MIXING PATTERNS IN A LARGE SOCIAL NETWORK*

Andrzej Grabowski[a], Robert Kosiński[a,b]

[a]Central Institute for Labour Protection — National Research Institute
Czerniakowska 16, 00-701 Warsaw, Poland

[b]Faculty of Physics, Warsaw University of Technology
00-662 Warsaw, Poland

(Received April 29, 2008)

We study mixing in a large real social network consisting of over one million individuals, who form an Internet community and organise themselves in groups of different sizes. We consider mixing according to discrete characteristics such as gender and scalar characteristics such as age. On the basis of the users' list of friends and other data registered in the database we investigate the structure and time development of the network. We found that in the network under investigation assortative mixing is observed, *i.e.* the tendency for vertices in network to be connected to other vertices that are like them in some way.

PACS numbers: 89.75.Da, 89.75.Hc, 89.65.Ef

1. Introduction

In recent years investigations of complex networks have attracted the physics community's great interest. It was discovered that the structure of various biological, technical, economical, and social systems has the form of complex networks [1]. The short length of the average shortest-path distance, the high value of the clustering coefficient and the scale-free distribution of connectivity are some of the common properties of those networks [1, 2]. Social networks, which are an important example of complex networks, have such properties, too.

The advent of modern database technology has greatly advanced the statistical study of networks. The vastness of the available data sets makes this field suitable for the techniques of statistical physics [2]. The study of

* Presented at the XX Marian Smoluchowski Symposium on Statistical Physics, Zakopane, Poland, September 22–27, 2007.

the statistical properties of social networks, *e.g.* friendship networks, is still very difficult. It is possible to assess the form of distribution of connectivity as a result of a survey, like in the case of web human sexual contacts [3]. However, other important properties describing the structure of the network are much more difficult to obtain, as a result of the lack of data on the whole network. A survey often provides data on a small sample of the whole network only.

Progress in information technology makes it possible to investigate the structure of social networks of interpersonal interactions maintained over the Internet. Some examples of such networks are e-mail networks [4], blog networks [5] and web-based social networks of artificial communities [6, 7]. The aim of this work is to introduce a data set describing a large social network of an Internet community (*Grono*), which consists of more than 10^6 individuals. The Grono (cluster) project was started in Poland in 2004 on the website `www.grono.net`. During its 36 months of existence, it has grown into a well-known social phenomenon among Internet users in Poland. Membership is strictly invitation only; existing members can invite an unlimited number of friends to the network via email, who, if they choose to do so, join the network by an initial link connecting to the person who invited them. All users can add, by mutual consent, and remove other people to their databases of friends. In this way undirected friendship network is formed. We show that the structure this network has similar properties to those of other social networks. In order to study the structure of the network, we analysed data containing *e.g.* the list of all friends, age and gender of an individual. The network under consideration consisted of a collection of individuals (network nodes) connected among one another by friendly relationships (network links). In our work we consider mixing according to discrete characteristics such as gender and scalar characteristics such as age.

2. Results

Basic network measures of the whole network and a Giant Component (GC) [1] are presented in Table I. The network consists of $N = 1002182$ individuals and the Giant Component (GC) contains almost all individuals (994381 individuals); only 7801 individuals do not belong to GC. The number of females (F) registered in the web-service is larger than the number of males (M). Females constitute 51% of the population. The value of the clustering coefficient C is two orders of magnitude larger than that of a random graph (RG). The average path length $\langle l \rangle$ in GC is very small and only slightly greater than that in a random graph. A high value of the clustering coefficient and a short average path length $\langle l \rangle$ are characteristic features of social networks [2, 8, 9]; they are typical for small-world networks [10].

TABLE I

Average properties of the whole network and the giant component (GC) and comparison with a random graph (RG) with the same number of nodes N and the same average connectivity $\langle k \rangle$. Only 7801 individuals do not belong to GC.

