A hierarchical network model for epidemic spreading. Analysis of A/H1N1 virus spreading in Romania

Silvia Rausanu[1] and Crina Grosan[23*]
[1]ISDC Cluj-Napoca, Romania
[2]Department of Information Systems and Computing
Brunel University London, UK
[3]Department of Computer Science
Babes-Bolyai University Cluj-Napoca, Romania

e-mail: silvia.rausanu@gmail.com, crina.grosan@brunel.ac.uk
[*]corresponding author

Abstract. The research in this paper presents a new approach for the modelling of epidemic spread by using a set of connected social networks. The purpose of this work is to simulate the spreading of the well know A/H1N1 pandemic virus. The case study analyzed in this paper refers to the spreading of A/H1N1 in Romania. The epidemic is followed from its beginning throughout its evolution in Romania, i.e. between May 2009 and February 2010. The evolution is performed in a hierarchical way, taking into account the state divisions, the influences among them, national level as well as influences from abroad (from other infected countries). Numerical experiments performed analyze the monthly evolution of the infection in each county and at the country level and compare the results with the real ones (collected during and at the end of the epidemic spread). The simulations results are closer to the reality than the ones provided by the Health Ministry in Romania.

I. INTRODUCTION

The appearance of the A/ H1N1 virus is dated at the end of April 2009 in Mexico, ever since this virus has spread around the globe with an amazing speed. The virus reached Romania only on the 23rd of May 2009. As this virus can be contacted by air, the spreading among the countries was facilitated by means of transportation, not only among the countries, but also inside them.

In the case of Romania, the virus has been contacted from abroad, firstly in the main cities of the country, then, following the same rule of moving masses of population, in almost each county. In addition to these factors, the time in which the spread got a higher speed was in the months when the likelihood of infection was greater and the crowding in different areas was much more common. These factors can form a pattern in the spread of the virus at any level of community: inside a county or inside a country. However, in Romania the spread did not respect entirely the patterns, nor the expectations of the Health Minister of Romania, as some counties, during the pandemic development, has no single case of infection with the A/H1N1 virus.

The current paper exposes a new perspective for the A/H1N1 epidemic spread simulation in Romania. A similar purpose has been established by the researchers in Vietnam, but using a different model for their simulation [3]. Social networks are a good model for epidemic spread simulations [1], but tend to have no connection to a real case and use only the basic definition of social networks, without any modifications according to the requirements of the experiment.

A model that proves appropriate for the requirements of the experiment and suitable for the observed patterns consists of a combination of existing models. The new model contains elements from the theoretical social networks, statistical network models, hierarchical networks, and epidemiological models. The choice for each of them came in a natural way: having insight the evolution of a human formed community in special conditions where the relations between the individuals inside the community can influence the future state of the entire group, social networks are the solution. The purpose of the application and the algorithm have imposed further theoretical grounds for the model: the individuals in a country are structured at an outer level in smaller communities, counties, forming themselves a social network and at the same time being decomposable in other smaller social networks; in other words, the whole group of individuals in a country is modeled as a hierarchical network with two levels, country level and counties level. The impossibility of tracking millions of connections for the individuals brought to the

decision of using generative random models based on statistical measures of a network. Aiming to simulate an epidemic, clearly, the epidemiological models played a role, being a base for the epidemiological states used in the algorithm.

The paper is organized as follows: in Section 2 the proposed model is presented in detail. Section 3 presents the simulations followed by Section 4 dedicated to experiments. Section 5 contains conclusions and future work ideas.

II. DESCRIPTION OF THE PROPOSED MODEL

For our case study, the spreading can be simulated over a network modeled at the level of the counties, taking into account all their connections with other counties in Romania and other countries affected by this epidemic. The final model is formed by a network of networks (or a *hierarchical network*), each county corresponding to a node in the big network, and at the same time it contains an inner network representing the network of inhabitants of the county.

A. Hierarchical networks

The simulation is performed at two levels, one being the counties level and the other the individual level. The individuals are tightly connected to the county they belong to, their individual evolution being influenced by the same characteristics as the county is, but in different proportions. One individual can be connected to one county only, in this way a more complex network is created (as in Figure 1), although there is no homogeneity of the nodes' types and relevance. The factors of influence in the spreading are transmitted from the county in general, to each individual located in it, in particular. This transmission order classifies the network in two hierarchies, corresponding to the levels of spreading simulations; in this way the network for spreading fits the model of the hierarchical social networks [8]. However, the flow of transmission is bidirectional, not only a county distributes its characteristics, but also the mass of individuals contribute to the final statistics computed for a county.

Having this interdependence, the two hierarchies can be isolated and separated physically, but sharing the context of spreading and keeping the exchange of information during the simulation.

.

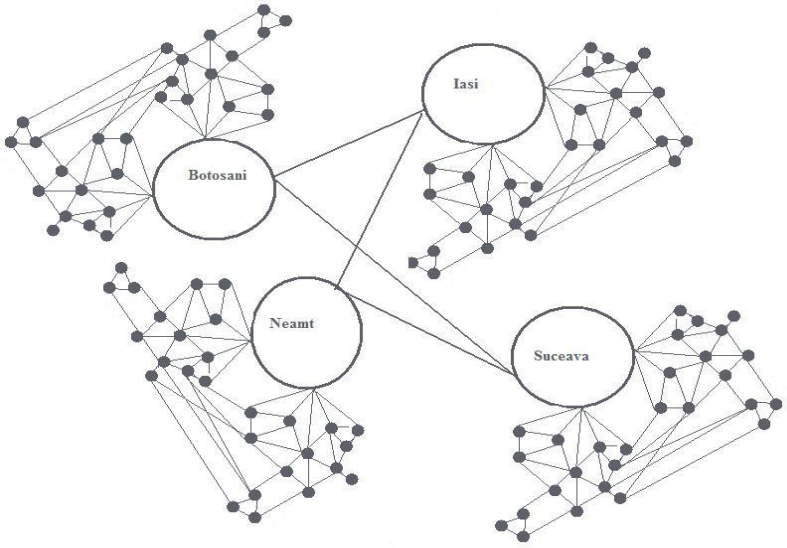

Figure 1. The hierarchical network model.

.

B. Outer network

The outer network is the network composed of the counties of Romania, which could be visualized as a map of the country. Although a social network has usually sole individuals as nodes, a group or an organized formation of individuals can be at the basis of a network, in our case the mass of inhabitants of a certain country. The connections

between these nodes are created based on many criteria, fact that changes the structure of the network from a simple graph – directed or undirected – to a directed multigraph as between two nodes exist more than one type of links:

$$V = \{v \ni nodes\}$$
$$E = \{(u,v) | u,v \ni V\}$$
$$f(s): E \to \{\{u,v\} | u,v \ni\}$$
$$f(r) = f(s) \Rightarrow r \downarrow\downarrow s$$

In this way, there are some properties assured to the network, such as general connectivity and, moreover, the network fits the model of *scale-free networks* [1][2][5] and the *small-world model* [4][12][15].

