New approaches to model and study social networks

P G Lind\textsuperscript{1,2} and H J Herrmann\textsuperscript{3,4}

\textsuperscript{1} Institute for Computational Physics, Universität Stuttgart, Pfaffenwaldring 27, D-70569 Stuttgart, Germany
\textsuperscript{2} Centro de Física Teórica e Computational, Universidade de Lisboa, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
\textsuperscript{3} Computational Physics, Institute for Building Materials, HIF E12, ETH Hönggerberg, CH-8093 Zürich, Switzerland
\textsuperscript{4} Departamento de Física, Universidade Federal do Ceará, 60451-970 Fortaleza, Brazil
E-mail: lind@icp.uni-stuttgart.de

\textit{New Journal of Physics} \textbf{9} (2007) 228

Received 9 January 2007
Published 12 July 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/7/228

\textbf{Abstract}. We describe and develop three recent novelties in network research which are particularly useful for studying social systems. The first one concerns the discovery of some basic dynamical laws that enable the emergence of the fundamental features observed in social networks, namely the nontrivial clustering properties, the existence of positive degree correlations and the subdivision into communities. To reproduce all these features, we describe a simple model of mobile colliding agents, whose collisions define the connections between the agents which are the nodes in the underlying network, and develop some analytical considerations. The second point addresses the particular feature of clustering and its relationship with global network measures, namely with the distribution of the size of cycles in the network. Since in social bipartite networks it is not possible to measure the clustering from standard procedures, we propose an alternative clustering coefficient that can be used to extract an improved normalized cycle distribution in any network. Finally, the third point addresses dynamical processes occurring on networks, namely when studying the propagation of information in them. In particular, we focus on the particular features of gossip propagation which impose some restrictions in the propagation rules. To this end we introduce a quantity, the spread factor, which measures the average maximal fraction of nearest neighbours which get in contact with the gossip, and find the striking result that there is an optimal non-trivial number of friends for which the spread factor is minimized, decreasing the danger of being gossiped about.

\textit{New Journal of Physics} \textbf{9} (2007) 228 PII: S1367-2630(07)41072-2

1367-2630/07/010228+17$30.00 © IOP Publishing Ltd and Deutsche Physikalische Gesellschaft
1. Introduction

Contrary to what may be perceived at a first glance, social and physical models have been brought together several times during the last four centuries. In fact, not only Maxwell and Boltzmann were inspired by the statistical approaches in social sciences to develop the kinetic theory of gases, but one can even cite the English philosopher Thomas Hobbes, who already in the seventeenth century, using a mechanical approach, tried to explain how people’s acquaintances and behaviours may contribute to the evolution towards a stable absolute monarchy [1, 2]. More than making a historical perspective if these approaches were successful and correct or not, it is almost unquestionable that, at a certain level, there are social phenomena that could be more deeply understood by using approaches of statistical and physical models. Recently [3]–[5], this perspective gained considerable strength from the increased interest in the—in several senses successful—network approach, where one describes complex systems by mapping them on a graph (network) of nodes and links and studies their structure and dynamics with the help of some statistical and topological tools from statistical physics and graph theory [6, 7].

When addressing the specific case of a social system, nodes represent individuals and the connections between them represent social relations and acquaintances of a certain kind. Social networks have been studied in different contexts [8]–[14], ranging from epidemics spreading and sexual contacts to language evolution and election voting. However, although they are ubiquitous, social networks differ from most other networks, yielding a still broad spectrum of unanswered questions and improvements to be done when studying their statistical and topological properties. In this paper, we will address three fundamental open questions related to the typical structure and dynamics associated with social networks.

The first open question has to do with the modelling of social networks. The recent broad study of empirical social networks has shown that they have three fundamental features common to all of them [15]. Firstly, they present the small-world effect [16] with small average path lengths between nodes and high clustering coefficients meaning that neighbours tend to be connected with each other. Secondly, they have positive correlations: the highly (poorly) connected nodes tend to connect to other highly (poorly) connected nodes. Thirdly and last, invariably one observes an organization of the network into some subsets of nodes (communities) more densely connected between each other. Although there are arguments pointing out that all these features could result from one another [15], the modelling of specific social networks reproducing quantitatively all these features has not been successful. Using a recent approach to construct networks, based on a system of mobile agents, it is possible to reproduce all these features. In section 2, we...
will further show that the degree distributions characterizing social networks typically follow a specific one-parameter distribution, a so-called Brody distribution.

