Molecular Model of Dynamic Social Network Based on E-mail Communication

Marcin Budka · Krzysztof Juszczyszyn · Katarzyna Musial · Anna Musial

Abstract In this work we consider an application of physically–inspired sociodynamical model to the modelling of the evolution of email–based social network. In contrary to the standard approach of sociodynamics, which assumes expressing of system dynamics with heuristically-defined simple rules, we postulate the inference of these rules from the real data and their application within a dynamic molecular model. We present how to embed the n–dimensional social space in Euclidean one. Then, inspired by the Lennard–Jones potential, we define a data–driven social potential function and apply the resultant force to a real e–mail communication network in a course of a molecular simulation, with network nodes taking on the role of interacting particles. We discuss all steps of the modelling process, from data preparation, through embedding and the molecular simulation itself, to transformation from the embedding space back to a graph structure. The conclusions, drawn from examining the resultant networks in stable, minimum–energy states, emphasize the role of the embedding process projecting the non–metric social graph into the Euclidean space, the significance of the unavoidable loss of information connected with this procedure and the resultant preservation of global rather than local properties of the initial network. We also argue applicabil-

M. Budka
Bournemouth University, Poole House, Fern Barrow, BH12 5BB Poole, UK
E-mail: mbudka@bournemouth.ac.uk

K. Juszczyszyn
Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
E-mail: krzysztof.juszczyszyn@pwr.wroc.pl

K. Musial
Kings College London, Strand Campus, WC2R 2LS London, UK
E-mail: katarzyna.musial@kcl.ac.uk

A. Musial
Wroclaw University of Technology, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
E-mail: anna.musial@pwr.wroc.pl

ity of our method to some classes of problems, while also signalling the areas which require further research in order to expand this applicability domain.

Keywords Complex systems · Link prediction · Molecular modelling · Social networks

1 Introduction

The emergence of complex behaviour in a system composed of many interacting elements is one of the most fascinating phenomena and recently also a prominent area of research. There are many types of complex networked systems, which can be classified in many different ways. One of the approaches distinguishes infrastructural (Internet, WWW, energy and transportation networks) and natural complex systems (biological networks, social systems and ecosystems) [3]. Another classification divides complex networks into technological, social, biological, and information networks [27]. There is no commonly accepted definition of a complex networked system 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 individual components [20,3]. The elements of the network are not independent but are rather connected via relationships and in consequence they influence each other. The number of nodes in these networks can differ from hundreds to millions [43]. One of the challenges is to identify, which component influences the behaviour of other components, which is directly connected with the dynamics of such structures.

Complex systems that are investigated in this paper are social networks where nodes represent people (but can also be other social entities such as departments or even whole organisations), connected by different types of social relationships (e.g. friendship, co–working, family) [15,41]. Although the general concept of social networks seems to be simple, the fact that the underlying structure is a network implies a set of characteristics, which are typical to all complex systems, i.e. the sum of the interactions between the users does not allow to draw conclusions about the behaviour of the social system as a whole. The consequence of this is that tracking changes in social networks is a very challenging and resource consuming task, especially that the number of edges of the graphs representing social networks that are nowadays at our disposal can be counted in millions.

Due to the scale and complexity of such systems, computer simulations became an increasingly popular tool for investigating the dynamics of complex systems including social networks. Simulations supplement traditional approaches — formal theories and empirical studies, and serve as analytical models enabling making certain predictions about the future behaviour of complex systems. In this research, we focus on the predicting the changes in the network structure. This is especially important as the network structure affects the functions of the network [39]. We also face a typical trade–off between simulations that take into account the detailed, microscopic description of the

system (an approach, which in theory assures the most accurate predictions, often with an unacceptable computational overhead) and the minimal set of rules that allows to model the evolution of the system [35].

It should be emphasized that many properties of complex systems are hardly definable in terms of any analytical model. Therefore, computer simulations seem to be the only way to gain insight into global system dynamics [35, 6]. So far, physics has provided several methodological approaches to tackle this issue. We hence argue that the spatial mobility and concentration of interacting particles can be modelled by employing the molecular dynamics paradigm, leading to many interesting extensions of standard approaches, based on the reinterpretation of potentials and distance in a given space [44]. One proposition of such modification is described in detail in this work. Another family of approaches successfully applied to physics, biology, evolutionary biology and social sciences are cellular automata [45], starting from the famous *Game of Life* artificial life model of Conway. One of the first researchers who applied the particle–based approach to social dynamics was Dirk Helbing, who in [17] proposed a fundamental dynamic model which includes many established models as special cases, (e.g. logistic equation, gravity model, some diffusion models, the evolutionary game theory and the social field theory), and also implies numerous new results.

However, in this work we argue that the rapid development of social portals and social media gives us an unique opportunity of the investigation of social systems on the basis of real data. When we consider inferring social relations from the records gathered from systems providing communication and recommendation services, the relations may be quantified and directly measured. On the other hand, a standard approach of sociodynamics assumes a global (and relatively simple) definition of social potential (which reflects the character of "social force" driving the changes in the relations between the system components) which is used to simulate and analyze the collective behavior of system components [13]. This approach has been proved useful for many classes of social systems and the modeling of opinion dynamics [32].

Taking the above into account, we propose to infer the character of social potential from the real–life social system data (using an email–based social network as an example) and to verify the possibility of using it to determine the evolution of the system. This requires embedding n–dimensional social space in Euclidean space in order to apply the physically inspired methods. According to the best of our knowledge no computational models for assessing the evolutionary schemes of real–world internet–based social structures, in which the edges can not only be formed but can also fade, were developed so far. Hence, in this paper, we propose application of molecular dynamics to modelling the evolution of email–based social network. We focus on the equilibrium state of a network, i.e. the state after the molecular simulation has converged, and discuss various issues and challenges encountered during this research. Moreover, we argue that, in the presence of the data coming from real system, the verification of such a model should be done by means of checking if it is possible to recreate the social network from simulation results

and comparing it with the real network structures which have evolved in the period of time covered by the simulation.

The rest of the paper is structured as follows: in Section 2 related work in the fields of social networks dynamics, graph embedding, and dynamic molecular modelling are presented. In Section 3 methodology followed in this paper is outlined and Section 4 explains the experimental set–up. Section 5 is devoted to the molecular simulation and its outcomes. Section 6 aims at presenting the concept of social network recreation from the simulation results and Section 7 includes the analysis of the retrieved social networks. Finally in Section 8 results arising from the conducted research are summed up and the future work is presented.

2 Related Work

2.1 Dynamics of Social Networks

In the last few years the problem of predicting the future interactions between users in social networks has become an important research challenge. Due to the availability of datasets of online activities and communication between people, scientists try to describe both structure and evolution of such networks. Most of the approaches addressing the complex networks growth take into consideration a limited set of global characteristics of the networks and develop models that reproduce only these characteristics, e.g. node degree distribution [2], clustering coefficient [42] or network diameter [7].

There are some approaches that aim at developing specific models for on-line social networks and take into consideration some information characteristic to such networks [29], [30], [9], [8], [31], [12], [25]. Different models propose different methods of network growth. In [29], on the basis of the analysis of real–world networks such as *Flickr* and *Yahoo 360!*, the users have been divided into three different types: passive, linkers and inviters. The members of the first group (passive users) join the network out of curiosity or because of being invited by a friend. These users, as their name suggests, never engage in any significant activities within the network and do not interact with other users. Inviters on the other hand, are interested in migrating the group that they have in the real world into a virtual world, thus their actions focus on inviting their friends to participate in an online social network. Linkers actively connect themselves to other members within the online social network. Based on the analysis of datasets the authors define a rule-based system that follows specific rules used for evolution of the social network. The method that describes the network growth can be defined as the set of steps: (i) at each time step, a node arrives, and one of the statuses: passive, linker, or inviter is randomly assigned to it; (ii) during the same time step, x edges arrive and the source of each of the edges is chosen at random from the existing inviters and linkers in the network using preferential attachment. Depending on the chosen type of the source node (inviter or linker) different actions are performed. If

the source is an inviter, then it invites a non-member to join the network, and so the destination is a new node. If the source is a linker, then the destination is chosen from among the existing linkers and inviters, again using preferential attachment [29]. This model represents the growth of a network, i.e. it takes into account adding new nodes and edges. However, the problem of link prediction covers not only the creation of new links but also fading of existing relations.

In [9] the authors have presented another approach that defines a set of rules regarding how the network evolves. They focus on discovering patterns of interactions between users and their evolution over time. The authors propose to create a single graph that represents social network, which is supplemented with additional information – a time-stamp, added to each relation when it appears in the network for the first time. The experiments were performed on the DBLP database [9]. Similarly to the previous presented study, also this one assumes that the users and the relations between them can only be added to the system and will never be deleted. Moreover, both of approaches presented so far allow to investigate the creation of new edges but do not allow to follow the dynamics of the relationships strengths between users, which is one of their limitations.