The outer network suits the scale-free model as it gathers its most important features. Firstly, the existence of hubs is underlined by some counties that tend to have lots of connections with the other nodes; these nodes have actually a high importance not only in the network but also in the country (collegial centers, main city, tourism nodes, etc). Secondly, the connectivity of the network is easily assured only by one type of connection: the neighborhood between counties, which makes the multigraph underlying it strongly connected.

The small-world property is fulfilled not by taking into consideration a possible increase of the network (over the entire process of simulation the size of the network will remain constant) but by the relative "closeness" of each nodes, following the "six degree of separation" [10], again, through the multi-typed set of edges. The size of 41 (the number of counties in Romania) is constant and relatively small, encouraging the property to preserve its validity throughout the simulation, regardless of the operations on the network.

As it was stated before, the nodes forming the network will be the counties of Romania and besides them, some countries with which Romania has common borders or to which Romania has various kinds of connection and exist many air routes between them. The last type of nodes have a small but vital role in the network as there will be no spreading simulation over them, but they will be taken into account at the time the computation of some indexes will be made for the inner nodes (see Figure 2).

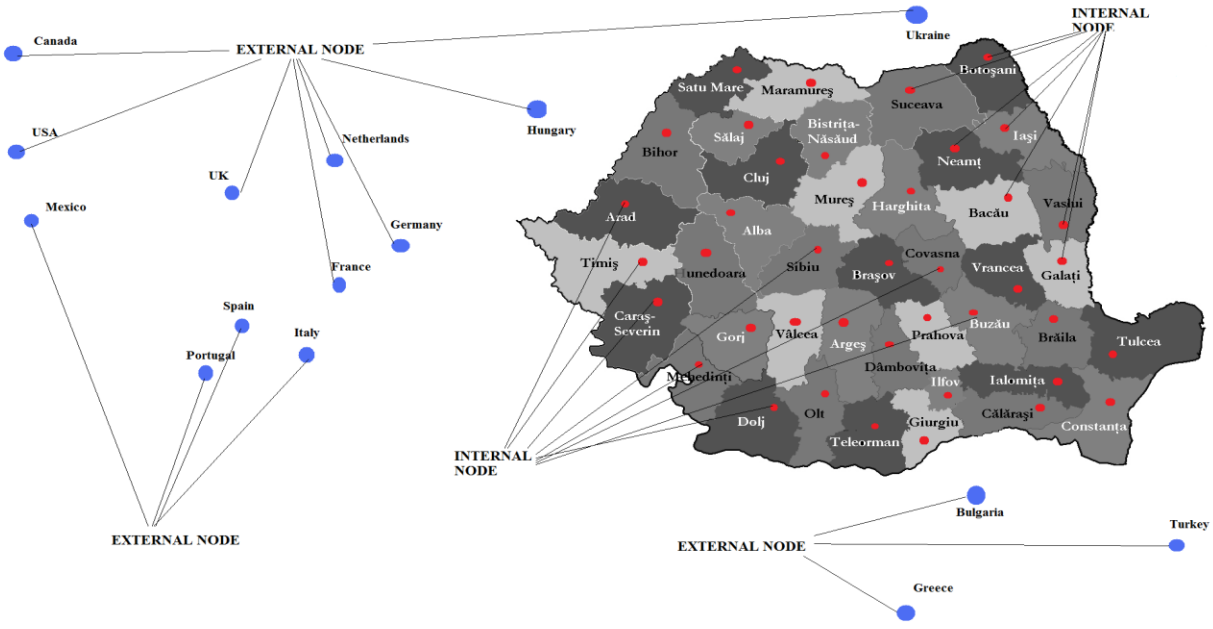

Figure 2 Types and distribution of nodes.

In order to avoid simulating the spread over other nodes than those corresponding to the national counties, a more adequate separation has been used, introducing a characteristic of graph node: type. The inner country nodes are called "internal" and all other ones are called "external".

Although the mathematical approach deals more with the connections between nodes [7][16], the case considered keeps a small part of the former sociologically approach of social networks by storing inside the nodes some characteristics which, during the simulation, will influence the evolution of the nodes in their neighborhood.

The most important attribute of a node is the number of already infected people inside the county. Based on the theory of moving masses of population, the number of infected people represents a percentage of the population and it can be part of the flow of moving individuals between nodes; this is why this value will be included in the computations not only for a current node, but also for all the surrounding nodes.

The population characteristics of an internal node (these characteristics are not considered for the external nodes) have an indirect influence on the evolution of the network during a simulation. The population size and population density give important information for constructing the corresponding inner network which recursively affects the inner evolution respecting the standards imposed by the entire network.

The links between the vertices are formed on some pre-established conditions that imply only characteristics taken from reality. The most important reasons for putting an edge between two nodes are:

- geographical closeness
- collegial surroundings
- nodes of means of transportation – railway, airport
- tourist attractions
- poverty level.

However, these are not the only connections; further connections can be established on the base of commercial centers, economical attraction (working places in relative nearby counties) and others.

As it can be seen from Figure 3, the edges have different characteristics, described in detail in what follows.

Internal Neighborhood type of edge appears between two internal nodes which are geographical neighbors – have a common border. The structure of the network is a directed multigraph and the relation of neighborhood is characterized by reciprocity, consequently, for one relation of this type, there will be two edges: one starting from node *X* to node *Y* and one starting from node *Y* to node *X*. This type of edge has been chosen as it is a well-known fact that between nearby counties there is a constant movement of masses; this means that a certain percentage of healthy or infected population can cross the border between the two counties, facilitating the spread of the virus. The properties of this type of edge remain consistent regardless of the time of the year.

External Neighborhood type of edge appears between one internal node and one external node situated at the border of Romania. Although the relation between these two nodes is reciprocal, there is one edge directed from the external node *E* to the internal node *I* as there is no interest for the evolution of the external node in the current situation. The influence of this type of edges if proven by taking into account the fact that, in some of the external nodes, the evolution of this virus had a more rapid evolution and there is, again, a common thing to have individuals passing the borders from one side to another.

Collegial Neighborhood type of edge appears between two internal nodes, one being a collegial center and the other being situated, usually, in the geographical closeness of that node. There is just one edge attached to this relation as during studies, the individuals from outside the collegial county come, so the edge is oriented from one nearby node towards the collegial node. In this case of population movement there is a greater percentage of population involved, but it is taken into consideration only in the months when the college courses are held.

Railroad Node type of edge appears between two internal nodes having more developed rail connections (rail hubs). The relation is characterized by reciprocity so there are two edges associated to this connection, one starting from one side and one from the other side. In bigger train stations the number of passengers is higher, but, on the other hand, the connection time (in the case of a stopover) gives passengers the chance to gain contact with the surrounding group of individuals around the station. There is no temporary limit of this connection, regardless of the time of the year.

