Molecular Dynamics Modelling of the Temporal Changes in Complex Networks

Krzysztof Juszczyszyn, Anna Musia, Katarzyna Musial, Piotr Bródka

Abstract—The dynamic of complex social networks is nowadays one of the research areas of growing importance. The knowledge about the temporal changes of the network topology and characteristics is crucial in networked communication systems in which accurate predictions are important. In this paper a physics-inspired method to track the changes within complex social network is proposed. This method is based on the dynamic molecular modelling technique used in physics for simulation of large sets of interacting particles. The data for the conducted research was derived from e-mail communication within big company (Wroclaw University of Technology). From this information the social network of employees was extracted. The created social network was utilized to evaluate the methodology of social network dynamics modelling proposed by authors.

I. INTRODUCTION

The emergence of complex behaviour in a system composed of many interacting elements (frequently referred to as agents) is one of the most fascinating phenomena of the world and also a prominent area of research. There are many types of complex networked systems, for which we have no commonly accepted definition but there is an agreement that such structure consists of multiple interacting components whose global behaviour cannot be simply inferred from the behaviour of the components [1]. The elements of the network are not independent beings but they are connected via relationships and in consequence they influence one another. The great challenge is to identify which component has caused that another component, which is connected with many elements, has changed its behaviour. The social networks are the example of such complex systems in which tracking changes is very resource consuming task.

Due to the scale and complexity of such systems, computer simulations became increasingly popular (in addition to traditional approaches – formal theories and empirical studies) in this area. Simulations serve as analytical models which make certain predictions about the future system’s behaviour. We also face a typical trade-off between simulations that take into account the detailed, microscopic description of the system (which approach, in theory assures the most accurate predictions, often with an unacceptable computational overhead) and the minimal set of rules that allows the modelling of system’s evolution [2].

It should be noted that many properties of multi-agent systems (including social systems in which agents stand for social entities) are hardly definable in terms of any analytical model. Therefore, computer simulations seem to be the only way to insight into global system dynamics [2][3]. So far, physics has provided several methodological approaches to tackle this issue. Spatial mobility and concentration of interacting (via interaction potential like Coulomb forces) particles may be modelled with molecular dynamics paradigm. It allows many interesting and applicable extensions based on the reinterpretation of potentials and distance in the given space [4]. The proposition of such modification is described in details in this article. Another approaches successfully applied to physics, biology, evolutionary biology and social sciences are cellular automata [5], starting from famous Game of Life artificial life model of Conway.

According to our best knowledge no computational models for assessing the evolutionary schemes of large real-world internet-based social structures were developed so far. In this paper we propose the application of molecular dynamics to modelling the evolution of large email-based social network. A methodology of constructing the evolutionary model from email server logs is proposed and the various problems and challenges met when dealing with this hardly addressed problem are discussed.

The rest of the paper is organized as follows: the creation of social network on the basis of e-mail data is described in section 2. Section 3 introduces the basic concepts of molecular modelling. The following section 4 presents the molecular simulation that was performed during the experiments. Finally section 5 defines the procedure of discovering the parameters of social network’s molecular model and section 6 includes the experiments conducted on the e-mail based social network derived from the e-mail logs.

II. CREATING AN EMAIL-BASED SOCIAL NETWORK

The network that has been chosen for experiments, is based on the email logs of Wroclaw University of Technology (WUT). The network was created in the course of the data cleansing process and removing fake and
external email addresses. The experimental data were collected during the period of 21 months (February 2006 – October 2007). Emails sent to each other by the employees of WUT were used to infer their relationships thus allowing the creation of the WUT social network.

Note that although every single email provides information about the sender activity, it can simultaneously be sent to many recipients. An email sent to only one person reflects strong attention of the sender directed to this recipient. The same email sent to 20 people does not respect each individual recipient so much. For that reason, the strength of email communication $S(x, y)$ from email user $x$ to $y$ has been defined in the following way:

$$ S(x, y) = \frac{\text{card}(EM(x,y))}{n_i(x,y)} $$

where:

$EM(x,y)$ – the set of all email messages sent by $x$ to $y$;

$n_i(x,y)$ – the number of all recipients of the $i$th email sent from $x$ to $y$.

In consequence, every email with more than one recipient is treated as $1/n$ of a regular one ($n$ is the number of its recipients).

The strength of the relationship $RS(x,y)$ between $x$ and $y$ is calculated as follows:

$$ RS(x, y) = \frac{S(x, y)}{n(x)} $$

where:

$n(x)$ – the total number of emails sent by user $x$, was introduced. The value of this function is from range $[0,1]$.