	N	C	$\langle l \rangle$	$\langle k \rangle$
Network	1002182	0.2	4.3	46.3
GC	994381	0.2	4.3	46.3
RG	994381	0.001	4.0	46.3

The degree distribution of the network for females (F) and males (M) is plotted in Fig. 1. In both cases the graph shows power law regime $P(k) \sim k^{-\gamma}$ with $\gamma_{\mathrm{F}} = \gamma_{\mathrm{M}} = 0.75$ for low k ($k < 100$). Such a power law is common in many types of networks [1], also in social networks [3, 4, 6]. Thus, such a relation can be a consequence of the fact that in the case of individuals with low k the majority of the links represent genuine preexisting social acquaintances. However, for large k ($k > 100$) the degree distribution has exponential form $P(k) \sim e^{-0.01k}$ with the same value of constant in the exponent for (F) and (M) (Fig. 1 (b)). This result is quite different from that

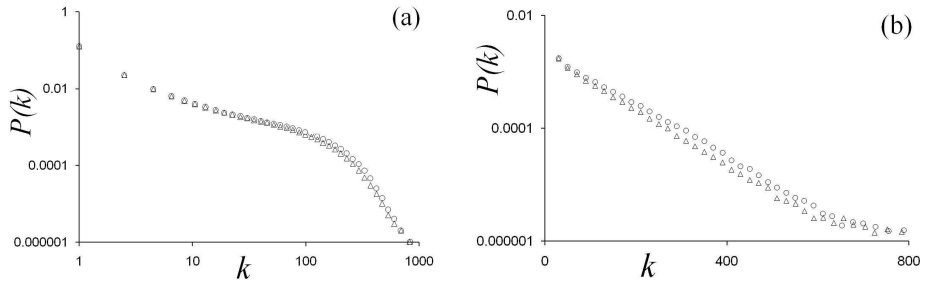

Fig. 1. The degree distribution for females (circles) and males (triangles). In the case of low k results can be approximated with power law $P(k) \sim k^{-\gamma}$, where $\gamma = 0.75$. For large k ($k > 100$) the degree distribution has exponential form $P(k) \sim e^{-ak}$, where $a = 0.01$.

presented in Ref. [6], where degree distribution exhibits two power scaling regimes separated by a critical degree. We suggest, that the discrepancy between results explains the relatively short age and small size (12×10^3 nodes) of the network presented in Ref. [6]. The form of degree distribution indicates the existence of two underlying networks that are simultaneously present in the social network under investigation: the network of links that represent preexisting social contacts for low k and the network of links represent social contacts maintained only via Internet for large k. It is visible that in the

regime of large k the values of degree distribution are slightly larger in the case of females than males. In the consequence there is visible discrepancy in value of the average connectivity, which equals $k_{\mathrm{F}} = 20.5$ and $k_{\mathrm{M}} = 17$ for females and males, respectively.

The network under investigation is a growing network, a new individual can join when he or she is invited. Each individual can invite to the network unlimited number of friends. It should be noted, that the invitation links represent a part of preexisting social acquaintances. The average number of invited friends equals $k_{\mathrm{F}}^{\mathrm{I}} = 1.1$ and $k_{\mathrm{M}}^{\mathrm{I}} = 0.8$ for females and males, respectively. These results show that females are clearly more active in this artificial society. They are more prone to use the web-service in order to communicate with friends and to make new acquaintances.

The average connectivity of the nearest neighbours k_{NN} of a node with k connections for females and males is shown in Fig. 2. It can be seen that the greater the k, the greater the k_{NN}. Hence, the network under investigation is assortative mixed by degree; such a correlation is observed in many social networks [12]. In social networks it is entirely possible, and is often assumed in sociological literature, that similar people attract one another. On the other hand individuals organise in groups of different sizes in order to chat together. Since it is highly probable that members of a group are connected to other members, the positive correlations between degrees may at least in part reflect the fact that the members of a large (small) group are connected to the other members of the same large (small) group. The relation $k_{\mathrm{NN}}(k)$ can be approximated by power-law relation $k_{\mathrm{NN}}(k) \sim k^{0.2}$ with the same value of exponent in the case of females and males. It should be noted that similar value of the exponent (0.18) was found in other social network [11].

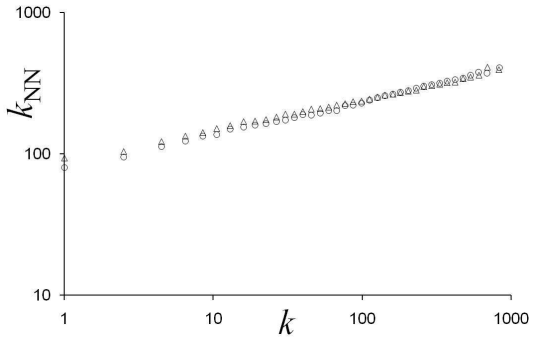

Fig. 2. The relation between the average connectivity of the nearest neighbours k_{NN} of a node and its connectivity k for females (circles) and males (triangles). The relation can be fit to power law $k_{\mathrm{NN}}(k) \sim k^{0.2}$.