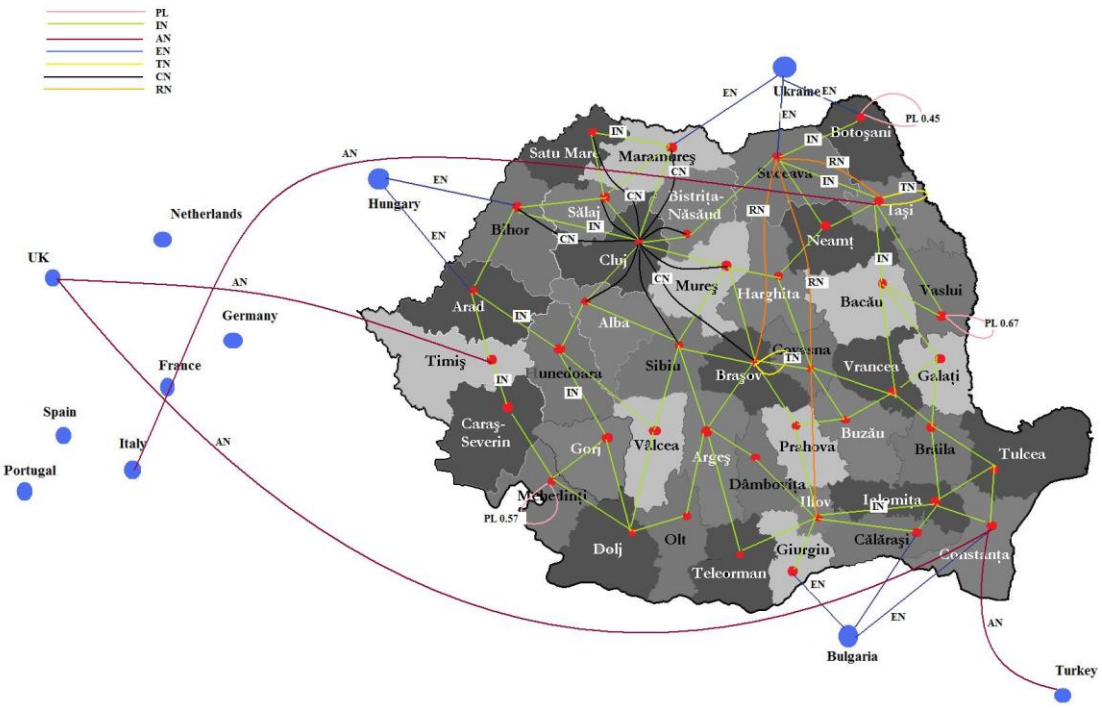

Figure 3 Edge types
IN – Internal Neighborhood, EN – External Neighborhood, CN – Collegial Neighborhood, RN – Railroad Node, PL – Poverty Level, AN – Air Node, TN – Tourism Node

Poverty Level is rather a characteristic of a county – an internal node – but its influence is general, at the level of the entire group of individuals, affecting the power of the spread in the node. The influence of this type of edge is done recursively, and it forms a loop, an edge from vertex *V* to vertex *V*, this being allowed by the multigraph structure. The choice of this type of edges is motivated by the way the virus affects individuals with less decent living standards, who cannot protect themselves properly. This edge has a weight attached to it that represents the percentage of population affected by poverty in a county and further affects the change of the evolution inside the county. This type of edge is the only one that does not involve the movement of masses.

Air Node type of edge is set between an external node and an internal node between which there are a significant number of flights. The edge is directed from the external nodes towards the internal nodes, as only for the last one the simulation is performed. This type of edge relies on the transportation of groups of individuals from the firstly affected countries in the world, this way being actually the one in which the virus was brought to Romania. The internal node is characterized by the existence of an important international airport that makes the movement of a considerable part of population easier.

Tourism Node type of edge implies only one internal node, being, from one point of view, a property of the node, but from another point of view it facilitates a certain movement of masses. There is only one edge from the node *N* to node *N* forming a loop. For this type of edge it is assumed that tourists that impacts on the virus actions over the node form a percentage of the healthy or infected population. The connection is taken into account as a factor of influence during the holiday months only.

Besides the normal components of a social network, nodes and edges, a new component – the time of the year (represented as months) – is considered. For a more accurate simulation, this temporal component is required, as in the autumn and winter months the infection likelihood rises. However, there is another reason for implying some characteristics according to the current month: in the final months of the epidemics, the vaccine against the virus has been used worldwide, consequently, the pandemic no longer spread at the same rate.

C. *Inner network*

The inner network is developed from one internal node of the outer network, inheriting simulated and/or original attributes from it. The creation of this type of network is quite difficult as tracking the connection at the individual level is complicated. This is mostly the reason for which a random model [13] for generating a network involving only the information available from the upper level will be used. We chose the Erdos - Renyi [4] model as it suits the requirements with the minimum amount of computation and analysis.

The nodes of the social network will represent the individuals of a county without containing any extra information. The edges will be simple connections between individuals, generated according to the chosen model. The two parameters required for the construction of the network are received from the corresponding node in the outer network and are:

- the number of nodes in the network and
- the probability that any two nodes are connected.

The probability received is not of much help, but using one of the properties of the Erdos – Renyi model it will be translated into clustering index [4], clusters being the point of interest in the simulation. Seeing clusters as crowds of individuals makes a logical connection with the population density that is known from the very beginning. However, the domains of definition of the two functions are different; consequently the population density must be scaled to the domain of the clustering index which is the interval [0, 1]. A scale of the density will be assumed as follows: the county with the highest density will have the highest clustering index and the rest will be proportional with this one; this implies that the maximum clustering index will be 1 but the highest clustering index between all counties will be in fact around 0.6 as any higher value will transform the network into an almost complete graph [4], situation which is mostly uncommon for large real-world social networks (like a county):

$$\frac{M}{d_i} = \frac{C}{c_i} \Rightarrow c_i = \frac{C \cdot d_i}{M},$$

where:
M denotes maximum density,
d_i is the density of node i,
C is the maximum clustering index and
c_i represents the clustering index of node i.

The other parameter, the number of nodes in the network, is the size of the population of that particular county. We scale this parameter according to the maximum size allowed and the maximum size of population among all counties [17]. Applying this theory, an individual will actually represent n individuals and this could cause problems during simulations when scaling the number of already infected individuals. For example: there is one infected individual, but following the scaling rule, one real individual is 0.2 of one individual in the network, therefore there will be 0.2 individuals infected. Obviously, this method will fail in most of the input cases.

Thus, each inner network will be split into m smaller inner network that can be considered to be independent communities or clusters of individuals inside a county. The number of communities in a county will be equivalent with the population density in that county. However, not all the inner-inner networks will be used for the simulation, but a percentage of them, the same percentage as the number of already infected people represents from the entire population of the county:

$$\frac{Infected}{Population} = \frac{p}{100}$$

$$Infected = Population \cdot p\% \Rightarrow cluster\_no = density \cdot p\%$$

The infected individuals will be distributed almost equally among these communities (see an example in Figure 4). In order to have a distinction between all the clusters, the probability for creating the connections between individuals, received from the upper level, will suffer minor transformations, being considered as a random number in the interval of a small neighborhood of the initial clustering index. Following this theory, we are assured that the infected individuals will be all distributed and the simulation will not be affected by the limitation of the programming environment.