Yet another framework for the network growth was developed in [30] where the authors studied four online social networks: *Flickr*, *Delicious*, *Answers* and *LinkedIn*. They proposed to apply the maximum-likelihood estimation principle to compare a family of parameterised models in terms of their probability of generating the observed data, and as a result to select the model that reflects the data in the best possible way. The task in this framework was to predict which nodes will a new edge connect. For every edge arriving to the network the likelihood that it will connect two given nodes under some model is assessed. The product of these values over all edges gives the likelihood of the model and the model with the highest likelihood is chosen. Similarly to the previously presented methods this one also does not consider the strength of the relation as well as the fact that an edge can disappear from the network.

A survey of other link prediction methods can be found in [31], where the approaches like common neighborurs, Jaccard's coefficient and Adamic/Adar method, preferential attachment, Katz method, PageRank and its variants, low-rank approximation, unseen bigriams, and clustering, are discussed.

A set of approaches that take into consideration the fact that links can disappear from the network have been proposed in [19,8] where the authors have detected a dramatic time dependance in network centrality and the role of nodes, something not apparent from static analysis of node connectivity and network topology. Their experiments studied a large-scale email networks consisting of 57,000 users based on data gathered over a period of 113 days. They found that although the daily networks were scale-free, the well-connected nodes in these networks changed from day to day.

A recent method also accounting for the disappearing links has been proposed and investigated in [21,23,24], where based on the changes in the local structure, a 1st order probabilistic model of transitions between various triad

types has been derived. The model results from an observation that there exist distinctive patterns which drive the evolution of connections between nodes. Node disappearance has also been addressed in [34], but in a somewhat different the context of disruption of the information flow.

Our approach differs from these presented above as we do not propose a model for network growth per se but we investigate the limitations of sociodynamic model verified on data coming from real system. Our proposition takes into account both creation and vanishing of the relationships. Additionally, the network investigated in this work is a structure where strength of the relationships changes over time, which is an important factor in social networks due to the cognitive limitations of people [18]. In our approach, we do not assign roles to users as this may be misleading. We rather assess, based on the current interactions between users, how the relations strength and the structure of the network may look like in the future.

2.2 Distance Preserving Graph Embedding

Following the in-depth discussion presented in [42] we cannot expect the social space to be metric i.e. the triangle inequality between any three nodes does not hold. On the other hand, as it was mentioned above, molecular modelling assumes the interaction between the particles embedded in the Euclidean space. For this reason, in order to apply molecular modelling we must first perform embedding of the social network graph in a metric, Euclidean space. Numerous embedding methods exist, an overview of is presented below.

The **Big Bang embedding algorithm (BBE)** presented in [36] simulates an explosion of particles that represent network users under a force field that is derived from the embedding error. Each particle is the geometric image of a vertex. The force field reduces the potential energy of the particles which is related to the total embedding error of all particle pairs. In the Big Bang Simulation (BBS) all particles are initially placed in the same location in space. The whole process is performed in an iterative manner and each iteration moves the particles in discrete time intervals. Every iteration begins with calculation the field force on each particle at the current particles' positions (for the first iteration forces are chosen randomly). As it was mentioned, the forces are derived from the potential energy. In the next step, the positions and velocities at the next time step are calculated. The final step of each iteration is to evaluate the new potential energy. This method allows to embed the network into a freely selected number of dimensions.

Another method that can be used to embed a graph in Euclidean space is called the **Multidimensional Scaling (MDS)** ([40]). MDS defines a suite of methods often used in information visualization and exploration of similarities or dissimilarities in data. There are two variations of MDS, i.e. classical multidimensional scaling (CMDS) algorithm and standard MDS [10], [28]. Classical metric MDS develops the metric as a symmetric bilinear form and calculates the leading d eigenvalues of the corresponding matrix [40]. An MDS algorithm

starts with a matrix of similarities between objects (similarity relation does not have to be symmetrical), then assigns a location of each item in a low-dimensional space. It hence 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. Both methods allow to embed graph into different numbers of dimensions. However, the problem that we faced during our experiments with MDS was that the computational overhead was very high and we were not able to obtain results within reasonable time.

In **High–Dimensional Embedding** [16], which is a fast method for creating 2D representations of large graphs, the graph is first embedded into a very high dimensional space — usually associated with the number of nodes — and then projected into a 2D plane using Principal Components Analysis. This method is used for undirected graphs. It will not be useful from the perspective of our experiments as one of the goals of this study is to embed the graph into different dimensions and verify which number of dimensions helps to achieve best results from the link prediction perspective.

Minimum Volume Embedding (MVE) presented in [37,38] is an algorithm for non–linear dimensionality reduction that uses semi–definite programming (SDP) and matrix factorization to find a low–dimensional embedding that preserves local distances between points while representing the dataset in fewer dimensions. Authors of MVE emphasise that in all cases MVE in comparison with Semi-definite Embedding and Kernel Principal Component Analysis is able to capture more of the variance of the data in the first two eigenvectors, providing a more accurate 2-dimensional embedding [37,38]. The main features of the minimum volume embedding approach are: (i) MVE for a given dataset returns always the same set of coordinates, (ii) isolated nodes are neglected in the embedding process, and (iii) MVE is stable, i.e. adding one node with very weak connections does not influence significantly the positions of the remaining nodes. Enumerated characteristics of MVE means that the graph can be embedded only into 2D space which is not enough from the perspective of the proposed in this paper experiments as one of the goals is to find out what is the best number of dimensions to which the graph should be embedded. Moreover in the case of not connected graphs the algorithm does not work.

2.3 Dynamic Molecular Modelling and Simulation

Dynamic molecular modelling is one of the simulation methods applicable to large ensemble of interacting objects. It was historically used to model physical systems with huge number of particles. In its most classical version the particles are identical and indistinguishable, and interact with each other through two–particle mutual symmetrical potential, which is identical for every pair of

interacting particles and only distance–dependent. This model can be further extended and modified and it has already proven its applicability to more complex systems. The exact form of the interaction potential can differ depending on the details of the modelled system. In some cases it can be obtained experimentally if two–particle interaction can be separated, extracted and the dependence on their basic properties (e.g. mass, charge etc.) and inter–particle distance can be determined or is known from theoretical considerations. Unfortunately it is not always the case. Usually the microscopic details of interaction potential are not directly accessible experimentally and only the macroscopic characteristics (which can be described as the statistical mean values) of the whole particle ensemble are known (e.g. temperature, energy, entropy etc.). Although the behaviour of each particle on a microscopic scale is fully deterministic due to the inter–particle interaction being driven and governed by the second Newton's dynamics principle, it is only possible in very limited number of cases to deduce the form of interaction potential from the macroscopic behaviour of the particle ensemble, i.e. if sub–ensembles, characteristic clustering effects or short–range ordering can be observed. In most cases one has to assume a form of interacting potential (basing on the boundary behaviour of analysed system or some descriptive characteristics that can be deduced from macroscopic observations or general features of two–particle behaviour), perform the simulations of the system of interest and check if the behaviour of macroscopic observables can be reproduced. The problem of finding the interaction potential, in the case when the trajectory/time–dependence of particle position is known, is solvable by a number of differentiation and integration steps. Although, this procedure is well defined mathematically it cannot be conducted in the case of many interacting particles as the trajectory is not a simple analytical function but rather seems random (similar to Brownian motion) due to the complexity of analysed system, in which every particle responds to a force originating from all other particles. Based on the potential, the force acting between particles can be calculated using the following formula:

$$\mathbf{F} = -\nabla U \tag{1}$$

where U is the interaction potential.

If the force is known, the time evolution of the system can be obtained by solving for each particle separately the classical equation of motion (2nd Newton's principle of dynamics):

$$\mathbf{F} = \frac{\mathbf{dp}}{dt} \tag{2}$$

where p denotes 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 the following familiar form:

$$\mathbf{F} = m\mathbf{a} = m\frac{\mathbf{d^2s}}{dt^2} \tag{3}$$

by using the definition of the momentum and acceleration a, defined as a second derivative of the position vector s. The above equation must be solved for every particle in every simulation step. In order to start the algorithm, the initial positions of all particles, the formula for the force which is identical for all particle pairs and the interacting potential need to be specified. One of the standard potential functions used to describe the many-particle problems is Lennard–Jones potential which is given by:

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

where r denotes the distance between particles. The Lennard–Jones potential, which has been depicted in Figure 1, is fully defined by two parameters: the depth of the potential ε – responsible for the strength of interactions between particles and σ – related to the minimum distance between two particles. As it can be seen the potential has a global minimum equal to ε for $r_{min} = 2^{1/6}, \sigma = 1.12\sigma$. An important characteristic of this potential is that the nature of interaction between two particles depends on their distance. Namely, for distances bigger than σ the particles attract each other, while for distances smaller then σ the character of the potential changes to strongly repulsive.
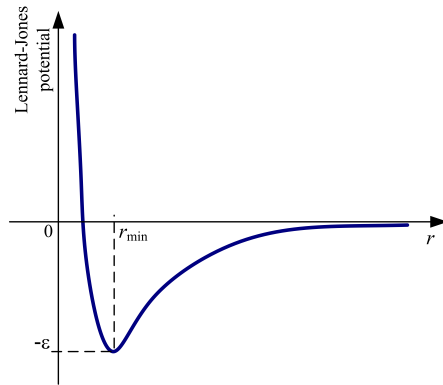

Fig. 1 Lennard–Jones potential

Having the analytical formula for the interaction potential one can easily obtain the formula for the force by simple differentiation, which should be performed analytically to avoid accumulation of numerical approximations. As a first approximation all particles are assumed to have unit mass. Knowing the forces, the Verlet's algorithm may be used to obtain the position and the velocity of each particle in consecutive time steps [22].