The second question is related to the intrinsic nature of the nodes. For certain social networks there are intrinsic features of the individuals which must be considered in the analysis. For instance, the gender in networks of sexual contacts [14] or the hierarchical position in a network of social contacts inside some enterprise. From the network point of view this distinction means to introduce multipartivity in the network, biasing the preferential attachment between nodes that tend to connect with nodes of a certain type. When there are two types of nodes, e.g. men and women, and the connections between them is strongly related to this type, e.g. men can only match women and vice-versa, the standard measures to analyse network structure fails. In particular, the standard clustering coefficient [16], is unable to quantify the connectedness of broader neighbourhoods that typically appear in multipartite networks. In section 3, we will revisit some of the clustering coefficients used to study clustering in bipartite networks, and show how the combination of both clustering coefficients can yield good estimates of normalized cycle distributions. Moreover, we will discuss a general theoretical picture of a global measure of increasing order of clustering coefficients according to some suitable expansion.

The third open question has to do with the heterogeneity of nodes in what concerns their influence in the connections and therefore in the propagation phenomena on social networks. In rumour propagation [17], for instance, one usually treats all connections equally in the spread of some signal (opinion, rumour, etc). This is a suitable assumption for situations like the spread of an opinion which is equally interesting to all nodes in the network, for example political opinions in some election. However, there are also several social situations where the signal is not equally interesting to all nodes, such as the case of spreading of some gossip about some common friend. In these cases there are connections which will be more probably used to spread the signal than others, since not all our friends are also friends of the particular person who is being gossiped about and therefore, either we tend not to tell the gossip to them or they tend not to spread it even if they hear it. In section 4, we will present a simple model for gossip propagation and describe some striking features. Namely, that there is an optimal number of friends, depending on the degree distribution and degree correlations of the entire network, for which the danger of being gossiped about is minimized.

Finally, in section 5, we make final conclusions, giving an overview of future questions which could be studied in social networks arising from the topics studied throughout the paper.

2. Modelling social networks: an approach based on mobile agents

Since the study of social networks is mainly concerned with topological and statistical features of people’s acquaintances [8, 9], the modelling of such networks has been done within the framework of graph theory using suitable probabilistic laws for the distribution of connections between individuals [3]–[5], [7]. This approach proved to be successful in several contexts, for instance to describe community formation [18, 19] and their growth [20].

However, they present two major drawbacks. Firstly, the graph approach may be suited to describe the structure of social contacts and acquaintances, but it does not give insight into the social dynamical laws underlying the structure. Secondly, these models seem to be unable to reproduce all the main features characteristic of social networks, at least at the fundamental level. In this context, it was pointed out that [21]–[23] dynamical processes based on local information...
should be also considered when modelling the network. Our recent proposal to overcome these shortcomings was to construct networks, from a system of mobile agents following a simple motion law [13, 14]. Mobile agents represent the nodes (persons) of the social network and the connections between pairs of nodes are ‘generated’ by the collisions occurring among the agents. With such an approach, one is able to construct a network, whose topology naturally emerges just from the collision law chosen for the agents. In other words, no global information is needed to address connections between pairs of nodes. In this section, we briefly review this model and further present the analytical expression that fits the obtained degree distributions. In particular, we show that the degree distribution typically follows a Brody distribution [24].

The model is given by a system of particles (agents) that move and collide with each other, forming through those collisions the acquaintances between individuals. Consequently, the network results directly from the time evolution of the system and is parameterized by two single parameters, the density \( \rho \) of agents characterizing the system composition and the maximal residence time \( T_\ell \) controlling its evolution. Each agent \( i \) is characterized by its number \( k_i \) of links and by its age \( A_i \). When initialized, each agent has a randomly chosen age, position and moving direction with velocity \( v_0 \) and one sets \( k_i = 0 \). While moving, the individuals follow ballistic trajectories till they collide. As a first approximation, we assume that social contacts do not determine which social contact will occur next. Therefore, after collisions, the total momentum should not be conserved, with the two agents choosing completely random new moving directions. Figure 1 sketches consecutive stages of the evolution of such a system of mobile agents.