The behaviour of the clustering coefficient C is an interesting problem. Fig. 3 plots the correlation between the local clustering coefficient $C(k)$ and the node degree k, showing the existence of a power law $C(k) \sim k^{-\alpha}$ with $\alpha_{\mathrm{F}} = 0.32$ and $\alpha_{\mathrm{M}} = 0.36$, for females and males respectively. Similar value of the exponent α has been observed in other social networks ($\alpha = 0.33$ [6] and $\alpha = 0.44$ [11]). The power-law relation $C(k)$ is similar to the relationship observed in hierarchical networks [13]. Such power laws hint at the presence of a hierarchical architecture: when small groups organise themselves into increasingly larger groups in a hierarchical manner, local clustering decreases on different scales according to such a power law. This may be connected with the fact that individuals can freely make acquaintances, without barriers connected with spatial distance between individuals. The influence of spatial distance between nodes in Euclidean growing scale-free networks on $C(k)$ relation is described in Ref. [14].

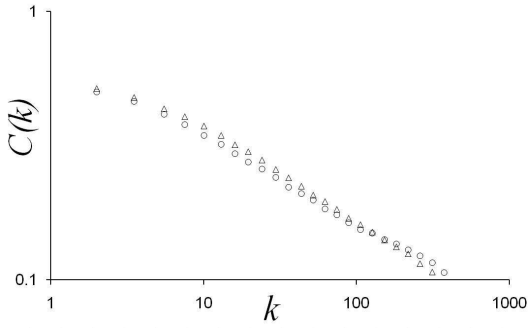

Fig. 3. The relation between the clustering coefficient of a node $C(k)$ and its connectivity k for females (circles) and males (triangles). The relation can be fit to power law $C(k) \sim k^{-0.32}$ (females) and $C(k) \sim k^{-0.36}$ (males).

The value of the clustering coefficient decreases faster in the case of males, however the average value of the clustering coefficient is greater and equals $C_{\mathrm{F}} = 0.19$ and $C_{\mathrm{M}} = 0.2$, for females and males, respectively. Males often are in small, but highly interconnected groups of friends. On the other hand females are more likely to find new friends (*cf.* Fig. 1), therefore they are less clustered.

Important factor significantly influencing the evolution and the form of social network is an age A of each individual. The age distribution is shown in Fig. 4 (a). It is visible that females registered in the web-service are generally younger than males. The average age equals $A_{\mathrm{F}} = 20.5$ and $A_{\mathrm{M}} = 22$ years, for females and males, respectively. The average age of nearest neighbours A_{NN} is highly correlated with the age of an individual (see Fig. 4 (b)). In the range of age between 13 and 27 years (almost 90% of

users are in this range) the A_{NN} increases approximately linearly with age of an individual increasing ($A_{\mathrm{NN}} \sim 0.7 \times A_{\mathrm{F}}$ and $A_{\mathrm{NN}} \sim 0.6 \times A_{\mathrm{M}}$, in the case of females and males, respectively). In the network under investigation also correlation between age of an individual and its connectivity are visible (Fig. 4 (c)). The maximal number of connections is observed for $A = 16$ years. This is so because young people use the web-service to communicate with their schoolfellows and invite most of their classmates.