The resulting social network $SN = \langle N, RS \rangle$ is defined as a set of all network nodes $N$ and a set of weighted relationships $RS : N \times N \rightarrow [0,1]$ given by the equation 2. The social distance between two neighbouring nodes is defined as the value of $RS$ complemented to one. The above scheme will serve as input data to create 60 temporal images of social network addressed in section 6.

III. DYNAMIC MOLECULAR MODELLING

Dynamic molecular modelling is one of simulation methods widely used in physics. Simulations are very important because of the fact that many physical problems are too complex to obtain an analytical solution for them. One of them are as a class called many-body problems. Whenever we are interested in a dynamic of physical system that contains of many interacting objects we are dealing with a many-body problem. Molecular dynamic modelling is a powerful method to describe such systems. It is important that starting point for such simulations is a microscopic scale (a position and velocity of each particle), but it provides a macroscopic characteristics of the system (obtained based on rules and laws of statistical mechanics), e.g. temperature, energy, specific heat of the system etc.

As previously mentioned a system of many interacting objects, e.g. particles is discussed. To start simulations the initial position and velocity of each particle must be set. They can be chosen freely. Only conditions that need to be fulfilled is that two particles cannot have the same position and the overall velocity of the system in each direction should be equal zero. This is to provide that the system is not moving as a whole, but every part of it is moving independently. The physical parameters are: the density of the particles, defined simply as:

$$ \rho = \frac{N}{V} $$

where $N$ is a number of particles (in our case: the nodes of social network) and $V$ stands for the size of the system (volume for three dimensions and area for two). The density is important, because it gives information of the state of the system (whether it is a solid, liquid or gas). On the other hand there are some simulation parameters: the value and number of time steps. What is more the interaction between particles should be described. The interaction is usually given by a two-particle potential (meaning that each two particles are interacting, three-particle interactions are not very common in physics) which depends on the distance between the particles and some characteristics of objects in the system (e.g. their mass, charge etc.). Our approach assumes that the potential reflects the tendency of two particles (nodes) to change their distance in social space. The potential is commonly known as a analytical formula obtained from either theory or experiment (phenomenologically), e.g. Coulomb potential for charged particles, Lenard-Jones’s potential. During our experiments the character of the potential will be estimated on the basis of known history of social network (past evolution which is known). From the potential the force acting between particles can be calculated using well-known formula:

$$ \overrightarrow{F} = -\nabla U $$

where:

$U$ – the interaction potential.

Knowing the force the time evolution of the system can be obtained by solving the classical equation of motion (2nd Newton’s principle of dynamics) for each particle separately:

$$ \overrightarrow{dp} = m \overrightarrow{a} $$

where:

$p$ – the momentum of the particle and $F$ is a vector sum of forces from all other particles in the system. For objects with constant mass this formula takes simpler form:

$$ \overrightarrow{F} = ma = m \frac{d^2 \overrightarrow{r}}{dt^2} $$

by using the definition of the momentum and acceleration $a$ (which is defined as a second derivative of the position vector $r$). During the simulation above equation must be
solved for each particle in each simulation step. This set of equations are solved by numerical integration using well-known algorithms (e.g. Euler’s method, Verlet’s algorithm).

This method is fully deterministic and allows to predict changes of particles’ positions in time – the time evolution of considered system. The biggest advantage of molecular dynamic modelling is that the simulation step is directly connected with time, so we are really able to obtain the time evolution of the system in contrary to other simulation methods, e.g. Monte Carlo simulation.

IV. MOLECULAR SIMULATION

To verify the created molecular model (see Section 3) theoretically obtained data are needed. It is possible to calculate the positions of all particles after given time using the Verlet’s algorithm. Input data indispensable for starting the algorithm are: initial positions of all particles, number of particles, formula for interaction force between each two particles. The number of time steps and the time interval between those steps can also be set. Usage of an additional parameter is possible – it is called the cut-off radius and defines the nearest neighbourhood for each and every particle, meaning the maximum distance of influence. Two particles away from each other for a distance bigger than the cut-off radius do not interact. This parameter is extremely important for large simulation when the number of particles is too big for taking all interactions into consideration. This radius should be large enough to treat the interactions with more distant particles as unimportant. It is possible to introduce this parameter only if the interactions shows short-range character.