The concept of molecular modelling and simulation is used in this study to model the dynamics of a social network. The users who are the nodes of the network become particles in Euclidean space and the distance between particles will be determined based on the relationship strength between the

users. The changes in the distances between particles over time will be the basis for inferring the potential in a purely data-driven way and in consequence for determining the force between particles.

3 Methodology

The approach proposed in this paper is to reformulate the problem of time evolution in social networks and interpersonal relations into the language of multiple–particle interacting system. This is achieved by assigning the position of the node in the social network graph to the position of the particle in a metric space, so that the inter–particle distance reflects the strength of the relation.

In our previous attempts the form of the interacting potential was assumed to reflect the tendency of two particles (nodes) to change their distance in social space. The experimental data was used to extract some characteristic features of the interaction and a modified Lennard–Jones potential was used to reproduce the time evolution at the macroscale. Another possibility to gain an insight into the character of social interaction is to examine in details the distance between each pair of nodes in consecutive time windows. The experimental data i.e. the positions of each particle (node) in each time window are sufficient to obtain the dependence of the variation of the distance between two particles on their distance (see Section 4.2 where the distance matrix is created from the adjacency matrix).

Using only this dependence it is possible to simulate the behaviour of interacting particle ensemble in the following way. Knowing the initial distances between all pairs of particles the change of two-particle distance can be read from the experimental curve. The change of the distance between each two particles can be easily transformed into the displacement vector. The displacement vector has its beginning at initial position of the particle and its end in its final position (it is defined as a difference between the initial and final position vector). Its direction coincides with the direction of a line connecting two particles under consideration and it is pointing into the direction of the centre of mass when the distance between two particles is decreasing, and in opposite direction if the distance is increasing. Because of the equal masses of both interacting particles each of them changes its position by the half of the calculated distance change between them. This procedure allows to define the displacement vector for a considered particle and one of all the other particles from the ensemble. Such an operation should be repeated for all other particles to obtain all displacement vectors for a given particle. Since the displacement vectors calculated in this way represent the forces exerted by other particles, their superposition determines the direction, in which the particle under consideration should be moved. We ignore the magnitude of the total force, as moving any particle by this value, which is the length of a negative gradient of the field potential, would most likely result in overshooting and lack of convergence. Instead, we optimize our system in an iterative

manner, shifting all particles by a small, fixed step at a time until it reaches a steady state. In that way we are able to simulate the time dependence of the position of each particle knowing only the initial positions of all particles and experimentally obtained relation between the change of the distance between two particles and their distance.

The methodology followed in this study is summarised in Figure 2. The consecutive stages of this research are presented below.
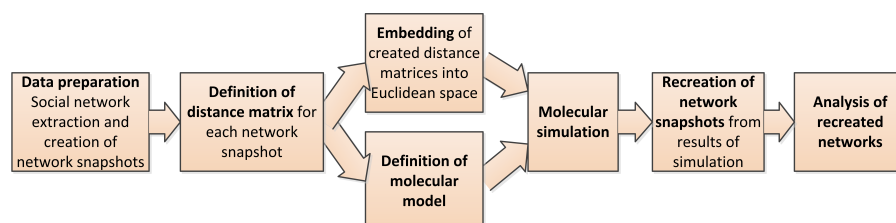

Fig. 2 Methodology followed in this paper

Data preparation: As the real–world evolving network is investigated in this paper, the first step is to prepare data in a way that they can be used in further parts of the experiments. This includes extraction of the interactions and time stamps of their occurrence from email logs dataset and dividing this set into time windows of a given size. From each time frame a single social network is created. Note, that in the case of email communication the underlying social network is directed and weighted. However, adjacency matrix feeded to the embedding process has to be symmetrical. Thus the directed social network is transformed into undirected one by aggregating the communication between every two nodes.

Distance Matrix Creation: Creating a distance matrix for each social network snapshot, in which the distances between nodes reflect the intensity of communication between them, is the next step. The distance needs to be calculated for all pairs of nodes, including the pairs which are not connected directly or at all. In our approach the distance between two particles reflects the length of the shortest weighted path linking the two nodes in question.

Embedding Distance Matrices in Euclidean Space: The goal of this step is to project the created distance matrices into the Euclidean space in a way that the distances between nodes are reflected in the best possible way. After a review of existing embedding methods, Big Bang Simulation and Classic Multidimensional Scaling were chosen. These two methods facilitate embedding of non-metric spaces into almost arbitrary number of dimensions, limited by characteristics of the network under consideration, with moderate computational requirements.

Definition of Molecular Model: In this step the potential field describing the evolution of interactions between nodes is defined. On this basis the whole molecular model of email based social network is created. The potential

function is determined from the changes in distances between two consecutive network snapshots. As the shape of the potential function depends on past data, the force governing the molecular simulation is different for each dataset.

Molecular Simulation: This part of the experiments utilizes the outputs of the previous steps: the embedded social network windows and the potential force which is used to move particles in the Euclidean space. The simulation terminates when the set of particles achieve a stable state.

Recreation of Network Snapshots and Analysis of the Results: The study aims at assessing the characteristics of a network in a stable state, which is an outcome of the molecular simulation. In order to do that, the reverse process to the embedding has to be performed. For each time window the results of simulation, which are the set of nodes' positions in Euclidean space, are taken and the network graph is created based on the distances between nodes (particles) after simulation. This process is straightforward: if the distance between two particles is lower than a given threshold value, the link between these particles in social network is created. The experiments were performed for different values of the distance threshold. Finally, the properties of these retrieved networks are investigated. Two main properties were taken into account: node degree distribution and clustering coefficient.

4 Experiment setup

4.1 Data Preparation — Creating Email based Social Network

The network that has been chosen for experiments, was extracted from the email logs of the Wroclaw University of Technology (WrUT). The experimental data were collected during the period of 21 months (February 2006 – October 2007). The network was created in the course of the data cleansing process and removing fake and external email addresses. The employees of WrUT are the nodes of the network, whereas email messages exchanged between them were used to infer their relationships (edges in the network). Although every single email message provides information about the sender's 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, while the same email sent to 20 people does not. For that reason, the intensity of email communication $I(x,y)$ between email user x and y has been defined as:

$$I(x,y) = \sum_{i=1}^{card(EM(x,y))} \frac{1}{n_i(x,y)} \tag{5}$$

where $EM(x,y)$ is the set of all email messages sent between x and y and $n_i(x,y)$ denotes the number of all recipients of the ith email sent between x and y [26].

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). Although 'to–list' recipients are likely to be of much greater message–network importance than the 'cc-list' recipients, both groups are treated in the same way, i.e. the total number of the recipients of an email is always taken into account. Such approach results from the fact that the obtained data does not include information if the recipient of the email is on the 'to–list' or 'cc–list'.

The resulting social network $SN = \langle N, I \rangle$ is defined as a tuple consisting of a set of network nodes N and a set of relationships that are described by their mean intensity $I : N \times N \rightarrow \mathbb{R}^+ \cup \{0\}$, given by Equation 5. Note that the resulting structure is a non–directed graph with intensity I as a label assigned to the relationships.

It should be emphasized that the social network derived from the email logs does not have a static structure. The existence of any link in such a graph (i.e. relationship) is a result of a series of discrete events (email messages) which occur in certain time instants and usually with changing frequency. We may also think of the computed relationships' intensity as of the social distance between network members (nodes). Greater I reflects smaller distance in the social space. In order to track changes in relationship strength we have used a sliding window approach.

For the experiments the data from a period of 84 days was selected and divided into frames covering seven days each. This allowed to create 12 social network graphs $SN(t_0), SN(t_1), \cdots SN(t_n)$ where $t_0, t_1, \cdots, t_n$ are discrete instants of time. Each network is created according to the procedure defined above on the basis of seven–day period starting in $t_0, t_1, \cdots, t_n$. The networks $SN(t_0), SN(t_1), \cdots SN(t_n)$ are temporal images of evolving social structure which was build on the basis of email communication. In addition only users who were active in all time windows were taken into account as they constitute the core of the network.