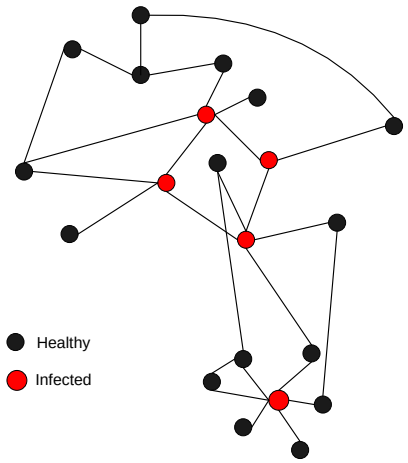

Figure 4 Infected individuals in an inner network. A network with 5 infected nodes.

Having a vast random network, the problem of placing the infected individuals in the network occurs. This issue can be solved either by selecting randomly *n* individuals to carry this special property, or by picking them according to a ranker of size *n*. The ranker can classify the nodes according to some criteria such as:

- betweeness centrality
- closeness centrality or
- degree centrality [9].

We opted for the possibility of selecting the first *n* individuals with the highest degree. In this way, the highest number of acquaintances and a relatively larger power of spreading the infection have been assigned to infected individuals.

III. DETAILS ABOUT SIMULATIONS

The simulations are intended to approximate the evolution of the infected individuals in each county over the course of a month. The complex model used is populated with real data collected from the national health repositories. However, the formulas used do not assure the same accuracy in computing, as they were developed from separate mathematical theoretical foundations and the connections to the subject in cause – viral spreading – were made with some modification over the social network epidemic spreading model [6]. This reasoning over the used formulas and other unwillingly ignored factors are sources of errors that oscillate during the tests.

A. Factors of influence

An epidemic is characterized firstly by the way the virus can be contacted; the easier the virus is contacted, the more factors encourage the epidemic. In the case of A/H1N1 virus, the spreading is done by air, a common and successful medium for an epidemic to pass to a pandemic spreading. The influence of the continuous movement of population masses is sustained by the multiple means of transportation between communities of individuals and, moreover, by the increase of the crowding coefficient that traveling with most of those means involves.

The types of edges represent the ways in which individuals change their current node location to another, carrying along the virus from one infected community to another one still uninfected. Among the percentage of individuals that move from one node to another there might exists infected individuals, and, as the virus is transmitted by air, any short or long contact of that individual can add a new victim to the general statistics. However, this theory is not totally real, depending actually on the particular characteristics of the individual, e.g. the power of its immunity system.

The weather (or season) is another positive factor of influence for the spreading. In the cold months when it is often raining, snowing of wind blowing, the human immunity system fails to keep the same properties as in the warmer months, so the probability of viral infection is increased for the majority of the population.

One factor that independently rises from the context is the apparition of a vaccine against this virus. The factors of influence remain valid for all the cases, but the number of susceptible individuals to the disease decreases drastically as an enormous part of the population of Romania has taken this vaccine. Consequently, in the final months of the epidemic, the spreading has reached the lowest level of activity and finally became inactive.

B. Spreading in the outer network

The purpose of the simulation is to estimate the number of newly infected individuals in a month and in a county, therefore, one part of the outer simulation will succeed to do this for each internal node in the network. A probability of infection is computed and transmitted to the successor inner network in order to perform the right simulation.

The computations for approximating the number of cases of infection are done only on the basis of the factors that encourage the spreading through the movement of masses. In other words, from the outer network there are taken into consideration the weights attached to the types of edges. Each type of edge has a fixed weight that can express either the percentage of population involved in the movement action, or a probability of movement of the infected individuals towards another node. The one exception from the rule, the PL edge type, carries two types of weights, one variable representing the percentage of poverty in the county, and the other one fixed representing the probability of movement of the infected people.

The number of newly infected individuals in one internal node "inherited" from one of the connected (by any type of edge) nodes is computed as a simple product between the number of infected individuals in the neighbor node and the weight attached to the corresponding weight:

$$product_i = Infected_i \cdot w(t)$$
$$t \ni T = \{IN, EN, CN, AN, TN, RN\}$$
$$w : T \to \{0,1\}$$

For the PL edge the product will have the variable weight as an extra factor:

$$product'_i = Infected_i \cdot w(t) \cdot W(i)$$
$$t = PL$$
$$W : \{counties\} \to \{0,1\}.$$

This product is computed for a number of times equal to the number of existing type of edges between any two nodes. Consequently, the entire number of infected individuals in a node is computed as a sum of all products for each edge connected to the node:

$$new\_Infected = \left[\sum_{i=0}^{i<\|no\_edges\|} product_i \right]$$

The final result is sometimes affected, positively or negatively, by the external factors.
When the index for the infection likelihood increases, so should the general number of infected individuals:

$$new\_Infected = new\_Infected \cdot Infection\_likelihood\_index$$

When a cure (or vaccine) has been found, the number of infected individuals should decrease drastically:

$$new\_Infected = \frac{new\_Infected}{cure\_found\_index}$$

Also, a probability of infection must be delivered for the inner network as standard for the infection spreading. The product described above has as factors an integer number – the number of already infected individuals –, and a sub-unitary factor – the probability of movement of the infected individuals. For the number of newly infected people we take only the integer part (whole individuals) and the factionary part will be used in the computation of the infection probability.

The probability to be set for the inner network as standard infection probability is used as a bound: if the infection probability for one node is equal to or greater than the bound, then the node is considered infected. Consequently, for having infected individuals, the probability must be as small as possible. The translated connection with the computations at the level of the outer network is that if a node has many connections with other infected nodes, then the probability of infection decreases, although logically it should increase. A fixed upper bound is set for the probability. For each edge connected to the internal node the probability will decrease, as explained above:

$$Infection\_probability = bound - \sum_{i=0}^{i<\|no\_edges\|} product_i$$

The infection probability is affected in an indirect proportion by external factors, infection likelihood and the existence of a cure. When the infection likelihood increases, the probability decreases in a reverse proportion and when a cure is found the probability increases considerably.

These two values are transmitted as parameters for the spreading simulation in the corresponding inner network.

C. *Spreading in the inner network*

The spreading simulation inside the inner network has a structure quite similar to the one of the epidemic spreading over cellular automata [11]: one node becomes infected when a defined number of neighbors are known to being infected. However, the simulation is considered to be over social networks as the number of neighbors is distinct among the components of the network and the number of edges from a node has a greater importance than the node itself. The relative similarity is underlined at the moment of infection when the probability of local infection is compared with the general probability of the network.