The essential from the conceptual point of view part of the procedure is formula used to describe interactions between the particles. This force is identical for all two interacting particles within the system. What is more this force depends only on the distance between the particles. To model described system the Lennard-Jones potential is used:

\[
V(r) = 4\varepsilon \left( \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right)
\]

where:
- \( r \) – the distance between particles. The character of the potential is shown below (Fig. 1).

\[
\begin{align*}
F_x(r_{ij}) &= 48\frac{\varepsilon}{\sigma^2} \left( x_i - x_j \right) \left( \frac{\sigma}{r_{ij}} \right)^{14} \left( \frac{\sigma}{r_{ij}} \right)^6 \\
F_y(r_{ij}) &= 48\frac{\varepsilon}{\sigma^2} \left( y_i - y_j \right) \left( \frac{\sigma}{r_{ij}} \right)^{14} \left( \frac{\sigma}{r_{ij}} \right)^6
\end{align*}
\]

where:
- \( x, y \) – mark the particular component of the force vector;
- \( r_{ij} \) – is the distance between two interacting particles.

The next step is to transform the forces into dimensionless form. This way the magnitude of used values, normally extremely small, are better represented in a computer and the operations on them can be more precise. To do this the distances are introduced in [\( \sigma \)] units and time in \( \sqrt{\frac{m\sigma^2}{48\varepsilon}} \).

Where \( m \) is the mass of interacting particles. As a first approximation the masses of all particles are taken equal 1.

Now as the interaction force between each two particles is known we can move on to the Verlet’s algorithm. This algorithm is used to integrate the equation of motion for each and every particle in the system.
\[ m \frac{d^2 r}{dt^2} = F \]  

(9)

The second derivative should be discretized using e.g. the three-point formula based on the Taylor’s expansion. The algorithm allows us to obtain the position and the velocity of each particle in following time steps:

\[
\begin{align*}
\vec{r}(t + h) &= \vec{r}(t) + v(t) \cdot h + \frac{\vec{F}(t)}{2m} \cdot h^2 \\
\vec{v}(t + h) &= \vec{v}(t) + \frac{\vec{F}(t) + \vec{F}(t + h)}{2m} \cdot h
\end{align*}
\]

(10)

where:

\( h \) – the magnitude of the time step.

The main part of the algorithm is the time step loop. First the program calculate the initial forces acting on each particle on the basis of their initial positions. The force acting on a single particle is the sum of forces originating from all the other particles in the system. When the forces are known, the positions of the particles can be calculated. The new positions are used to obtain the forces in an actual step which are required for velocity calculations, because positions depend only on variables’ values in the previous time step whereas the velocities depend also on the forces in the actual time step. This procedure is repeated in every time step. As an output this algorithm returns the positions and velocities (velocities are not crucial for the interpretation in our case, at least during the first experiments) of all particles in the system in following time steps.

To compare the experimental data with model discussed above the positions of the particles should be generated for the same time step as in the experiment. Then the standard deviation of the two data series for each time step need to be calculated and optimized. By optimization process the adjustment to the experimental data is understood. This can be proceed by optimizing the value of the parameters of the Lennard-Jones potential. Values of this parameters can be changed in wide range to minimize the standard deviation between the theoretical and experimental data.

V. MOLECULAR MODEL OF SOCIAL NETWORK

Application of the molecular dynamics to social network modelling requires completing several tasks which will be defined in the further subsections.

The overall methodology assumes the four main stages:

1. Creating temporal images of social network.
2. Embedding created social network graph in 2D 3D Euclidean space.
3. Estimating the character of social force potential between network nodes.
4. Using dynamic molecular modelling to predict the evolution of the network.

A. Creating temporal images of the social network

It should be emphasized that the social network derived from the e-mail logs is not a static structure. The existence of any link in such a graph (i.e. relationship) is a result of a series of discrete events (sending e-mail messages) which occur in certain moments and (usually) with changing frequency. We may also think of the computed relationships’ strengths as of the social distance between network members (nodes). Greater RS reflects smaller distance in social space. In order to track changes in relationship strength we used the sliding time frame approach. Mail server log data from the period of 21 months (February 2006 - October 2007) were divided into frames covering one month each, and starting in consecutive moments which differ by ten days.

This allowed the creation of 60 social network graphs \( SN(t_0), SN(t_1), \ldots SN(t_n) \) where \( t_0, t_1, \ldots, t_n \) are discrete moments of time and each network is created (according to the procedure defined in section 3 on the basis of one-month period starting in \( t_0 \)). In our case \( t_0 = \text{“the 1st of February, 2006”} \).

The networks \( SN(t_0), SN(t_1), \ldots SN(t_n) \) are temporal images of evolving social structure which was built on the basis of the entire 21 months-long email communication. The relationship strengths are computed for each network separately.