4.2 Distance Matrix Creation

The distance between two nodes should reflect their proximity. The most obvious choice – graph distance expressed as the length of the shortest path between the nodes, does not really fit the problem of modelling the dynamics of an email network, especially if the graph is weighted.For example, suppose that the shortest path between nodes x and y has a total weight of 0.7, but it passes through 2 intermediate nodes. At the same time the shortest path between v and w has a total weight of 0.3, but there are no intermediate nodes at all. In the context of an email network it means that v and w communicate directly, but not very often. On the other hand x and y do not communicate directly with each other, but their nearest neighbours do it frequently, and x and y communicate with the neighbours frequently too. In practice it means that x and y may not even know each other, while v and w certainly do. Hence in this case the standard graph distance is misleading and for our

experiments we propose an alternative definition of social distance. Denoting by $D_{EC}(x \leftrightarrow y)$ the number of edges in the shortest undirected path between nodes x and y and by $D_{EW}(x \leftrightarrow y)$ the sum of weights along the same path, normalized to the $(0,1)$ range, the total distance between nodes x and y is given by the following formula:

$$D(x \leftrightarrow y) = \begin{cases} D_{EC}(x \leftrightarrow y) + D_{EW}(x \leftrightarrow y) \; if \; \exists (x \leftrightarrow y) \\ \max(D_{EC}) + \max(D_{EW}) + 1 \; if \; \nexists (x \leftrightarrow y) \end{cases} \quad (6)$$

As a result the distance will always fall into the $(1,2)$ interval for directly connected nodes, $(2,3)$ if there is one intermediate node etc. Note, that in this setting the number of edges in the shortest path contributes the most, while the additional information given by the edge weights is also taken advantage of. Equation 6 also assigns some finite distance value to all pairs of nodes not connected by any path, as one of the requirements imposed by the embedding algorithm we have used was that the distance should be defined for every pair of nodes.

4.3 Embedding Networks in the Euclidean Space

With the distance matrices in place the graphs $SN(t_0), SN(t_1), \cdots, SN(t_n)$ can be embedded in the Euclidean space (two or more dimensional), where each node is represented by a point with given coordinates. The resulting sets of points $SN_0, SN_1, \cdots, SN_n$ represent the temporal network images.

An important issue, which should be discussed here, is the dimensionality of the embedding space. Most embedding algorithms have been designed for the purpose of graph visualization. This naturally implies a two or three dimensional embedding. However, the higher the dimensionality of the embedding, the more accurately the social distances are mapped into the Euclidean space. Figure 3 depicts the average distance distortion[1] of the embedding as a function of dimensionality for the WrUT email network and for (a) BBS, (b) CMDS methods. As expected, in both cases the accuracy of the embedding grows with dimensionality. Please note that the scales on the vertical axes in the Figure 3 are different. It should be emphasized that the pace of accuracy growth with increasing dimensionality is much faster in the case of BBS than CMDS. Intuitively we should choose the number of dimensions to be as high as possible. There is a limit however, which results from the so called 'curse of dimensionality' [5], and especially the 'distance concentration' phenomenon, which as demonstrated in [11] is particularly relevant in the context of dynamic molecular simulation of potential fields in the Euclidean space.

It has been observed that as the number of dimensions grows, the Euclidean distance looses its discriminative power, regardless of the characteristics of the dataset [1,14]. The reason for this is that under a broad set of conditions

---

[1] The distance distortion is defined for each pair as the maximum of the ratio between the original and Euclidean distance and its inverse.

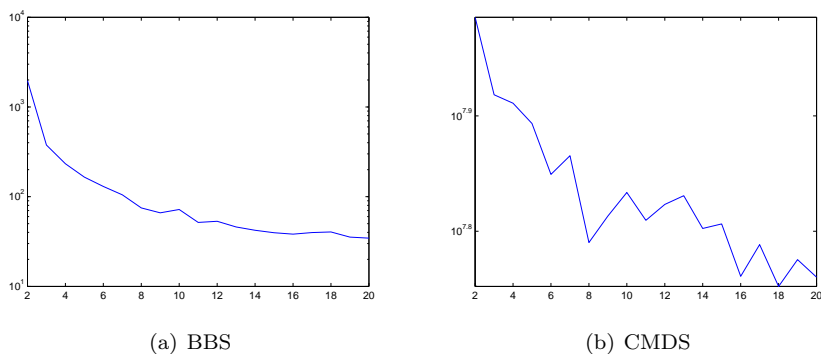(a) BBS   (b) CMDS

Fig. 3 Distance distortion as a function of dimensionality

the mean value of the L_2-norm distribution grows with data dimensionality while the variance remains approximately constant (Figure 4) [14]. As a result, the nearest and furthest neighbours of any molecule appear to be at approximately the same distance, which makes the ratio of distances to the nearest and farthest neighbour tend to converge to 1. As argued in [4], it can occur even for sets with as few as 10 dimensions and the decrease in the ratio between the farthest and nearest neighbour distance is steepest in the first 20 dimensions. The effect is additionally magnified by the limited precision of calculations a computer can handle and often leads to the molecular simulation failing to converge [11]. Hence in practice the embedding dimensionality needs to be a compromise between the distance distortion and negative effects of high dimensionality. For this reason we have decided to embed each graph into $2, 3, \ldots, 20$ dimensions to investigate the mapping between graph distances and distances in embedded graph.

Embedding algorithm has to assure that the Euclidean distances between points (nodes) fit in the best possible way the distances in a social space (relation strengths in original graphs). As a result one obtains the representation of social system in which the network is seen as an assembly of N particles, representing the nodes of a social network.

After reviewing several embedding methods, it has been decided that two sets of experiments will be performed: (i) the Big-Bang Simulation and (ii) CMDS as these methods enable to embed graph into an arbitrary number of dimensions. Additionally, BBS models the network nodes as a set of particles, which is consistent with the next part of the experiments where molecular modelling approach is used to determine the dynamics of a social network.

Embedding was performed on 12 extracted previously social networks. Each of the networks was embedded into $2, 3, \ldots, 20$ dimensions using BBS. CMDS inherently selects the best number of dimensions (in excess of 400 in our case), so in this case the parameter was not set during the experiments, but only first $2, 3, \ldots, 20$ dimensions produced by CMDS have been used in our simulations.
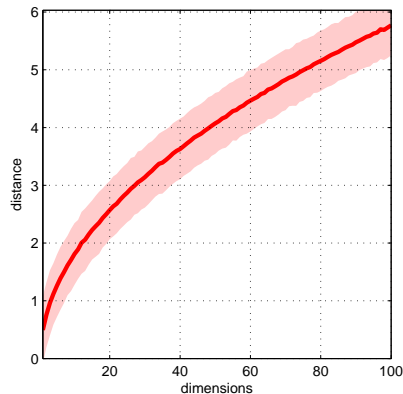
Fig. 4 Distance concentration for the Euclidean norm, for a random vector drawn from a unit hypercube (solid line denotes the mean value, shaded region denotes the mean $+/-$ 2 standard deviations).

For each of the dimensionalities given above we have analysed how well the distances between particles from the social networks (graph) are reflected in the embedded space. To avoid negative effects of high dimensionality we decided to select the lowest number of dimensions that allowed to embed the graph in a way that the mean values of the distances after embedding, which correspond to the graph distances in the ranges $< 1; 2), < 2; 3), < 3; 4)$ etc. were well separated. This has been achieved for 12 dimensions, where for both BBS (Figure 5) and CMDS (Figure 6) the the distributions of distances in the embedding space are approximately unimodal and their expected values are in the required range.

Due to the above, the actual molecular simulation has been performed in 12–dimensional space. However, for the visualisation purposes, where appropriate and to present general idea, the figures were presented for the two–dimensional embedding and molecular simulation.

After selecting the number of dimensions, the next stage of the experiments was to embed the created social networks snapshots into Euclidean space. As discussed in Section 3, embedded graphs serve as an input to the molecular simulation process.

4.4 Setting up the Dynamic Molecular Model

Because the sets of network nodes in $SN(t_0), SN(t_1), \cdots, SN(t_n)$ are equal, each point (node) is represented in any of the sets $SN_0, SN_1, \cdots, SN_n$ and is active in each of the windows. We may think of these points as of particles moving as a result of interactions (email communication) between them. At this point we use the formalism of molecular dynamics to associate a potential
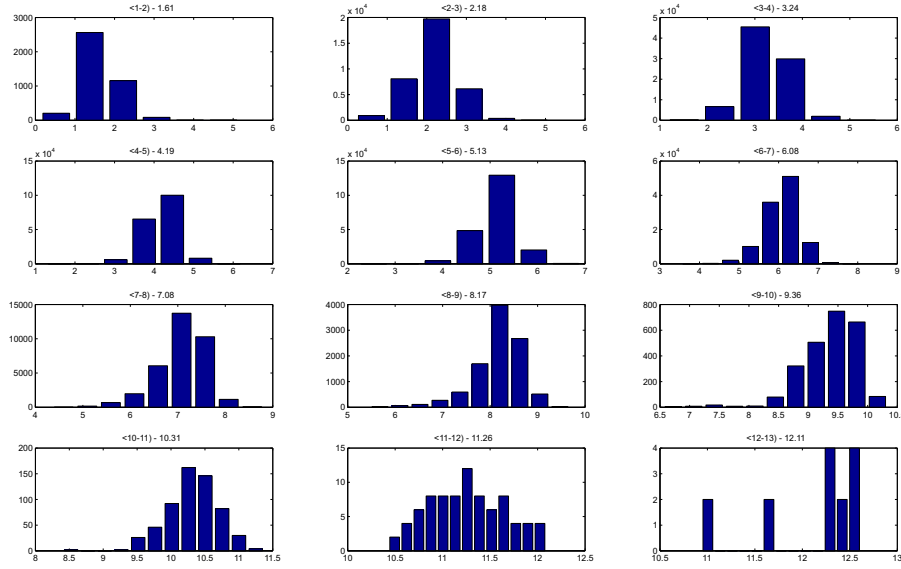
Fig. 5 Mapping between graph distance and distances after embedding graph into 12 dimensions using BBS algorithm.

