are composed of only 140 characters and may transmit humoristic or sarcastic messages, labeling them can be complex even for human beings. For example, in the misinformation diffusion and control case, the labeling involves deciding if each tweet is a rumor, anti-rumor, or other. Some authors skip the labeling by identifying hearsay or memes in Twitter with hashtags as Weng et al. (Weng et al., 2013); or considering all tweets in a specific temporal window as Yang et al. (Yang et al., 2003). Another option is to use only popular tweets and their retweets, making the labeling process simpler by assuming that the retweets meaning is the same as their original tweets. Nonetheless, this is not necessarily true since some comments can be added in retweets by using *"RT"* between the extra comment and the cited tweet.

3.2 Exploratory data analysis

Once the Twitter data has been retrieved and preprocessed, the exploratory data analysis is suggested to gain insights into the modeled market. When data is studied, new questions arise, causing specific parts to be viewed in more detail and changing or refining the design decisions made in the models. Some interesting data views for modeling Twitter users are:

- Tweets for each type (label), to study if there is enough information to model them or if some types can be joined.
- Number of users per tweets sent, to study the possibility of having different agents' behaviors in function of the information available.
- Tweets per minute / hour / weekday / day of the month, to decide the simulation time step. This also allows observing patterns such as the decrease or increase of messages at nights or in work hours.
- Metrics for experiments. Scraped data is composed of information such as tweets, users attributes, network topology, etcetera. Some further processing may be necessary for comparing the ABSS model output with the real data in the validation experiments.

The Python programming language, recommended for the scraping and preprocessing, is also a powerful tool for the exploratory data analysis through packages such as Pandas and NumPy (Rossant, 2013). However, the authors found that the *R*[9] programming language is more convenient for this specific task. This is because, unlike the aforementioned Python packages, R has interfaces for other programming languages employed for ABSS such as *Java*. This allows the simulation code to be connected to the data analysis code for, among others, comparing the simulation output with the real data.

4 Model construction

As explained, this task consists of translating the model into something which can be used by a computer (i.e. programming the model). There are a large number of agent-based modeling software frameworks (*Agent-based modeling software frameworks list*, 2015) which aid developers to implement these systems. Some of the typical features include: simulation examples; predefined agents; scheduling algorithms; and, support for batch experiments. Some of the most extended platforms for the ABSS development are *NetLogo* (Tisue & Wilensky, 2004), *Mason* (Luke et al., 2004) and *Repast* (North et al., 2006).

Although the aforementioned frameworks have support for modeling and displaying networks, the *social network analysis* (SNA) tools (Abraham et al., 2010) are clearly superior in displaying and studying large-scale graph. This is, among others, because of: the ready-to-use centrality metrics, which identify the most important vertices within a graph; the community detection methods, which can be used to split the network in well defined groups densely connected

[9] The R Project for Statistical Computing website: `www.r-project.org`

internally; and, more importantly, the variety of force-directed graph drawing algorithms that these packages implement, allowing a more understandable network visualization. Understandable simulation displays are crucial because they provide the basic mechanism to verify the ABSS, i.e. checking that it meets the model specification and requisites. Some popular SNA frameworks (*Social network analysis software*, 2015) are: Gephi, iGraph, GraphStream, and NetworkX.

One of the most important and commonly forgotten ABSS implementations requisites is to offer repeatability and reproducibility. In an ABSS, there are a number of stochastic processes such as: the agents' initial positions and states; the order the agents gain the turn of execution; the order the agents neighbors are visited; or, the agents' state evolution. When combining different software packages such as ABSS and SNA frameworks, developers have to ensure a single and parametrized random seed to control these processes. A recommended practice is the use of *unit testing*. This allows developers to automatically test that each individual simulation unit does not return different outputs when the same random seed is used as input.

Acknowledgments

This research work is supported by the Spanish Ministry of Economy and Competitiveness under the R&D project CALISTA (TEC2012-32457); by the Spanish Ministry of Industry, Energy and Tourism under the R&D project BigMarket (TSI-100102-2013-80); and, by the Autonomous Region of Madrid through the program MOSI-AGIL-CM (grant S2013/ICE-3019, co-funded by EU Structural Funds FSE and FEDER).

References

Abraham, A., Hassanien, A.-E., & Snasel, V. (2010). Computational social network analysis. *Computational Social Network Analysis, Computer Communications and Networks*, *1*. Retrieved 2014-02-25, from http://link.springer.com/content/pdf/10.1007/978-1-84882-229-0.pdf

Agent-based modeling software frameworks list. (2015). http://en.wikipedia.org/wiki/ABM_Software_Comparison. (Accessed July 20, 2015)

Campuzano, F., Garcia-Valverde, T., Botia, J. A., & Serrano, E. (2015). Generation of human computational models with machine learning. *Information Sciences*, *293*(0), 97 - 114. Retrieved from http://

www.sciencedirect.com/science/article/pii/S0020025514009049 doi: http://dx.doi.org/10.1016/j.ins.2014.09.008