**Figure 1.** Illustration of the two-dimensional mobile agents system. Initially there are no connections between nodes and nodes move with some initial velocity \( v_0 \) in a randomly chosen direction (arrows). At \( t = 1 \) two nodes, \( P_1 \) and \( P_2 \) collide and a connection between them is introduced (solid line), velocities are updated increasing their magnitude and choosing a new random direction. At \( t = 2 \) two other collisions occur, between nodes \( P_2 \) and \( P_4 \) and between nodes \( P_1 \) and \( P_3 \). In this way a network of nodes and connections between them emerges as a straightforward consequence of their motion (see text).
Assuming that large numbers of acquaintances tend to favour the occurrence of new contacts, the velocity should increase with degree $k$, namely

$$
\vec{v}(k_i) = (\bar{v} k_i^\alpha + v_0) \vec{\omega},
$$

where $\bar{v} = 1 \text{ m s}^{-1}$ is a constant to assure dimensions of velocity, $\vec{\omega} = (\vec{e}_x \cos \theta + \vec{e}_y \sin \theta)$ with $\theta$ a random angle and $\vec{e}_x$ and $\vec{e}_y$ are unit vectors. The exponent $\alpha$ in equation (1) controls the velocity update after each collision. Here, we consider $\alpha = 1$. Further, the removal of agents considered here is simply imposed by some threshold $T_\ell$ in the age of the agents: when $A_i = T_\ell$ agent $i$ leaves the system and a new agent $j$ replaces it with $k_j = 0$, $v_j = v_0$ and a randomly chosen moving direction. The selected values for $T_\ell$ must be of the order of several times the characteristic time $\tau$ between collisions, in order to avoid premature death of the nodes. Too large values of $T_\ell$ are also inappropriate since in that case each node may on average collide with all other nodes yielding a fully connected network.

Similarly to other systems [25, 26], this finite $T_\ell$ enables the entire system to reach a non-trivial quasi-stationary state [13]. In fact, only by tuning $T_\ell$ within an acceptable range of small density values can one reproduce networks of social contacts. In figure 2(a) one sees the normalized residence time $T_\ell/\tau$ as a strictly monotonic function of the average degree $\langle k \rangle$. From the residence time it is also possible to define a collision rate, as the fraction between the average residence time $T_\ell - \langle A(0) \rangle = T_\ell/2$ and the characteristic time $\tau$, namely $\lambda = T_\ell/(2\tau) = (\bar{v} T_\ell/(2v_0 \tau_0))$, where $\tau_0$ is the characteristic time of the system at the beginning when all agents have velocity $v_0$. Figure 2(b) shows clearly that $\lambda = 2\langle k \rangle$.

By looking at figure 2 one now understands the main strength of the mobile agent model here described: when taking a real network of social contacts and measuring the average degree $\langle k \rangle$ the correspondence sketched in figure 2 straightforwardly returns the suitable value of $T_\ell$ that reproduces the topological and statistical features.

It has already been reported [13, 27] that empirical networks extracted from a survey among 84 American schools are easily reproduced with this mobile agent model, as far as the degree distribution, second-order correlations, community structure, average path length and clustering coefficient are concerned. As an illustration, figure 3 shows the degree distribution $P(k)$ of nine

![Figure 2](http://www.njp.org/)
Figure 3. Degree distributions of nine different schools (symbols) from an in-school questionnaire involving a total of 90 118 students which responded in a survey between 1994 and 1995. Each school comprises a number $N$ of interviewed students and from their questionnaires an average number $\langle k \rangle$ of acquaintances is extracted. With solid lines, we represent the fit obtained with a Brody distribution, equation (2), whose parameter value is computed in figure 4.

Such distributions are well fitted by Brody distributions (solid lines) defined as [24]:

$$P_B(\bar{k}) = \frac{1}{B} (\beta + 1) \eta \bar{k}^\beta \exp\left(-\eta \bar{k}^{\beta+1}\right),$$

(2)

with $\bar{k} = k/\langle k \rangle$ and

$$\eta = \Gamma\left(\frac{\beta + 2}{\beta + 1}\right)^{\beta+1}$$

(3)

and $B$ a normalization constant. Roughly, the Brody distribution in equation (2) is, apart from some special constants, the product of a power of $k$ with an exponential with a negative exponent proportional to a higher power of $k$. For the particular case $\beta = 0$, the Brody distribution reduces to the exponential distribution having always a non-positive derivative.