The similar approach was utilized in [14] where the temporal changes of betweenness centrality were tracked for organizational email network of 200 users. The time frame of 30 days moving by one day was used. In our case, due to the size of the system and observed behaviour of several users (who happened to be inactive for weeks, for example) we decided to increase its speed to ten days per step.

B. Embedding networks in Euclidean space

Following the in-depth discussion presented in [6] we cannot expect the social space to be metric i.e. the triangle inequality between any three nodes may not be fulfilled. On the other hand, molecular modelling assumes the interaction between the particles moving in the Euclidean space. For this reason, in order to apply the molecular modelling we must conduct embedding of the social network graph in metric, Euclidean 2D space.

Numerous methods for completing this task were developed so far; the best will be chosen in the course of experiments, good candidates were proposed in [8] (so-called Big-Bang embedding algorithm) and [9] (a fast method for creating 2D representations of large graphs, which consists of two steps: embedding the graph in a very high dimension – usually associated with the number of nodes- and then project it into the 2D plane using principal components analysis).

In this step we embed the graphs \( SN(t_0), SN(t_1), \ldots SN(t_n) \) in Euclidean 2D space, where each node is represented by a point with given coordinates. The resulting sets of points
SN0, SN1, ..., SNn represent the temporal network images. Embedding algorithm assures us that the Euclidean distances between points (nodes) fit in the best possible way the distances in a social space (relation strengths in original graphs). We obtain the representation of social system in which the network is seen as an assembly of N (the nodes of social network). The procedure we have chosen for our experiments was a classic multidimensional scaling algorithm (MDS [15], [16]). Multidimensional scaling defines a call of methods often used in information visualization and exploring similarities or dissimilarities in data. An MDS algorithm starts with a matrix of similarities between objects (similarity relation doesn’t have to be symmetrical), then assigns a location of each item in a low-dimensional space. In other words – it estimates the coordinates of a set of objects in a space of specified dimensionality on the basis of measuring the distances (which, however do not have to be metric) between pairs of objects. A variety of models can be used that include different ways of computing distances and various functions relating the distances to the actual data. We computed a square non-symmetrical (because our graph is directed) matrix of the distances between the nodes in social graph as an input for the MDS procedure.

C. Setting up the dynamic molecular model

Because the sets of network nodes in SN(t0), SN(t1), ..., SN(tn) are equal, each point (node) is represented in any of the sets SN0, SN1, ..., SNn. In fact, in email-based social network it doesn’t have to be so, but to check our model we excluded users which were not continually active (about 10% of all) from our analysis. It is planned to model dynamic sets of the users in the future model development. However, the position of each node, depending of its distances to other points, is subject to change. We may think of these points as of the particles moving in result of interactions (email communication) between them.

At this point we use the formalism of molecular dynamics (see section 3) to associate a force potential U with any particle (network node). The actual characteristic of this potential depends on the behaviour of the particles changing their positions in time instants t0, t1, ..., tn. First experiments were carried using standard Lennard-Jones potential function, which resembles some features of the dynamics of email communication – the the growing intensity of communication is always followed by the periods of less frequent email activity. This resembles the repelling force emerging between particles when their distance becomes less then some rmin.

The force potentials introduced in section IV allow us to simulate the changes between communication patterns in consecutive time instants. The force potentials associated with the nodes reflect their abilities and tendency to establish future connections with the neighbours – the nodes which are close in terms of social space (thus changing the distances in social space which is analogous to the behaviour of particles moving under influence of electrical/gravitational forces). Note, that this approach allows also general interpretations because “being close in 2D social space” does not imply direct connection between nodes (users). Belonging to the same clique of the social graph has similar effect.

VI. EXPERIMENTS

The experiments were conducted in order to track the temporal changes in the network structure. This was performed by utilizing the time frame of 30 days moving by 10 days. A set of 60 separate (partially overlapping, as in [14]) frames was created from the email logs gathered in the period of February 2006 – October 2007. As stated above, in order to enable modelling the dynamics via the molecular simulation the users who were active in all 60 periods were chosen and this resulted in the creation of the set of 262 users and the relationships existed between them. This subnetwork was used in the further stages of our experiments.

Fig. 2 presents the results of running non-metric MDS procedure on the matrix of distances between social graph nodes.

![Fig. 2 Initial positions of 262 particles](image)

Molecular simulation was run on the set of particles with initial coordinates as showed on the Fig.2 and assuming that their interactions were characterized by standard Lennard-Jones potential function. However, first results have shown, that the exponent 14 in one of the terms of Equation 8 has to be decreased. It was responsible for rapidly growing repelling force between particles close to each other and has to be decreased to 10. After this change in the definition of the potential the system became stable thus allowing further simulations.