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in twitter: The million follower fallacy. In *4th international aaai conference on weblogs and social media (icwsm).* Retrieved from http://scholar.google.de/scholar.bib?q=info:rqhbqWEH79kJ:scholar.google.com/&output=citation&hl=de&as_sdt=0&ct=citation&cd=10

De Domenico, M., Lima, A., Mougel, P., & Musolesi, M. (2013, October 18). The Anatomy of a Scientific Rumor. *Scientific Reports*, *3*. Retrieved from http://dx.doi.org/10.1038/srep02980 doi: 10.1038/srep02980

Flentge, F., Polani, D., & Uthmann, T. (2001). Modelling the emergence of possession norms using memes. *J. Artificial Societies and Social Simulation*(4). Retrieved from http://dblp.uni-trier.de/db/journals/jasss/jasss4.html#FlentgePU01

Garcia-Valverde, T., Campuzano, F., Serrano, E., Villa, A., & Botia, J. A. (2012, August). Simulation of human behaviours for the validation of ambient intelligence services: A methodological approach. *Journal of Ambient Intelligence and Smart Environments*, *4*(3), 163–181. Retrieved from http://dl.acm.org/citation.cfm?id=2350776.2350778

Gatti, M. A. d. C., Appel, A. P., dos Santos, C. N., Pinhanez, C. S., Cavalin, P. R., & Neto, S. B. (2013). A simulation-based approach to analyze the information diffusion in microblogging online social network. In *Proceedings of the 2013 winter simulation conference: Simulation: Making decisions in a complex world* (pp. 1685–1696). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2675983.2676193

Gilbert, N., & Troitzsch, K. G. (2005). *Simulation for the social scientist.* Open University Press. Hardcover. Retrieved from http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0335216013

Gupta, A., Lamba, H., & Kumaraguru, P. (2013, September). $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on twitter. San Francisco, CA. Retrieved from http://precog.iiitd.edu.in/Publications_files/ecrs2013_ag_hl_pk.pdf

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on world wide web companion* (pp. 729–736). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2487788.2488033

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, *42*, 599–653.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis* (pp. 8:1–8:9). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2501025.2501027 doi: 10.1145/2501025.2501027

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In H. Xiong, G. Karypis, B. M. Thuraisingham, D. J. Cook, & X. Wu (Eds.), *2013 IEEE 13th international conference on data mining, dallas, tx, usa, december 7-10, 2013* (pp. 1103–1108). IEEE Computer Society. Retrieved from http://dx.doi.org/10.1109/ICDM.2013.61 doi: 10.1109/ICDM.2013.61

Liu, D., & Chen, X. (2011). Rumor propagation in online social networks like twitter – a simulation study. In *Proceedings of the 2011 third international conference on multimedia information networking and security* (pp. 278–282). Washington, DC, USA: IEEE Computer Society. Retrieved from http://dx.doi.org/10.1109/MINES.2011.109 doi: 10.1109/MINES.2011.109

Long, C., & Wong, R. C.-W. (2014). Viral marketing for dedicated customers. *Information Systems*, *46*(0), 1 - 23. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306437914000751 doi: http://dx.doi.org/10.1016/j.is.2014.05.003

Luke, S., Cioffi-Revilla, C., Panait, L., & Sullivan, K. (2004). Mason: A new multi-agent simulation toolkit. In *Proceedings of the 2004 swarmfest workshop.*

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics* (pp. 71–79). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1964858.1964869 doi: 10.1145/1964858.1964869

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653 - 7670. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417414003455 doi: http://dx.doi.org/10.1016/j.eswa.2014.06.009

Nekovee, M., Moreno, Y., Bianconi, G., & Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 457 - 470. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437106008090 doi: http://dx.doi.org/10.1016/j.physa.2006.07.017

North, M. J., Collier, N. T., & Vos, J. R. (2006, January). Experiences creating three implementations of the repast agent modeling toolkit. *ACM Trans.*