For each node in the network, the number of neighbors already infected is computed and is considered a local infection probability, determined as the ratio between the number of infected neighbors and the total number of neighbors. This local infection likelihood is compared with the one received from the upper level. If the local one is equal to or greater than the general one, then the current node becomes infected, otherwise, the node remains healthy:

$$pl_t = \frac{Infected\_neighbours\_no_t}{neighbours\_no_t}$$

$$pl_t \geq pg \Rightarrow i - Infected$$

The inner network size is considerable and testing whether each node becomes infected or not involves an exponential parsing of the network. Thus, the checking will start with a set of already infected nodes by parsing their non-infected neighbors for computing the local probability. Whenever a node gets infected, it is added to the set for checking its neighbors further.

D. *Algorithm outline*

The main core of the simulation theory can be described using the algorithm below:

```
 SimulationAlgorithm(pastMonth)
beginAlgorithm
       M(V, E) ← getMonthSituation(pastMonth)
       currentMonth ← next(pastMonth)
       for v ∈ V do
              noInfected ← getNoInfected(M,v)
              generalProbability ← computeProbability(M,v,currentMonth)
              newInfected ← computeNewInfected(M,v,currentMonth)
              noInfected ← noInfected+newInfected
              noClusters ←( noInfected*100/ getPopulation(v)) * getDensity(v)
              clusteringIndex ← scaleDensityToClusteringIndex(M,v)
```

```
            simulatedRes ← 0
            for i ← 1 , noClusters do
                  if  noClusters=i then
                        network ← createInnerNetwork(MAX_SIZE,
                        generateAround(clusteringIndex),
                         (noInfected mod noClusters), generalProbability )
                  else
                        network ← createInnerNetwork(MAX_SIZE,
                        generateAround(clusteringIndex),
                         (noInfected div noClusters), generalProbability )
                  endif
                  simulatedRes ← getSimulated(network) + simulatedRes
            endfor
            simulatedRes ← simualtedRes – noInfected + newInfected
            storeResult (v, simulatedRes)
      endfor
endAlgorithm
```

where:
`getMonthSituation(pastMonth)` returns the multigraph having in its nodes the required information for the *pastMonth*
`getNoInfected(M,v)` returns the number of infected individuals for node *v* in the structure *M*
`computeProbability(M,v,currentMonth)` computes the probability for the node *v*
`computeNewInfected(M,v,currentMonth)` computes the new infected individuals for node *v*
`createInnerNetwork(MAX_SIZE, generateAround(clusteringIndex), (noInfected mod noClusters), generalProbability)` builds the inner network
`scaleDensityToClusteringIndex(M,v)` scales the density

and `generalProbability` refers to the probability received from the outer network and is set as standard for the inner networks.

IV. Numerical Experiments

Numerical experiments are performed in Romania, over a hierarchical network with two layers corresponding to the country level (this network has 41 nodes corresponding to the 41 counties) and the county level (for each of the 41 counties, the network has a variable number of nodes according to the population size of each of them.
The results are simulated over 9 months, between May 2009 (the starting months which is not taken into consideration for simulations) and February 2010.

A. The Data

The data collected includes the monthly situation of newly infected individuals for each county from Romania and of the countries included in the network with which our country has stronger connections.

Table 1 contains the number of newly infected people in each county and for each month between May 2009 and February 2010. The data in this table has been collected from the Romanian Health Ministry official website; the section of press communicates [14].

Table 1. The number of newly infected individuals in each county and for each month between May 2009 and February 2010.

County Country	**Number of newly infected individuals**									
	May 2009	June 2009	July 2009	August 2009	September 2009	October 2009	November 2009	December 2009	January 2009	February 2010
Alba	0	0	0	0	0	0	22	44	13	0
Arad	0	0	0	0	0	0	80	50	10	0
Arges	0	0	0	1	1	0	29	54	23	1
Bacău	0	0	0	2	0	0	130	111	29	1

Bihor	0	0	0	0	0	1	10	4	2	0
Bistrita-Nasaud	0	0	0	0	0	0	0	1	3	3
Botoşani	0	0	0	0	0	0	269	133	111	0
Braşov	0	0	27	9	3	1	29	43	14	4
Brăila	0	0	0	6	0	0	2	11	14	1
Buzău	0	0	0	0	0	0	25	121	5	0
Caras-Severin	0	0	0	0	0	0	19	63	14	1
Calarasi	0	0	1	2	0	0	3	0	5	0
Cluj	0	0	0	6	1	3	58	84	45	1
Constanţa	0	0	5	2	1	0	21	55	48	4
Covasna	0	0	0	0	0	0	5	22	11	4
Dâmboviţa	0	0	0	2	0	0	94	107	52	1
Dolj	0	0	8	0	0	1	148	98	8	0
Galaţi	0	0	1	4	0	0	85	33	22	1
Giurgiu	0	0	0	0	0	0	20	16	8	0
Gorj	0	0	0	0	0	0	0	0	0	0
Harghita	0	0	0	0	0	0	52	39	23	0
Hunedoara	0	0	0	8	0	0	94	40	27	0
Ialomiţa	0	0	0	0	4	0	16	22	5	0
Iaşi	0	7	8	6	8	40	171	58	39	3
Ilfov	5	15	62	63	9	12	725	612	260	15
Maramureş	0	0	0	0	0	0	18	19	3	0
Mehedinţi	0	0	2	1	1	0	11	12	1	0
Mures	0	0	7	7	0	1	27	66	56	3
Neamţ	0	0	0	0	0	0	56	33	24	1
Olt	0	0	0	0	0	0	10	24	40	0
Prahova	0	0	1	4	0	41	51	83	72	0
Satu Mare	0	0	0	0	0	0	0	0	3	1
Sălaj	0	0	0	0	0	0	3	3	1	0
Sibiu	0	0	1	3	0	1	24	79	16	0
Suceava	0	0	0	2	0	0	10	70	30	1
Teleorman	0	0	0	3	1	0	10	10	3	0
Timiş	0	5	2	8	0	0	23	62	38	3
Tulcea	0	0	0	0	0	0	6	23	5	1
Vaslui	0	0	0	0	0	0	45	27	4	0
Vâlcea	0	0	0	3	3	0	22	11	6	0
Vrancea	0	0	0	0	0	0	62	43	4	0
Ukraine	0	1	1	0	6250	850000	11005	1230	301	22
Hungary	0	7	11	138	1250	1877	1107	203	70	3
Bulgaria	0	5	10	47	470	100000	2307	967	111	25
SUA	2254	20000	33902	6700	3200	1050	320	115	67	5
Canada	280	5438	7983	2060	986	320	98	20	0	0
UK	40	1540	7447	5957	8960	17325	6015	2000	200	0
Spain	93	430	760	838	2600	17303	1230	700	183	7
Mexico	1626	4957	10262	2350	1739	600	121	67	21	0
France	12	171	300	825	3024	7017	659	226	105	0
Turkey	0	27	40	50	180	625	303	29	15	0
Greece	0	58	109	1340	2506	2030	270	37	25	0
Germany	11	291	470	12320	1445	4445	750	217	32	12
Italy	9	86	130	1138	7213	21207	3070	375	93	31

Portugal	1	6	27	1960	1530	1248	625	123	75	1
Netherlands	3	100	134	1368	1020	2364	950	99	65	5

Another set of data that remains constant during the simulation contains the characteristics of each internal node in the outer network. These data consist of the population size and population density, as seen in Table 2.