U with every particle (network node). The actual characteristic of this potential depends on the behaviour of the particles changing their positions in time instants $t_0, t_1, \cdots, t_n$.

First experiments were performed using standard Lennard–Jones potential function [22, 33]. The analysis of server logs has revealed some features of the dynamics of email communication – 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 minimum. We noticed that intense email communication (which results in very small distances in social graph) is never sustained for a longer period of time. On the other hand fading communication is (in most cases) followed by frequent message exchanges.

It should be stressed that the Lennard–Jones potential was used only for the first experiments and did not accurately fit the underlying data. In the experiments presented in this paper the social network–specific potential function on the basis of available data was developed. In order to do that, first the distance transition probability defined as the probability that a given distance in one window will change into another given distance in the next time frame, was calculated. For the 12 time windows, 11 transition probability matrices were obtained. The matrices were then averaged. The resulting final matrix is presented in Figure 7a. The force that governs the changes of the location of the particles in the Euclidean space is proportional to the distance change. Hence the third degree polynomial presented in Figure 7b, inferred from the
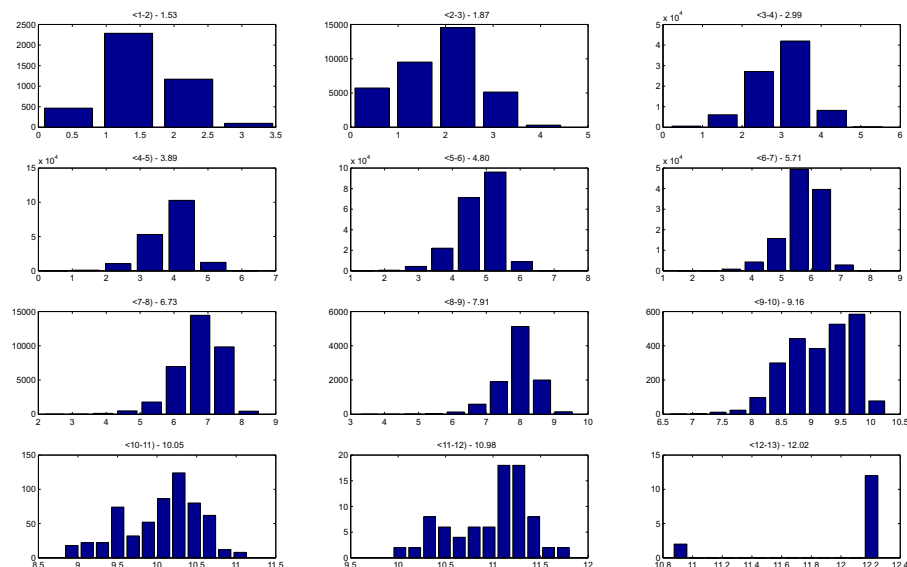
Fig. 6 Mapping between graph distance and distances after embedding graph into 12 dimensions using CMDS algorithm.

distance transition probability, describes the force used in the molecular simulation. Please note that the force will be different for different datasets.
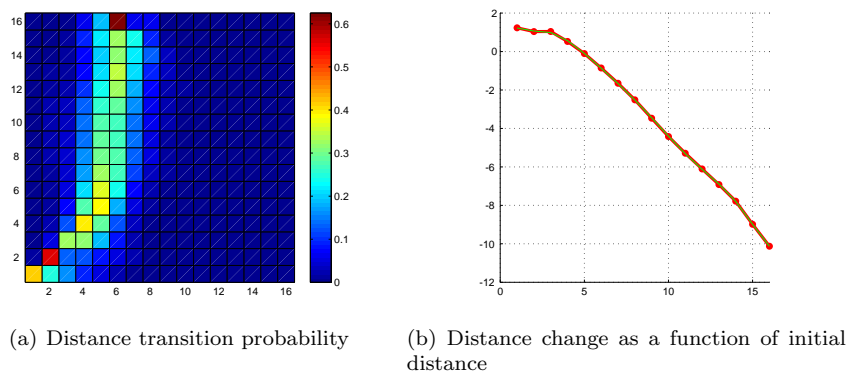

(a) Distance transition probability

(b) Distance change as a function of initial distance

Fig. 7 Graph distance transition probability and expected graph distance change for the WrUT email network

The presented force allows to simulate the changes between communication patterns in consecutive time instants. The potentials associated with the nodes reflect their abilities and tendency to establish future connections with their 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).

5 Molecular Simulation

After the network graphs have been embedded in the Euclidean space and the force function has been established the molecular simulation can be performed. The goal of the simulation is to obtain the network that is in a stable state, i.e. does not change from one simulation step to another. In practice it means that the particles oscillate around the point of equilibrium. The whole process is performed iteratively until oscillation is detected by checking if the mean displacement of all particles between two non–consecutive steps of the simulation is below a threshold value (0.0005 in our experiments). Figure 8 presents the mean displacement changes during the simulation of the first time window, embedded using CMDS. This is just one example of the obtained during the experiments; for each time window and for each embedding algorithms this function will have a different shape and the stop condition will be met for different number of iterations. The common feature of all simulations is that mean displacement converges to 0.
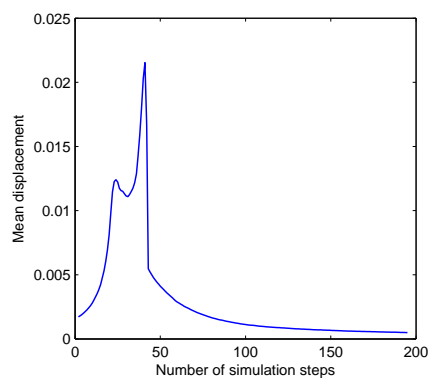

Fig. 8 Mean displacement of all particles between two non–consecutive steps of the simulation as a function of number of simulation steps for social network from Window 1 embedded using CMDS method

Figures 9 and 10 present the results of the molecular simulation in 2D space for windows 2 and 4 respectively (the windows have been selected for illustration purposes). For each of these windows the BBS and CMDS embedding as well as the result of the simulation are presented. Although the embedded graphs look differently for CMDS and BBS, the final outcomes of simulations are similar. These two windows were chosen to present how different shapes of embedded graphs behave during the simulation process.
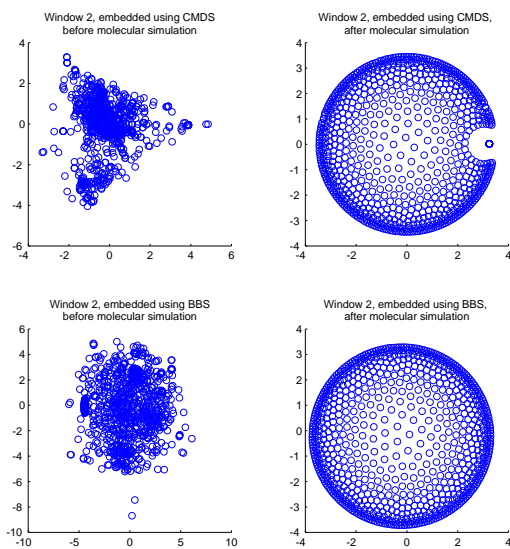
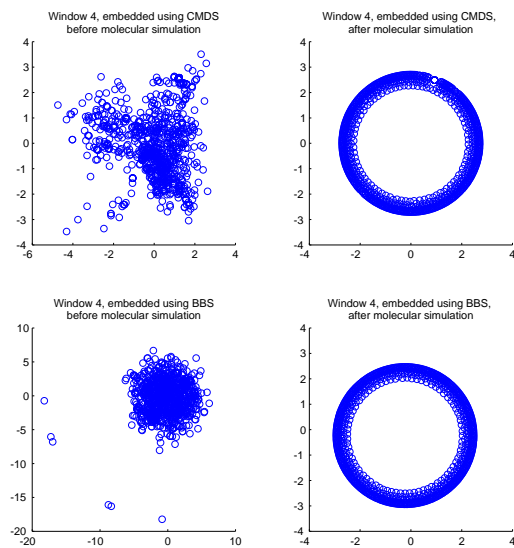**Fig. 9** Window 2 before and after molecular simulation

Fig. 10 Window 4 before and after molecular simulation

6 Reconstruction of Social Networks

The result of the molecular simulation is a set of particle collections in their stable states. In order to investigate the characteristics of obtained structures a reverse–embedding process needs to be performed. During this phase the social

non–metric graph is created from the particles embedded in Euclidean space whose positions were determined during the molecular simulation. The graph is recreated using the pairwise Euclidean distances between the particles.