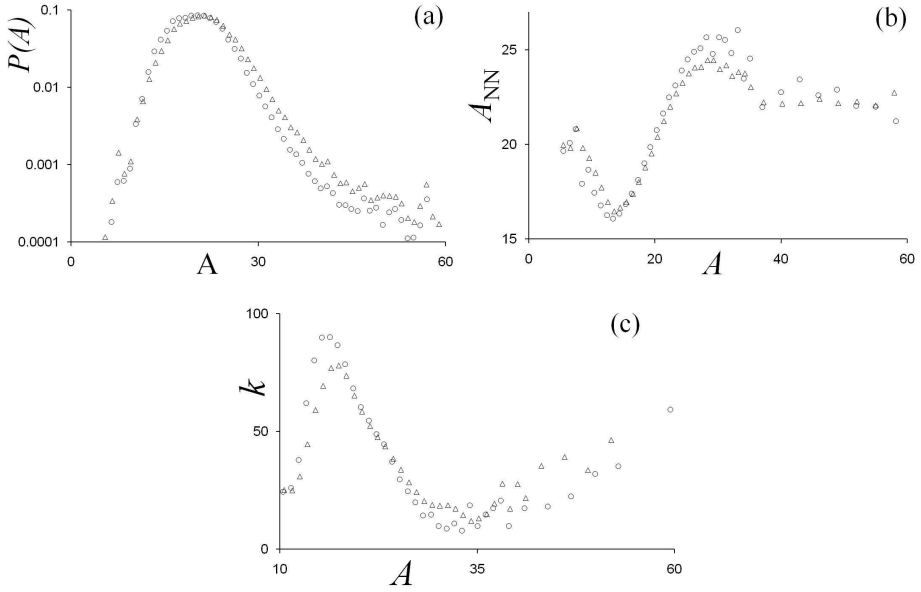

Fig. 4. Age distribution (a), the relation between the average age of the nearest neighbours of a node A_{NN} (b), connectivity of a node k (c) and its age A for females (circles) and males (triangles).

The probability that connection between individual will be created depends also on their gender. Fig. 5 illustrates the relation between connectivity of an individual and the probability p_{G} that the neighbour has the same gender. In the case of females the value of the probability p_{G} decreases with k increasing and is larger than in the case of males. It is visible that the behaviour of females with low k is different than females with large k. Females with small number of friends prefer to make acquaintances with other females and up to 63% of friends are females. On the other hand approximately 60% of friends of females having large k are males. Similar tendency for large k is observed in the case of males, however the value of the probability $p_{\mathrm{G}} < 0.5$ irrespective of k.

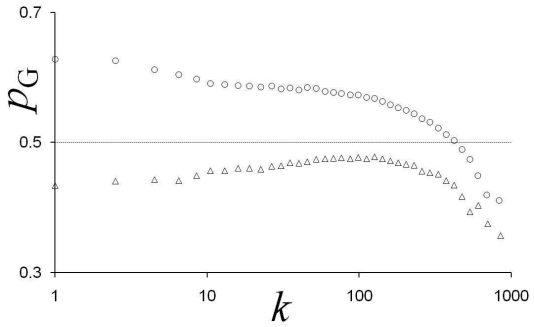

Fig. 5. The relation between probability p_G that a neighbour has the same gender for females (circles) and males (triangles).

The website contains additional services such as discussion forums; hence users can organise themselves into groups of different sizes. Each group is called *Grono* (cluster), and members of that group chat together on a single topic (music, movies, health, politics, *etc.*). Each individual can join to unlimited number of groups. Because an individual can be a member of many groups it is interesting to plot the distribution of number of groups of an individual n_G (see Fig. 6). Most of individuals belong to small number of groups, $n_G < 10$. For greater values of n_G, the probability decreases abruptly and the distribution can be fit to a power-law. It is visible that the number is highly correlated with gender and the average number of groups equals $n_G^F = 4.7$ and $n_G^M = 3.5$, for females and males, respectively. Therefore females are much more prone to join to new groups than males. The number n_G is also positive correlated with degree of an individual.

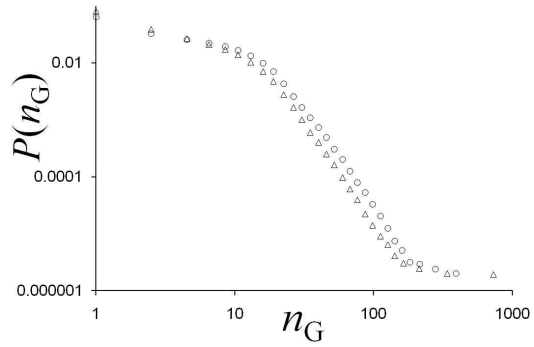

Fig. 6. Distribution of number of groups of an individual for females (circles) and males (triangles). The relation can be fit to power-law $P(n_G) \sim n_G^{-3.3}$ (females) and $P(n_G) \sim n_G^{-3.4}$ (males).