Table 2. Population size and density of each county.

County Name	Population Size	Population Density	County Name	Population Size	Population Density
Alba	382.747	61	Hunedoara	485.712	69
Arad	461.744	60	Ialomiţa	296.572	67
Argeş	652.625	95	Iaşi	826.552	150
Bacău	706.623	113	Ilfov	2.221.860	389
Bihor	600.223	84	Maramureş	510.110	81
Bistriţa-Năsăud	317.254	58	Mehedinţi	306.732	62
Botoşani	452.834	91	Mureş	580.851	86
Brăila	373.199	78	Neamţ	557.000	99
Braşov	596.140	110	Olt	489.274	89
Buzău	495.325	81	Prahova	829.945	183
Călăraşi	324.617	64	Sălaj	248.015	64
Caraş-Severin	333.219	39	Satu Mare	367.281	83
Cluj	692.316	105	Sibiu	421.724	78
Constanţa	715.151	101	Suceava	688.435	80
Covasna	222.449	60	Teleorman	436.025	75
Dâmboviţa	541.763	134	Timiş	659.512	76
Dolj	734.231	99	Tulcea	265.349	31
Galaţi	619.556	139	Vaslui	455.049	72
Giurgiu	297.859	84	Vâlcea	413.247	86
Gorj	387.308	69	Vrancea	391.833	80
Harghita	326.222	52			

The simulation does not take into account the population information (size and density) for the countries with which Romania has connections. Among the data sets that remain unchanged and are collected from different sources, we also have the connections between the nodes of the outer network. Each node in the outer network – the multigraph – has a set of connections with the other nodes. These connections represent the edges for all multigraphs created for each month, although not always all of them are taken into consideration (an example is given in Figure 5). Table 3 contains the degree distribution for the internal nodes of the outer network, including both the indegree and the outdegree.

Table 3. The degree distribution of the internal nodes.

Internal Node	Outdegree	Indegree	Internal Node	Outdegree	Indegree
Alba	10	3	Hunedoara	4	5
Arad	4	3	Ialomita	3	4
Arges	6	2	Iasi	11	17
Bacau	7	4	Ilfov	13	27
Bihor	4	6	Maramures	4	2
Bistrita-Nasaud	5	0	Mehedinti	3	5
Botosani	4	2	Mures	4	6
Brasov	9	9	Neamt	3	5
Braila	9	1	Olt	2	2
Buzau	5	2	Prahova	1	7

Caras-Severin	5	0	Satu Mare	3	2
Calarasi	5	1	Salaj	1	4
Cluj	10	23	Sibiu	5	5
Constanta	4	15	Suceava	9	11
Covasna	3	3	Teleorman	1	4
Dambovita	3	3	Timis	1	18
Dolj	9	5	Tulcea	1	3
Galati	8	4	Vaslui	3	4
Giurgiu	2	3	Valcea	2	7
Gorj	4	2	Vrancea	3	6
Harghita	5	3			

B. Parameter setting for the algorithm

The simulation involves a number of parameters. The numerical ones are described in Table 4 together with their values.

Table 4. The numerical parameters involved in the simulation.

Name	Value	Description
MAX_CLUSTERING_ INDEX	0.6	The maximum clustering index over all counties in the network, as any higher value will result in an complete graph
P_INTERNAL_ NEIGHBORHOOD	0.0001	The percentage/probability attached to the IN edge type
P_EXTERNAL_ NEIGHBORHOOD	0.0003	The percentage/probability attached to the EN edge type
P_COLEGIAL_ NEIGHBORHOOD	0.002	The percentage/probability attached to the CN edge type
P_RAILROAD_NODE	0.0005	The percentage/probability attached to the RN edge type
P_POVERTY_LEVEL	0.007	The percentage/probability attached to the PL edge type
P_AIR_NODE	0.0003	The percentage/probability attached to AN edge type
P_TURISM_NODE	0.0001	The percentage/probability attached to the TN edge type
MAX_PROBABILITY	0.9	The maximum probability allowed for the spreading probability transmitted to a corresponding network.

Apart from the numerical parameters, temporal parameters, corresponding to the month for which the simulation is performed, are considered (a description of them is provided in Table 5).