First, the distance between each pair of particles is calculated. After that a threshold for the distance is set and a link is created between pairs of nodes for which the distance is below this threshold. Each social network was reconstructed using 100 different threshold values. First the difference between maximum and minimum distance in a given time window was calculated and then this number was divided in 99 equal parts. Different values of distance threshold influence the number of links in the recreated network. Figure 11 shows the number of links in the 1st window of social network (embedded using CMDS) as a function of distance threshold. Note that, most of the distances is in the range $(0; 8)$. For each time window and embedding method, 100 networks with different distance threshold were created and these networks and their characteristics are investigated in the next section.
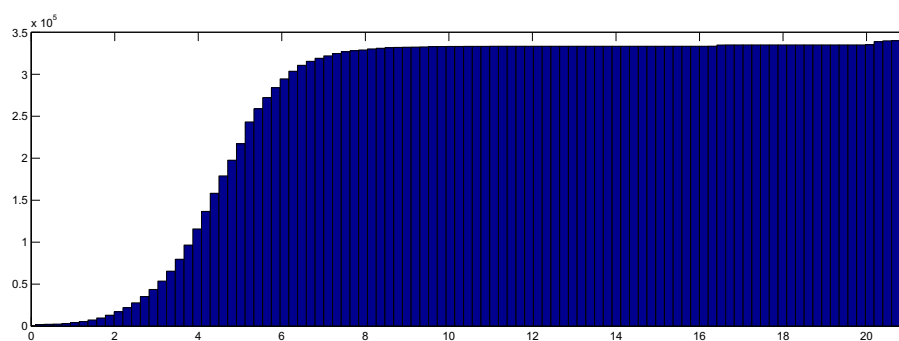

Fig. 11 Number of links in the 1st window of social network (embedded using CMDS) as a function of distance threshold

7 Analysis of the Experimental Results

The recreated networks have been examined with respect to node degree distribution, clustering coefficient and shortest paths. Although there were 1,100 reconstructed networks (100 threshold values times 11 windows) and all three characteristics were calculated for each of them, only a subset of networks was selected for subsequent analysis . As it was pointed out, during the reconstruction process, the value of the distance threshold directly influences the number of links in the recreated network. The general and intuitive rule is that the higher the threshold the denser the reconstructed network. Because the structures analysed in this study are social networks, and according to Dunbar's number every human being can on average maintain 150 meaningful social relations [18], in further analysis we have selected a threshold for each

embedding method and for each time window for the average node degree to be as close to 150 as possible. Please note that the average node degree is meaningful in this case as the node degree distributions follow Poisson distribution. In Figure 12 the node degree distributions for Window 2 for all 100 distance thresholds are presented, with different colours denoting distributions for different distance thresholds. It is clearly visible that the average node degree increases with the growth of the threshold. On the left side of the plot, for threshold equal to 1 the recreated network is empty (probability that a node has 0 edges is 1) and on the other hand for threshold 100 the network is represented by a fully connected graph (probability that a node is connected to all other nodes in a network is 1). For all intermediate thresholds the node degrees oscillate insignificantly around a given number which grows together with the distance threshold. Hence the value of the average node degree can be used to select networks for further analysis.
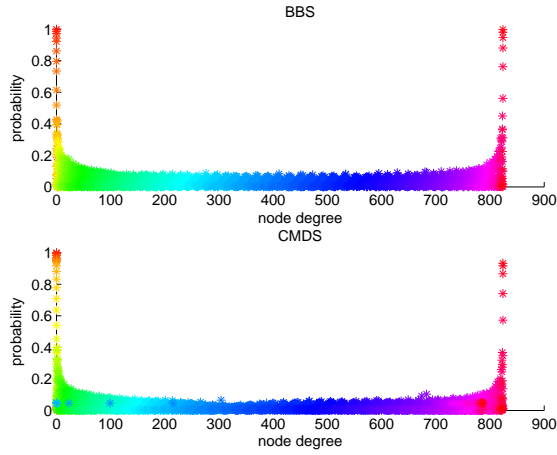

Fig. 12 Node degree distributions for reconstructed social networks from Window 2 for all 100 distance thresholds (different colours represent node degree distributions for different thresholds; distance threshold increases from 1 to 100 looking from left to right).

The distance thresholds chosen for further analysis are presented in Table 1. For all of them, for a given embedding method and for each time window the mean value of node is closest to 150.

Table 1 Distance thresholds for each time window in which average node degree is closest to 150 in comparison to other thresholds.

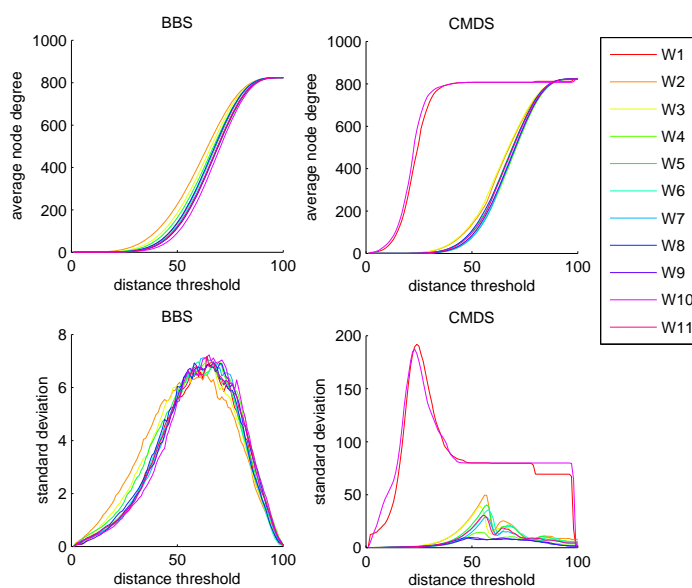| Window no.	1	2	3	4	5	6	7	8	9	10	11
BBS threshold no.	51	45	47	52	49	51	52	51	53	54	52
CMDS threshold no.	17	51	50	56	53	55	56	53	55	16	54

Fig. 13 Average node degree distribution for recreated social networks after the BBS and CMDS embedding and the molecular simulation as a function of distance threshold. Results for Time Window 2.

The mean node degree values for each threshold together with their standard deviation are depicted in Figure 13 for both BBS and CMDS. There is not much variation for the BBS embedding. In the case of CMDS, windows 1 and 10 can be perceived as outliers, which follow a different node degree distribution than the remaining windows. This difference is also visible in Table 1 where in the case of these two windows, the threshold values which result in the average node degree of 150, are considerably smaller then for other windows. The reason for this is that CMDS, due to its dependence on eigendecomposition of the dissimilarity matrix, is very sensitive to outliers, i.e. nodes that are at the periphery of the social network. The result of the embedding and molecular simulation in such a case is that most of the nodes are very close to each other, while a few are very far away. This explains the trend in node degree distributions for Windows 1 and 10 when CMDS was used.

Node degree distributions for the reconstructed networks for previously selected thresholds for each window are presented in Figure 14 (BBS) and in Figure 15 (CMDS). All of the networks (except Windows 1 and 10 for CMDS discussed before) follow Poisson distribution of the node degree. It means that the reconstructed networks that are in the stable state, in terms of node degree distribution, behave like random or small–world network (see Table 4) and not scale–free networks that have power–law node degree distribution. There are no hubs in the the reconstructed networks and the standard deviation of the node degree is low as it does not exceed 8 in the case of BBS and 50 in
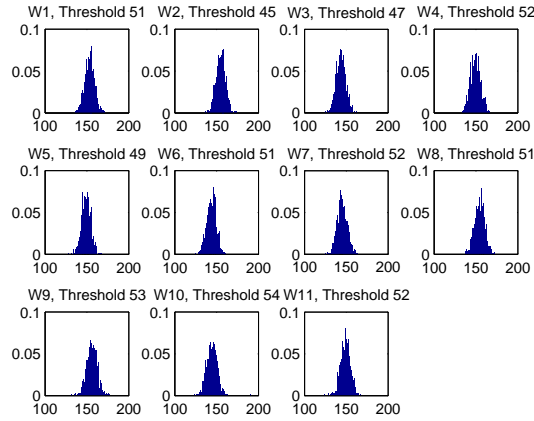
Fig. 14 Node degree distribution for recreated social networks after BBS embedding and the molecular simulation for distance thresholds from Table 1 in different time windows (x axis – node degree; y axis – probability).
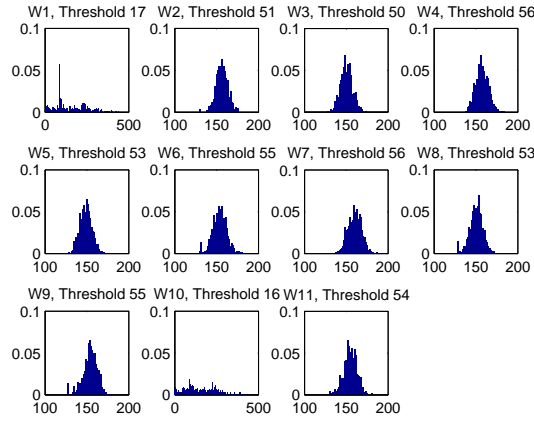

Fig. 15 Node degree distribution for recreated social networks after the CMDS embedding and the molecular simulation for distance thresholds from Table 1 in different time windows (x axis – node degree; y axis – probability).

the case of CMDS (neglecting Windows 1 and 10 where standard deviations peeks at 200). This indicates that the active core of the organisational email-based social network in its equilibrium states resembles a community where everybody has similar number of connections.