3. Human dynamics

On-line communities offer a great opportunity to investigate human dynamics, because much information about individuals is registered in databases. To analyse how long people are interested in a single task, we studied creation date and last login date registered in the database. Individuals can lose interest in using the website after some time. We consider an individual as active when it regularly uses web services, and we consider an individual as inactive when it do not login more than one month. The distribution of probability P_{L} that an inactive individual has the lifespan T_{L} for females and males is shown in Fig. 7 (a). The lifespan of an individual T_{L} is defined

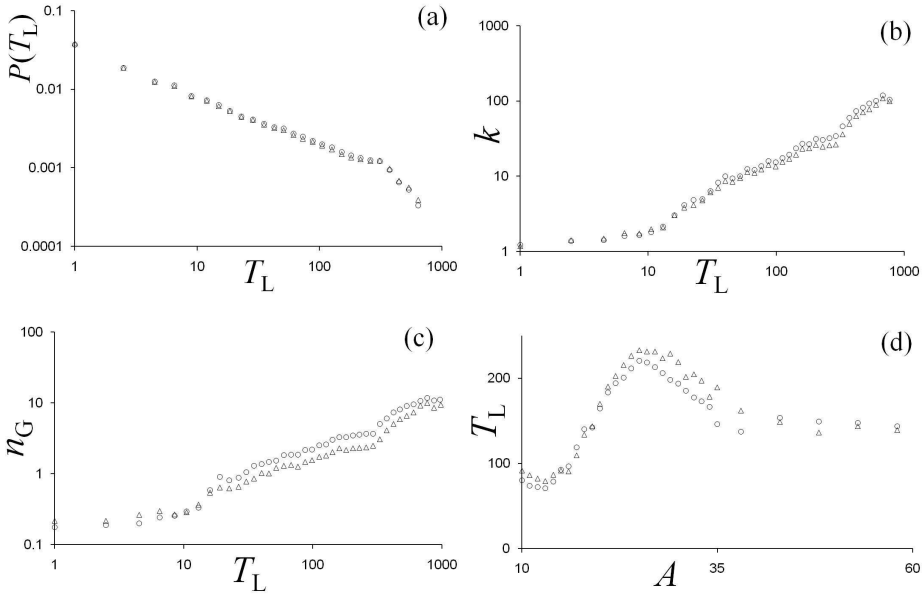

Fig. 7. The distribution of probability P_{L} that activity of an individual in the community last T_{L} days (a). Average time T_{L} equals 200 days. The relation between lifespan T_{L} of an individual and its connectivity k (b) and number of groups n_{G} (c). Correlations between age and lifespan are depicted in (d). Data are denoted by circles and triangles for females and males, respectively.

as the number of days since the time of an individual was added (invited to the network) to the date of last logging. This distribution can be approximated with the power-law $P_{\mathrm{L}}(T_{\mathrm{L}}) \sim T_{\mathrm{L}}^{-0.6}$ (with the value of the exponent irrespective of gender). Thus, the probability that a human will devote the time t to a single activity has a fat-tailed distribution. Many individuals lose interest in using the website very shortly, so they cannot make new ac-

quaintances. Therefore, abrupt decrease in $P(k)$ for very low k is visible (see Fig. 1). The average lifespan equals $T_{\mathrm{L}}^{\mathrm{F}} = 194$ and $T_{\mathrm{L}}^{\mathrm{M}} = 204$ for females and males, respectively.

The time development of the connectivity of an individual, *i.e.* the relation between lifespan of an individual and its connectivity, is shown in Fig. 7 (b). The relation $k(T_{\mathrm{L}})$ can be approximated by power-law relation $k(T_{\mathrm{L}}) \sim T_{\mathrm{L}}^{0.92}$ for females and $k(T_{\mathrm{L}}) \sim T_{\mathrm{L}}^{0.88}$ for males. Thus, the number of friends increases faster in the case of females. It should be noted that similar value of the exponent (0.8) was found in other social network [11].

We investigate the time development of number of groups of an individual and the results are shown in Fig. 7 (c). The relation $n_{\mathrm{G}}(T_{\mathrm{L}})$ can be approximated with power-law relation $n_{\mathrm{G}}(T_{\mathrm{L}}) \sim T_{\mathrm{L}}^{0.73}$ for females and $n_{\mathrm{G}}(T_{\mathrm{L}}) \sim T_{\mathrm{L}}^{0.7}$ for males.