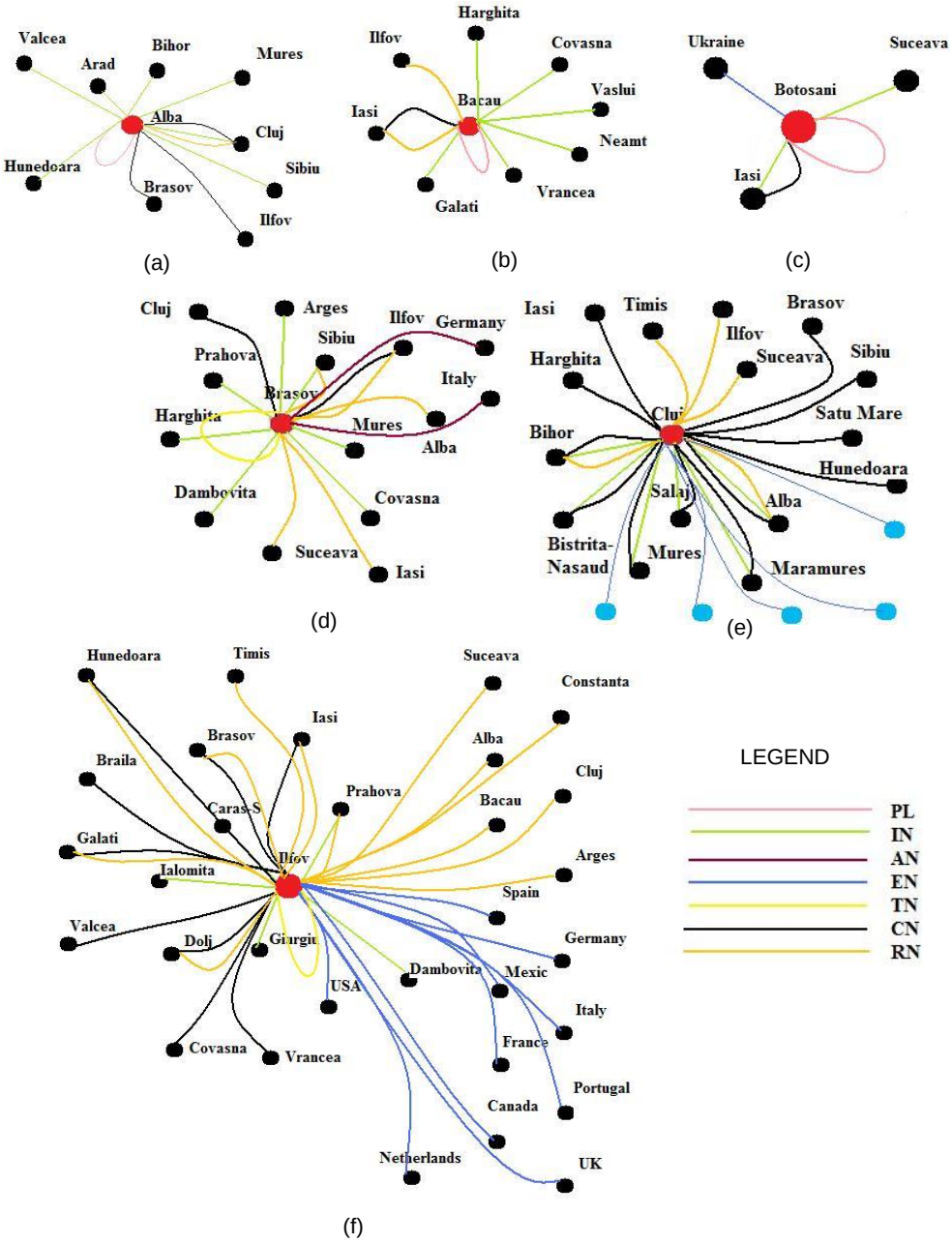

Figure 5. The connection set for some selected counties: Alba (a), Bacau (b), Botosani (c), Brasov (d), Cluj (e), Ilfov (f) (the capital Bucharest is included in Ilfov county). The blue nodes represent the external node to which a node is connected. For Cluj for instance, they are Spain, France, Germany, Italy, and Portugal.

Table 5. The temporal parameters.

Month	Attached Properties	Month	Attached Properties
May	- school month	*October*	- school month
June	- holiday month	*November*	- school month - infection likelihood month
July	- holiday month	*December*	- holiday month - school month - infection likelihood month
August	- holiday month	*January*	- school month - infection likelihood month
September	- holiday month	*February*	- school month - the month were the cure was distributed

The values of the numerical parameters have been set based on preliminary experiments. The temporal parameter values have been chosen from the real-world context.

C. *Errors Measures*

The simulation aims at approximating as well as possible the results of the A/H1N1 epidemic spread over the duration of a month for each county. The performance of our method is measured by the speed and convergence of the errors towards their lower limit (0 or 1, according to the definition interval).

We define three error measures, taking into account the following:

- the distribution between the counties of the extra or the missing infection cases simulated
- the ration between the real number of infections and the number of infections generated by our simulation
- the distribution of the absolute error value.

1) *Missing/extra cases*

This error represents the proportion of the missing/extra cases in comparison with the real number of cases of the current month and is given by:

$$err_1(M,M') = \frac{\sum_{i=1}^{i\le 41} |real(i) - simulated(i)|}{\sum_{i=1}^{i\le 41} real(i)}$$

$$i \in M, i \in M'$$

where M denotes the real set of nodes and M' denotes the simulated set of nodes. Values closer to 0 are proffered.

2) *Ratio between the real and the simulated number of infections*

This error is given by the ratio between the real number of infection per country and the simulated number of infection cases per country. In the ideal case, the two numbers would be equal; consequently, the lower bound is 1:

$$err_2(M,M') = \frac{\sum_{i=1}^{i\le 41} |real(i)|}{\sum_{i=1}^{i\le 41} simulated(i)}$$

$$i \in M, i \in M'$$

For even more accurate error computation, this error can be extended by computing the ration between the real and the simulated number of cases for each county in Romania and the final error will be computed as the sum of all ratios divided to 41, the number of counties. This error is desired to converge to 1 during the execution as if the sum of all ratios, in the best case, is 41 and after the final division will be 1:

$$err_3(M,M') = \frac{\sum_{i=1}^{i\le 41} \frac{real(i)}{simulated(i)}}{41}$$

$$i \in M, i \in M'$$

D. *Results of the simulations*

The results of our simulation report the number of infections for each internal node of the outer network from June 2009 until February 2010. May 2009 is given as input data, being the month in which the virus has first been reported in Romania.

1) June 2009

We construct the network whose nodes are the counties of Romania and some countries. The nodes contain as information name, population size, and population density (as given in Table 2). The situation of the month given as input (in this case May 2009) is also incorporated and consists in the number of already infected individuals.
Once the nodes are defined within the network, the connections are loaded, fulfilling the degree distributions listed in Table 3.

The errors obtained for this simulation are given in Table 6.

Table 6. Errors for month June 2009.

Error type	Value
err_1	1.11
err_2	1.58
err_3	0.36

It can be observed that the distribution of the extra/missing cases error is very small, indicating the fact that simulation is getting closer to the reality. err_3 indicates that the size of the simulated infected population is 1.58 times smaller than the real one.

2) July 2009

For the month *July* we have as input the situation June 2009. The CN edge type is not taken into consideration as it is a holiday month and for this reason the edge type TN is considered. Results of the simulation are given in Table 7.

Table 7. Errors for month July 2009.

Error type	Value
err_1	0.76
err_2	2.97
err_3	1.26

err_2 indicates that the simulation returned a smaller number of cases than the real situation: 2.97 times smaller.

There are 9 internal nodes which were reported to have been newly infected and the algorithm returned the following ones as newly infected: Brasov, Constanta, Dolj, Galati, Mehedinti, Mures, Prahova, Sibiu – which means that 8 out of 9 were identified by our simulation.

3) August 2009

August is the month when most of the people go on holiday, consequently there will be slight modifications on the computations: the edges having the type CN (Collegial Neighborhood) will not be taken into consideration, while the edge with the type TN (Tourism Node) will be used for the simulation, with the results presented in Table 8.

Table 8. Errors for month August 2009.

Error type	Value
err_1	0.72
err_2	0.75
err_3	0.43

The fact that the values for the last two errors are smaller than 1 indicates that the algorithm simulated a higher number of infections as they are in reality.

One argument for this could be that during the month of August it is possible that the theory of moving masses on the TN edges is not sustained as people can make their holidays abroad, not only within the country.

4) September 2009

For the month *September* we have as input the situation of August 2009; in this month school starts for undergraduates, for the college students the courses have not begun yet, so the CN edge type will not be considered and the TN one will.

Table 9. Errors for month September 2009.

Error type	Value
err_1	4.59
err_2	0.191
err_3	0.097

The errors of the simulation (as given in Table 9) are slightly worse for this case due to the fact that the number of infected counties has increased considerably and the simulation has been expanded throughout the county. The total size of the infected population (as returned by the simulation) is 0.191 times higher than the real one.

5) October 2009

The month given as input is *September*. The month October imposes some other limitations over the computations for the epidemic simulation. This month is known to be the month when the collegial year begins, so the CN edges are "means of transportation" of population from the nodes to the surroundings. The TN edges will not be considered for the computations. Table 10 contains the error values.