Next, we investigate the clustering coefficient (CC). Suppose a node v has neighbours $\mathcal{N}(v)$, with $|\mathcal{N}(v)| = k_v$. At most $k_v(k_v - 1)/2$ edges can exist between them (this occurs when v is part of a k_v–clique). The clustering coefficient of a vertex, CC_v, is defined as the fraction of these edges that actually exist. The clustering coefficient of the graph is defined as the average clustering coefficient of all the vertices in the graph. The distributions of the clustering

Table 2 Average clustering coefficient (ACC) and its standard deviation for the reconstructed networks

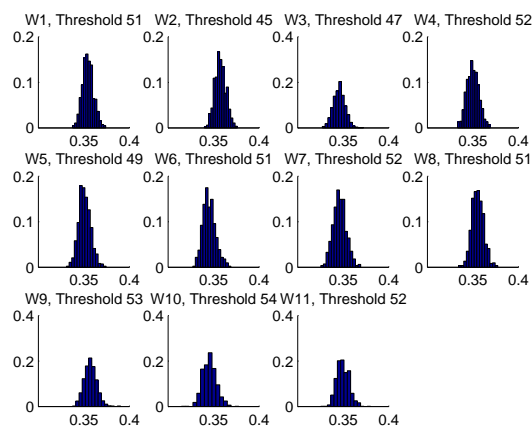| Window no.	1	2	3	4	5	6	7	8	9	10	11
ACC (BBS)	0.35	0.36	0.35	0.35	0.35	0.35	0.35	0.36	0.38	0.35	0.35
Std. Dev. (BBS)	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
ACC (CMDS)	0.70	0.35	0.34	0.35	0.34	0.34	0.35	0.35	0.35	0.70	0.35
Std. Dev. (CMDS)	0.16	0.08	0.07	0.02	0.07	0.07	0.07	0.01	0.01	0.16	0.06

Fig. 16 Clustering coefficient distribution for recreated social networks after the BBS embedding and the molecular simulation for distance thresholds from Table 1 in different time windows (x axis – node degree; y axis – probability).
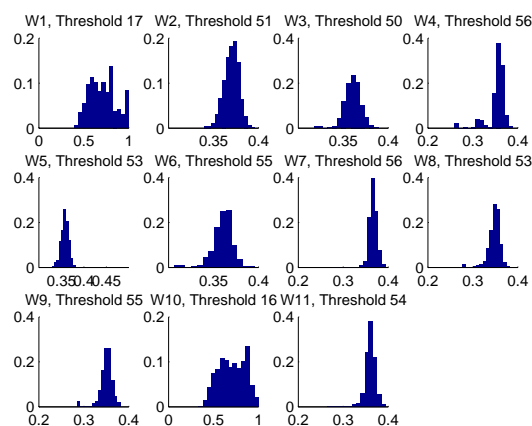

Fig. 17 Clustering coefficient distribution for recreated social networks after the CMDS embedding and the molecular simulation for distance thresholds from Table 1 in different time windows (x axis – node degree; y axis – probability).

coefficient for the selected networks are presented in Figure 16 (BBS) and
Figure 17 (CMDS). Similarly to the node degree distributions, this one also
follows Poisson distribution, i.e. most of the users have similar clustering coef-
ficient (at the level of 0.35 for both BBS and CMDS – see Table 2). Moreover,
the standard deviation of the clustering coefficient is low – 0.01 for BBS for
all windows, for CMDS it mostly varies between 0.01 and 0.08 and reaches its
maximum – 0.16 – for Windows 1 and 10. The clustering coefficient at this
level is characteristic for real–world social network. Comparing these results
with random and ordered networks of the same size (Table 4), it is clear that
all of the recreated networks share the features of both these types of net-
works: their clustering coefficient is larger than the one for random network,
which is 0.18 and smaller than in the ordered network – 0.74. We can conclude
by saying that the analysed networks follow a small–world network model in
terms of clustering coefficient.

Table 3 Average path length and its standard deviation for the reconstructed networks

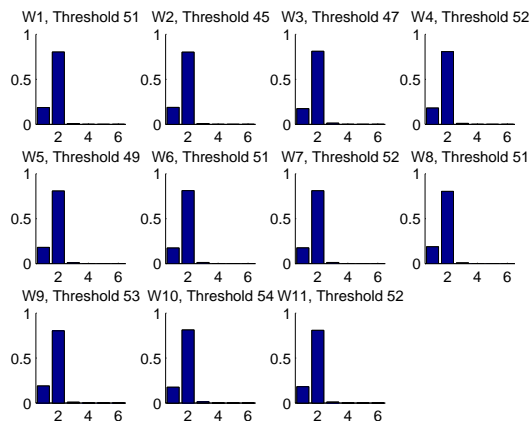| Window no.	1	2	3	4	5	6	7	8	9	10	11
APL (BBS)	1.82	1.82	1.84	1.83	1.83	1.84	1.83	1.82	1.82	1.84	1.83
Std. Dev. (BBS)	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41
APL (CMDS)	2.06	1.81	1.82	1.82	1.82	1.81	1.87	1.83	1.82	2.04	1.87
Std. Dev. (CMDS)	0.71	0.42	0.41	0.41	0.41	0.41	0.48	0.41	0.41	0.73	0.48

Fig. 18 Shortest paths distribution for recreated social networks after the BBS embedding
and the molecular simulation for distance thresholds from Table 1 in different time windows
(x axis – node degree; y axis – probability).

Last analysed characteristic that describes the reconstructed networks in
a comprehensive manner is the length of the shortest path. The experiments
were performed in the same way as in the case of clustering coefficient and
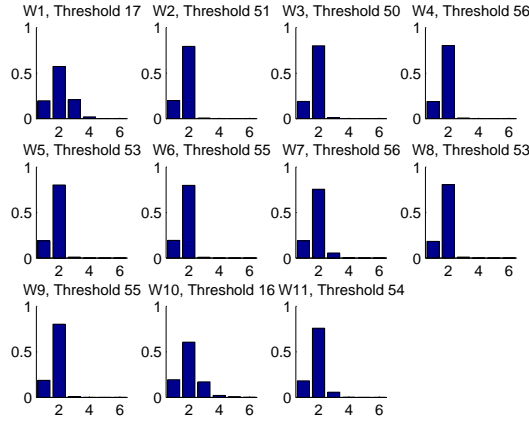the histograms are presented in Figure 18 (BBS) and Figure 19 (CMDS). The

Fig. 19 Shortest paths distribution for recreated social networks after the CMDS embedding and the molecular simulation for distance thresholds from Table 1 in different time windows (x axis – node degree; y axis – probability).

results indicate that the average path lengths (APL) are short for both BBS and CMDS as they are in the range $[1.82; 1.84]$ for BBS and $[1.81; 2.06]$ for CMDS. Also their standard deviation is rather modest: 0.41 for BBS and in the range $[0.41; 0.72]$ for CMDS (see Table 3). This low value of average path length indicates that small-world phenomena, where two people are separated just by few intermediates, is present. Similarly to the clustering coefficient, average path length puts the analysed networks somewhere in between order and randomness. APL is longer than in the case of random network (1.34) and shorter than in an ordered network (2.75), which means that also in regards to average path length the recreated networks are in fact small-worlds (see Table 4).

Table 4 Network models characteristics; $N = 825$ – number of nodes; $k = 150$ – average node degree

Feature	Random Network	Small–world Network	Ordered Network
Clustering Coefficient	Low $\frac{k}{N} = 0.18$	Between order and randomness BBS: $[0.35; 0.38]$ CMDS: $[0.34; 0.70]$	High $\frac{3(k-2)}{4(K-1)} = 0.74$
Average Path Length	Short $\frac{\ln N}{\ln k} = 1.34$	Between order and randomness BBS: $[1.82; 1.84]$ CMDS: $[1.81; 2.06]$	Long $\frac{N}{2k} = 2.75$

The performed analyses revealed that the networks recreated after the molecular simulation are small–world networks. They follow Poisson node degree distribution, have big clustering coefficient and small average path length. Molecular simulation terminates when the system achieves stable state. We showed that networks reconstructed after the simulations feature the three,

enumerated above characteristics of social networks. The fact that the final networks are small–world ones and resemble typical characteristics of real-world social networks can be an indication that molecular simulation can be a new way of generating this type of networks and may be effectively applied in sociodynamical analysis.

8 Conclusions

We have proposed to model the dynamics of a complex social system using molecular simulation, where the interactions between the individuals are determined from the data in a form of a social force, which corresponds to the particle interaction force used in the simulation. In our case the social relation was defined on the basis of communication events (message exchange) recorded in the computer system (email server). This allowed to define a social distance as inversely proportional to the number of messages sent between users and to estimate the character of the social force determining the changes of social distance. It was also shown, that the global dynamics of such system may be modeled by treating the users as interacting particles embedded in an Euclidean social space. The movement of particles is determined by the social force and their trajectories are determined by their initial positions, derived from the email server logs and allowing to create the social network.