The configuration of particles moving in social space after 100 time steps is shown on Fig.3.
As we may see from comparing Fig. 2 and Fig. 3, the configuration of particles is stable, however their relative positions are changing. The analysis of the results of the first 100 time steps of molecular simulations allowed drawing interesting conclusions about the behavior of the systems of social particles.

First, there is one important information which cannot be inferred from the above picture. Small number of particles (less than 6% of the total 262) is located far outside the area visible on Fig. 3, in few cases with their coordinates exceeding 1.0E+10. This observation led to analysis of their (and – afterwards, of all the others) properties as the elements of the social graph. An interesting conclusion is that the accuracy of the molecular model depends on the position of the nodes within the social graph. We observed that:

- Nodes with minimal (i.e. equaling 1) in- and out-degree centrality (connected only to one neighbor in social network) when moved to social 2D space tend to approach the other nodes (particles) to very short distances. Apparently, this reflects the fact that their relative distance is based only on single social relation. This in turn results frequently in the emergence of large repelling force. The particle gains high speed and is thrown out – after that it moves along linear path, in our case the length of simulations was not enough to cause such a particle return in result of attracting force coming from interaction of the others. Obviously, this is not a behavior observed in social space.
- On the other hand, nodes with high degree centrality were observed to be stable during the interactions with the others.

In order to check the ability of our model to predict the social distances between particles we checked if there are correlations between node-to-node distances stored in distance matrices generated for all 60 temporal images of considered social network and the distances in 2D social space in consecutive steps of simulation. This was done separately for the four classes of nodes:

  - **Class 1**: Nodes with degree centrality equaling one.
  - **Class 2**: Nodes with high in-degree centrality (many – at least two – relations pointing to them, zero or one outgoing relations).
  - **Class 3**: Nodes with high out-degree centrality (many – at least two – relations originating in given node, zero or one incoming relations).
  - **Class 4**: Nodes with balanced in- and out-degree centrality both equaling at least two.

In all cases distances between any given node and all its counterparts in the network were computed for all temporal images of social network and the equal number of time steps of the molecular simulation. These vectors of distances were compared using the Pearson correlation coefficient. The average results for the members of the all considered classes are as follows:

  - **Class 1**: -0.0794
  - **Class 2**: 0.3726
  - **Class 3**: 0.4311
  - **Class 4**: 0.6617

We may see that there is visible correlation between the predictions of the model and the behavior of the particles/nodes well-connected to the network. It maybe explained by their communication patterns being dependent from the other members of the community. For the nodes characterized by broadcast type of communication (out-degree hubs) or information sinks (in-degree hubs) this correlation is not so strong, however visible.

The communication patterns of the nodes loosely connected to the network are not predictable in terms of the proposed model, their behavior is independent from their network counterparts.

**VII. CONCLUSIONS AND FUTURE WORK**

The results presented above are preliminary and summarize the first experiments on creating the predictive evolutionary model for large internet-based social networks. The same formalism may be also applied to other social networks based on communication technologies, for example, network of mobile phone users etc. There are many details which have to be addressed during future works and experiments. Among them the most important are:

- **Force potential issues**: The actual shape and equation of the force potential characteristic for given network should be estimated by analysing the history of interactions between pairs of users (nodes, particles) – we used “classic”, modified version of the Lennard-Jones potential. The question if there is one definition of the potential which may be applied to all members of the network remains open.
- **Choosing the time step of the molecular simulation**: In the case of our first experiments we analysed only simple correlations between changes in the distances in social space.
space and communication graphs of time-changing social network.

- Dynamically changing sets of network nodes.
- 2D social graph embedding procedure. There are other options which may be applied instead of the MDS used in our simulations. They may also affect the results and the accuracy of the model.

Future experiments will be carried on larger social networks (data sets based on logs from a 5000-users corporate server are being prepared for the forthcoming research) and will be preceded by developing methods of tuning the potential equations and the length of time steps. The approach proposed in this article aims to predict the evolution of social network of arbitrary size on the basis of known records. The existence of email-based social networks gives us the opportunity to use precise data (mail server logs) which hold information about the activity of users. As a result it was possible to propose a computational approach with good simulation potential and compatible with known solutions from complex systems field. Additionally, our approach was general and may be applied to any internet-based social network (their size and the length of history records do not impact the procedure).

However, there are also several challenges and questions that must be answered in order to apply the method with required accuracy.

ACKNOWLEDGMENT

The work was supported by The Polish Ministry of Science and Higher Education, grant no. N N516 264935.

REFERENCES