The lifespan is not correlated with the gender, however strong correlation between lifespan and age of an individual are observed. Fig. 7 (d) illustrates the relation between age of an individual and its lifespan for inactive users. The maximum value of T_{L} is observed for $A \approx 25$. It is interesting result, that for $A > 35$ the lifespan is independent on age.

It should be stressed that we have found such a distribution of lifespan, with different values of the exponent (-1.0) in other social network [11]. Similar relations concerning human dynamics have also been observed elsewhere [15] and can be a consequence of a decision-based queueing process. The model of such a process was recently proposed by Barabási [16–18]. It indicates that scale-free distributions are common in human dynamics.

4. Conclusions

In conclusion, we have shown that a friendship network maintained in the on-line community has similar properties (*e.g.* large clustering, a low value of the average path length, assortative mixing by degree and a scale-free distribution of connectivity) to other social networks. Individuals organise themselves into groups of different sizes. The groups size distribution and distribution of number of groups of an individual have power-law form. On the basis of creation date and last login of each individual recorded on the server, we have presented results concerning human dynamics. The power-law form of distributions $P_{\mathrm{L}}(T_{\mathrm{L}})$, $k(T_{\mathrm{L}})$, $n_{\mathrm{G}}(T_{\mathrm{L}})$ and other authors' [11, 16] results indicate that such a scaling law is common in human dynamics and should be taken into account in models of the evolution of networks [19] and of human activity.

We found that in the network under investigation assortative mixing is observed, *i.e.* the tendency for vertices in network to be connected to other vertices that are similar to them in some way. The average connectivity

of nearest neighbours of a node increases with its degree increasing. Moreover the network is assortative mixed by age of a node. We observe also interesting correlation between age and degree of a node.

We believe that such a correlation observed in the system under investigation can have strong influence on dynamic phenomena in social networks. Therefore, they should be taken into account in order to create more plausible models of different dynamic phenomena observed in social networks, like rumour propagation, epidemic spreading or opinion formation [20].

We wish to thank Izabela Grabowska for help in the preparation of this article. A.G. acknowledges financial support from the Foundation for Polish Science (FNP 2007).

REFERENCES

[1] R. Albért, A-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
[2] S.N. Dorogovtsev, J.F.F. Mendes *Evolution of Networks*, Oxford Univ. Press 2004.
[3] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, Y. Aberg, *Nature* **411**, 907 (2001).
[4] H. Ebel, L.I. Mielsch, S. Bornholdt, *Phys. Rev.* **E66**, 035103(R) (2002).
[5] W. Bachnik, S. Szymczak, P. Leszczynski, R. Podsiadlo, E. Rymszewicz, L. Kurylo, D. Makowiec, B. Bykowska, *Acta Phys. Pol. B* **36**, 3179 (2005).
[6] G. Csanyi, B. Szendroi, *Phys. Rev.* **E69**, 036131 (2004).
[7] A. Grabowski, *Physica A* **385**, 363 (2007).
[8] M.E.J. Newman, J. Park, *Phys. Rev.* **E68**, 036122 (2003).
[9] M.C. González, P.G. Lind, H.J. Herrmann, *Phys. Rev. Lett.* **96**, 088702 (2006).
[10] S.H. Strogatz, *Nature (London)* **410**, 268 (2001).
[11] A. Grabowski, N. Kruszewska, *Int. J. Mod. Phys.* **C18**, 1527 (2007).
[12] M.E.J. Newman, *Phys. Rev.* **E67**, 026126 (2003).
[13] E. Ravasz, A-L. Barabási, *Phys. Rev.* **E67**, 026112 (2003).
[14] P. Sen, S.S. Manna, *Phys. Rev.* **E68**, 026104 (2003).
[15] T. Henderson, S. Nhatti, Modelling User Behavior in Networked Games, Proceedings of the 9th ACM International Conference on Multimedia, 2001, p. 212.
[16] A-L. Barabási, *Nature* **435**, 207 (2005).
[17] J.G. Oliveira, A-L. Barabási, *Nature* **437**, 1251 (2005).
[18] A. Vázquez, *Phys. Rev. Lett.* **95**, 248701 (2005).
[19] A. Grabowski, R. Kosiński, *Acta Phys. Pol. B* **38**, 1785 (2007).
[20] A. Grabowski, M. Rosińska, *Acta Phys. Pol. B* **37**, 1521 (2006).