Table 10. Errors for month October 2009.

Error type	Value
err_1	0.89
err_2	1.75
err_3	0.21

For this experiment, a very good result has been obtained, with 1.21 cases/county being wrongly distributed, which, in the country population context, the difference is hardly noticeable. The second type of error is in the normal range, with no spectacular value. The ratio, per the entire country, indicates that the simulation returned a smaller number of infection cases than the real situation.

One of the most relevant reasons for having these errors is the fact that October is an autumn month when the temperatures decrease and the infection likelihood increases, but not that much as in the months to follow.

6) November 2009

November is a school month; consequently the CN edge type is included in the computations. The temperatures in this month decrease drastically and the crowding coefficient (in terms of transportation) increases, therefore the likelihood of contacting the virus is greater, so the probability of infection being transmitted within the inner networks will be modified accordingly.

Table 11. Errors for month November 2009.

Error type	Value
err_1	0.86
err_2	3.86
err_3	7.87

The errors (see Table 11) may not look that promising (especially the last two ones), but they are not, the specifics of the month encourage the real epidemic spreading more than in the months before, although the initial values were in a normal range. The ratio between the real value and the simulated one indicate the fact that the evolution of the virus was unexpected, thus the size of the simulated infected population is 3.86 times smaller than the real value. The value of error1 is in the normal range, indicating a success of the simulation over this month.

One of the greatest achievements in this simulation was the infection of the county Botosani, which until this month was not infected and now, according to the simulation, was reported to have 267 cases, when the real value was 269.

7) December 2009

This month is one of the most complex ones. For the first half of this month the courses are held, so it can be considered a school month – the CN edge type is considered for computations – but the other half it is a general holiday (winter holiday), so the TN edge type is considered as well. Besides these two infection causes, the infection likelihood is considerable and it is added to the entire equation.

Table 12. Errors for month December 2009.

Error type	Value
err_1	10.91
err_2	0.08
err_3	0.65

The results reported in Table 12 show that the number of extra cases is higher than the expected values: the simulated infected population size is 1/0.08 times higher than the real size. There are many factors that influence positively the probability of infection – it increases the chance for any individual in any inner network to become infected.

8) January 2010

For January, the infection likelihood remains a preset factor (the temperatures are still low) and this month is a school month. The inherited set of infections is significant and encourages the spreading.

Table 13. Errors for month January 2010.

Error type	Value
err_1	8.83
err_2	0.101
err_3	0.097

The errors – presented in Table 13 – have an important divergence from the desired value, but smaller than the one from the previous month. The explication for these errors has its basis in the algorithm specific characteristics of the month. Another reason for having these errors when comparing with the real situation is that, towards the end of the month, the vaccine against the virus has already appeared and a percentage of the population has been vaccinated – this characteristic being ignored in the simulation.

9) February 2010

The last month considered in our experiments is *February 2010*. This month is mostly characterized by the fact that during this time the vaccine against the virus A/H1N1 was world-wide distributed, including Romania, therefore the simulation of the spreading did not function at the same parameters.

Table 14. Errors for month February 2010.

Error type	Value
Error1	1.14
Error2	1.42
Error2’	0.58

The errors for February – as in Table 14 – seem acceptable. The importance of the proportion of the missing cases against the total number of cases is 1.14, taking into account the fact that the number estimated by the algorithm is 1.42 times smaller than the real value of the epidemic result during this month.

V. CONCLUSIONS

The paper proposes a new approach for analyzing epidemic spreading over social networks by introducing a new model tested against real-world results. The model is based on intensive research in social networks and epidemic spreading, viewed from different aspects: the mathematical aspect and the sociologic-statistical aspect. The data selected for the model has been restricted to a number of characteristics and supports further extensions.

The model developed is general and can be applied to any hierarchical organizational structure similar to the one of Romania and it is valid for the simulation of the spreading of any other virus. The simulation algorithm can

support modifications to fit any other epidemiological model. The application, although it is presented as a case study, can be modified to have a general character: it can suit any country of the world, only with the change of data from the database and of the characteristics deduced from the time of year, which rather seem to be specific to Romania.

The results of our simulations have been compared to the real data and the real situation in Romania and shown to being very promising.

Acknowledgement: This work is partially supported by the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project number PN-II-PT-PCCA-2011-3.2-0917.

REFERENCES

[1] Badham J., Stocker R., *The impact of network clustering and assortativity on epidemic behaviour,* Theoretical Population Biology 77, pp. 71-75, 2010

[2] Barabasi, A.L., Scale-free networks: A decade and beyond, Science 325, pp. 412-413, 2009

[3] Boni M.F., Manh B.U., Thai P.Q., Farrar J., Hien T.T.**,** Hien N.T.**,** Van Kinh N., Horby P., *Modelling the progression of pandemic influenza A (H1N1) in Vietnam and the opportunities for reassortment with other influenza viruses,* BMC Medicine 7, pp. 43-47, 2009

[4] Caldarelli G., Vespignani A.: *Large Scale Structure and Dynamics of Complex Networks from Information Technology and Natural Science,* World Scientific, London, 2007

[5] Caldarelli G, Scale-Free Networks, Oxford Univ. Press, Oxford, 2007

[6] Daniel T., Bleckmann P.*: Epidemic algorithms,* Universitat PaderBorn Report, 2004.

[7] Degenne A., Forsé M., *Introducing social networks,* Sage, London, 1999

[8] Dehmer M., Emmert-Streib F.*: Structural similarity of directed universal hierarchical graphs: A low computational complexity approach,* Applied Mathematics and Computation 194, pp. 7-20, 2007.

[9] Keeling M.J., Eames K.T.D.: *Networks and epidemic models,* J R Soc Interface 2, pp. 295–307, 2005.

[10] Krebs V., *Social Network Analysis, A Brief Introduction,* http://www.orgnet.com/ # Milgram, S., The small world problem, *Psychology Today*, 1:1, pp. 60-67, 1967

[11] Mikler A.R., Venkatachalam S., Abbas K., Modeling infectious diseases using global stochastic cellular automata, Journal of Biological Systems, 13:04, pp. 421-439, 2005

[12] Milgram, S., The small world problem, *Psychology Today*, 1:1, pp. 60-67, 1967

[13] Newman M. E. J., Watts D. J.: *Random graph models of social networks,* PNAS 99:1, pp. 2566-2572, 2002

[14] Romanian Health Ministry, Press release, http://www.ms.ro/?pag=62

[15] Watts D. J., Strogtz S. H., *Collective dynamics of 'small-world' networks,* Nature 393, pp. 440-442, 1998

[16] Wasserman S., Faust K.: *Social network analysis: methods and applications,* Cambridge University Press, New York, 1994

[17] Wellman B., *Structural Analysis: From Method and Metaphor to Theory and Substance* Cambridge: Cambridge University Press, London, 1988