To the best of our knowledge this is the first attempt to apply a molecular modelling approach to the problem of social network dynamics. It has hence required careful verification, especially with respect to representation of the network evolutionary processes and chosen network structural properties, commonly used in network analysis. The experiments have shown that the proposed approach allows to reason about structural properties of evolving social network, while benefitting from the algorithmic simplicity of molecular modelling.

In this study we have presented the whole process of building and using a molecular model, while identifying the following key points:

– The embedding procedure projecting the non-metric social graph into the Euclidean space should be chosen with care, taking into account the inherent trade-off between preserving the distances from social graph with the required accuracy and limiting the dimensionality of the Euclidean space. This has proven to be especially difficult for network hubs, regardless of the embedding method used.
– The character of social force leading to changes in social distances can be generalized, however this process is inherently connected with the loss of information in the case of individuals who behave statistically differently from the mean pattern (typical behaviour) derived from the whole network data.
– The molecular model of social dynamics allows to reconstruct the social network from positions of the users (moving particles) in an Euclidean social space. While the reconstructed network preserves some of the global

characteristics, local properties at the level of individual nodes usually cannot be recovered.
- The reconstructed social network follows the small–world network model with large clustering coefficient and small average path length.

Our work aimed to show and demonstrate the possibilities and limitations of constructing an evolving sociodynamic model which is inherently data-driven and shows explanatory power. The key challenge was to establish a link between various approaches used in sociodynamics and network mining techniques which are based on the data acquired directly from contemporary computer-based social systems. We conclude that, using a modified molecular dynamic method, it is possible to create the evolutionary model of complex computer-based social network, but its applicability is restricted only to certain network properties measured in social network analysis. Taking into account, that there are few works dealing with the predictive modeling of complex social networks, we find these results promising and forming a basis for the future experiments and development of data-driven evolutionary network models. The obedience of Internet-based social networks provides a huge amount of data for the analysis, changing the paradigms for the description of behavioral changes based on computer-supported social interaction processes.

References

1. Aggarwal, C., Hinneburg, A., Keim, D.: On the surprising behavior of distance metrics in high dimensional space. In: J. Bussche, V. Vianu (eds.) Database Theory – ICDT 2001, *Lecture Notes in Computer Science*, vol. 1973, pp. 420–435. Springer (2001)
2. Barabasi, A.L.: Linked: How Everything Is Connected to Everything Else and What It Means. Plume (2003)
3. Barrat, A., Barthelemy, M., Vespignani, A.: Dynamical processes on complex networks. Cambridge University Press (2008)
4. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful. In: C. Beeri, P. Buneman (eds.) Databases Theory – ICDT 1999, *Lecture Notes in Computer Science*, vol. 1540, pp. 217–235. Springer (1999)
5. Bishop, C.: Neural Networks for Pattern Recognition. Oxford University Press, New York, USA (1995)
6. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics Reports **424**(4-5), 175–308 (2006)
7. Bollobas, B.: Random Graphs. Academic, London, UK (1985)
8. Braha, D., Bar-Yam, Y.: From centrality to temporary fame: Dynamic centrality in complex networks. Complexity **12**, 59–63 (2006)
9. Bringmann, B., Berlingero, M., Bonch, F., Gionis, A.: Learning and predicting the evolution of social networks. IEEE Intelligent Systems **25**(4), 26–35 (2010)
10. Bronstein, M., , Kimmel, R.: Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. Proc. National Academy of Sciences **103**(5), 1168–1172 (2006)

11. Budka, M., Gabrys, B.: Electrostatic field framework for supervised and semi-supervised learning from incomplete data. Natural Computing **10**, 921–945 (2011). URL http://dx.doi.org/10.1007/s11047-010-9182-4. 10.1007/s11047-010-9182-4

12. Davis, D., Lichtenwalter, R., Chawla, N.: Supervised methods for multi-relational link prediction. Social Network Analysis and Mining pp. 1–15 (2012). DOI 10.1007/s13278-012-0068-6. URL http://dx.doi.org/10.1007/s13278-012-0068-6

13. Epstein, J.: Why model? Journal of Artificial Societies and Social Simulation **11**(4) (2008)

14. Francois, D., Wertz, V., Verleysen, M.: Non–Euclidean metrics for similarity search in noisy datasets. In: Proceedings of the European Symposium on Artificial Neural Networks, pp. 339–334. d–side publications (2005)

15. Garton, L., Haythornthwaite, C., Wellman, B.: Studying online social networks. Journal of Computer-Mediated Communication **3**(1), 0 (1997)

16. Harel, D., Koren, Y.: Graph drawing by high-dimensional embedding. Journal of Graph Algorithms and Applications **8**(2), 195–214 (2004)

17. Helbing, D.: Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes. Springer (2010)

18. Hill, R.A., Dunbar, R.I.M.: Social network size in humans. Human Nature **14**(1), 53–72 (2002)

19. Hill, S., Braha, D.: Dynamic model of time-dependent complex networks. Physical Review E, arXiv:0901.4407v2 **82** (2010)

20. Holland, J.: Hidden Order: How Adaptation Builds Complexity. Basic Books (1996)

21. Juszczyszyn, K., Budka, M., Musial, K.: The dynamic structural patterns of social networks based on triad transitions. In: Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, pp. 581–586 (2011). DOI 10.1109/ASONAM.2011.50. URL http://dl.acm.org/citation.cfm?id=2055729

22. Juszczyszyn, K., Musial, A., Musial, K., Brodka, P.: Molecular dynamics modelling of the temporal changes in complex networks. In: IEEE Congress on Evolutionary Computing, Trondheim, Sweden, pp. 553–559. IEEE Computer Society Press (2009)

23. Juszczyszyn, K., Musial, K., Budka, M.: Link prediction based on subgraph evolution in dynamic social networks. In: The Third IEEE International Conference on Social Computing (SocialCom 2011), pp. 27–34 (2011). DOI 10.1109/PASSAT/SocialCom.2011.15. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=06113091

24. Juszczyszyn, K., Musial, K., Budka, M.: On analysis of complex network dynamics changes in local topology. In: The Fifth Workshop on Social Network Mining and Analysis co-located with The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SNA-KDD) (2011)

25. Kashoob, S., Caverlee, J.: Temporal dynamics of communities in social bookmarking systems. Social Network Analysis and Mining **2**, 387–404 (2012). DOI 10.1007/s13278-012-0054-z. URL http://dx.doi.org/10.1007/s13278-012-0054-z

26. Kazienko, P., Musial, K., Zgrzywa, A.: Evaluation of node position based on email communication. Control and Cybernetics **38**(1), 67–86 (2009)

27. Kolaczyk, E.: Statistical Analysis of Network Data. Springer (2009)

28. Kruskal, J.B., Wish, M.: Multidimensional Scaling, Sage University Paper series on Quantitative Application in the Social Sciences. Sage Publications (1978)

29. Kumar, R., Novak, J., Tomkins, A.: Microscopic evolution of social network. In: The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press (2006)

30. Lescovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2008)

31. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology **58**(7), 1019–1031 (2007)

32. Malarz, K., Gronek, P., Kulakowski, K.: Zaller-deffuant model of mass opinion. Journal of Artificial Societies and Social Simulation **14**(1) (2011)

33. Musial, A., Juszczyszyn, K., Musial, K., Brodka, P.: Utilizing dynamic molecular modelling technique for predicting changes in complex social networks. In: IEEE/WIC/ACM

Joint International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Press, pp. 1–4 (2010)

34. Sarr, I., Missaoui, R.: Managing node disappearance based on information flow in social networks. Social Network Analysis and Mining pp. 1–13 (2012). DOI 10.1007/s13278-012-0071-y. URL http://dx.doi.org/10.1007/s13278-012-0071-y

35. Schweitzer, F.: Brownian Agents and Active Particles – Collective Dynamics in the Natural and Social Sciences. Springer Series in Synergetics. Springer (2003)

36. Shavitt, Y., Tankel, T.: Big-bang simulation for embedding network distances in euclidean space. IEEE/ACM Transactions on Networking **12**(6), 993–1006 (2004)

37. Shaw, B., Jebara, T.: Minimum volume embedding. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (2007)

38. Shaw, B., Jebara, T.: Structure preserving embedding. In: Proceedings of the 26th International Conference on Machine Learning (2009)

39. Strogatz, S.H.: Exploring complex networks. Nature **410**(6825), 268–276 (2001)

40. Torgerson, W.: Multidimensional scaling of similarity. Psychometrika **30**(4), 379–393 (1965)

41. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)

42. Watts, D.: Small Worlds: Dynamic of Networks between Order and Randomness. Princeton University Press (2002)

43. Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–444 (1998)

44. Weidlich, W.: Physics and social science - the approach of synergetics. Physics Reports **1**(204), 1–163 (1991)

45. Wolfram, S.: Theory and Applications of Cellular Automata. World Scientific (1986)