Model. Comput. Simul., *16*(1), 1–25. Retrieved from http://doi.acm.org/10.1145/1122012.1122013 doi: 10.1145/1122012.1122013
Nwana, H. S. (1996). Software agents: An overview. *Knowledge Engineering Review*, *11*, 205–244.
Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1589–1599). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145602
Rand, W., & Rust, R. T. (2011, September). Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, *28*(3), 181–193. Retrieved 2014-02-27, from http://linkinghub.elsevier.com/retrieve/pii/S0167811611000504 doi: 10.1016/j.ijresmar.2011.04.002
Rodriguez, A., Kim, B., Lee, J.-M., Coh, B.-Y., & Jeong, M. K. (2015). Graph kernel based measure for evaluating the influence of patents in a patent citation network. *Expert Systems with Applications*, *42*(3), 1479 - 1486. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417414005338 doi: http://dx.doi.org/10.1016/j.eswa.2014.08.051
Rossant, C. (2013). *Learning IPython for Interactive Computing and Data Visualization.* Packt Publishing.
Russell, M. A. (2011a). *21 Recipes for Mining Twitter* (1st ed.). O'Reilly Media. Paperback. Retrieved from http://oreilly.com/catalog/0636920018261
Russell, M. A. (2011b). *Mining the Social Web* (1st ed.). O'Reilly Media. Paperback. Retrieved from http://oreilly.com/catalog/0636920018261
Seo, E., Mohapatra, P., & Abdelzaher, T. (2012). Identifying rumors and their sources in social networks. Retrieved from +http://dx.doi.org/10.1117/12.919823 doi: 10.1117/12.919823
Serrano, E., & Botia, J. (2013). Validating ambient intelligence based ubiquitous computing systems by means of artificial societies. *Information Sciences*, *222*(0), 3 - 24. Retrieved from http://www.sciencedirect.com/science/article/pii/S0020025510005578 doi: 10.1016/j.ins.2010.11.012
Serrano, E., Poveda, G., & Garijo, M. (2014). Towards a holistic framework for the evaluation of emergency plans in indoor environments. *Sensors*, *14*(3), 4513–4535. Retrieved from http://www.mdpi.com/1424-8220/14/3/4513 doi: 10.3390/s140304513
Serrano, E., Rovatsos, M., & Botía, J. A. (2013). Data mining agent conversations: A qualitative approach to multiagent systems analysis. *Information Sciences*, *230*(0), 132 - 146. Retrieved from http://www.sciencedirect.com/science/article/

pii/S002002551300011X doi: 10.1016/j.ins.2012.12.019

Shah, D., & Zaman, T. (2011). Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory*, *57*(8), 5163–5181. Retrieved from http://dx.doi.org/10.1109/TIT.2011.2158885 doi: 10.1109/TIT.2011.2158885

Social network analysis software. (2015). http://en.wikipedia.org/wiki/Social_network_analysis_software. (Accessed July 20, 2015)

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. (In iConference 2014 Proceedings (p. 654 - 662))

Tisue, S., & Wilensky, U. (2004). NetLogo: A Simple Environment for Modeling Complexity..

Tripathy, R. M., Bagchi, A., & Mehta, S. (2010). A study of rumor control strategies on social networks. In *Proceedings of the 19th acm international conference on information and knowledge management* (pp. 1817–1820). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1871437.1871737 doi: 10.1145/1871437.1871737

Twitter REST API console website. (2015). https://dev.twitter.com/rest/tools/console. (Accessed July 20, 2015)

Twitter REST API documentation website. (2015). https://dev.twitter.com/overview/documentation. (Accessed July 20, 2015)

Twitter terms of use website. (2015). https://dev.twitter.com/overview/terms/policy. (Accessed July 20, 2015)

Valecha, R., Oh, O., & Rao, H. R. (2013). An exploration of collaboration over time in collective crisis response during the haiti 2010 earthquake. In R. Baskerville & M. Chau (Eds.), *Proceedings of the international conference on information systems, ICIS 2013, milano, italy, december 15-18, 2013.* Association for Information Systems. Retrieved from http://aisel.aisnet.org/icis2013/proceedings/ResearchInProgress/96

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013, August). Virality prediction and community structure in social networks. *Scientific Reports*, *3*. Retrieved 2014-02-25, from http://www.nature.com/srep/2013/130828/srep02522/full/srep02522.html doi: 10.1038/srep02522

Woolridge, M. (2001). *Introduction to multiagent systems.* New York, NY, USA: John Wiley & Sons, Inc.

Yang, S. Y., Liu, A., & Mo, S. Y. K. (2003). *Twitter financial community modeling using agent based simulation* (SSRN Scholarly Paper). Rochester, NY. Retrieved from http://papers.ssrn.com/abstract=2358523 (IEEE Computational Intelligence in Financial Engineering and Economics, London, 2013)

Zhao, L., Cui, H., Qiu, X., Wang, X., & Wang, J. (2013). {SIR} rumor spreading model in the new media age. *Physica A: Statistical Mechanics and its Applications*, *392*(4), 995 - 1003.

Retrieved from http://www.sciencedirect.com/science/article/pii/S037843711200934X doi: http://dx.doi.org/10.1016/j.physa.2012.09.030

GNU Free Documentation License

Version 1.3, 3 November 2008

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The “**publisher**” means any person or entity that distributes copies of the Document to the public.

A section “**Entitled XYZ**” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “**Acknowledgements**”, “**Dedications**”, “**Endorsements**”, or “**History**”.) To “**Preserve the Title**” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document’s license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
D. Preserve all the copyright notices of the Document.
E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document’s license notice.
H. Include an unaltered copy of this License.
I. Preserve the section Entitled “History”, Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled “History” in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the “History” section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled “Acknowledgements” or “Dedications”, Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
M. Delete any section Entitled “Endorsements”. Such a section may not be included in the Modified Version.
N. Do not retitle any existing section to be Entitled “Endorsements” or to conflict in title with any Invariant Section.
O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version’s license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.