The distributions in figure 3 were obtained with values of $\beta$ slightly above zero, namely between zero and one as shown in figure 4. In this case one is able to obtain the nontrivial positive slope which is typically observed for small $k$ values in the degree distribution of such social networks. Interestingly, figure 4 also shows a linear trend between the average degree $\langle k \rangle$ in
Figure 4. The linear dependence between the parameter $\beta$ of the Brody distribution in equation (2) with the average number $\langle k \rangle$ of connections. Each bullet corresponds to one of the schools whose degree distribution is plotted in figure 3. The solid line yields the fit $\beta = 0.094 \langle k \rangle + 0.078$.

the network and the corresponding value of $\beta$ which fits the degree distribution. This guarantees that distribution in equation (2) has indeed one single parameter.

To understand how this model of colliding agents moving in a two-dimensional space is some form of mapping of the dynamics of social systems, three last remarks remain to be clarified. Firstly, we point out that the two-dimensional space where nodes move is not the physical space, but represents a projection of a highly dimensional Euclidean space whose metric is related to what is called the social distance [28, 29]. This social distance can be regarded as a measure of both the geographic distance between nodes together with their affinities. In this context, one easily understands that the smaller the social distance the more probable it is to establish an acquaintance, i.e. to collide.

Secondly, the residence time $T_\ell$ and the density $\rho$, as the two parameters of our model are closely related with the parameters of the social network, namely the average degree $\langle k \rangle$ and the clustering coefficient $C$. Increasing $T_\ell$ increases the average number $\langle k \rangle$ of connections, while increasing the density confines the accessible region of agents thus promoting the occurrence of collisions among them which are more confined in space, and therefore increases the clustering coefficient.

Thirdly, from both these first two remarks one then understands that the velocity of a given agent may be regarded as a measure of the region accessible to it. Therefore, increasing at each collision, the velocity illustrates the increasing ability of a person to establish new acquaintances. A deeper discussion of these points is given in [27].

3. Particular measures for social networks

To measure ‘the cliquishness of a typical neighbourhood’ in a network, Watts and Strogatz [16] introduced a simple coefficient, called the clustering coefficient, which counts the number of pairs
of neighbours of a certain node which are connected with each other, forming a cycle of size \( s = 3 \). While such a tool enables one to access the structure of complex networks arising in many systems [4, 7], helping to characterize small-world networks [16], to understand synchronization in scale-free networks of oscillators [30] and to characterize chemical reactions [31] and networks of social relationships [32, 33], there are other situations where this measure does not suit. Namely, when the network presents a multipartite structure. For instance, when there are two different kinds of nodes and connections link only nodes of different type, the network is bipartite [32]–[34] and the bipartite structure does not allow the occurrence of cycles with odd size, in particular with \( s = 3 \).

Bipartite networks are quite common for social systems [34, 35] where the two different kinds of nodes represent e.g. the two genders. While the standard clustering coefficient in such networks is always zero, they have in general non-vanishing clustering properties [33]. When the nodes are of completely different nature (e.g. actors and movie networks or collaboration networks composed by papers linked to the corresponding authors) one simple way to overcome this shortcoming is to project [9] the bipartite network into a monopartite network, composed of only one kind of nodes, whose connections link those nodes connected to a common node of the other kind. However, there are also several situations where bipartite and monopartite counterparts may co-exist, and comparisons between them should not take any projection. It is for instance the case for networks of sexual contacts [14] where some of them have two genders (bipartite) and some of them consider only homosexual contacts. For these situations, more appropriate quantities to access such networks have been proposed, namely coefficients counting larger cycles [14, 33]. In this section, we will discuss how these different clustering coefficients are related to each other and how one can use them to improve the knowledge of the network structure.

The standard clustering coefficient \( C_3 \) is usually defined [16] as the fraction between the number of cycles of size \( s = 3 \) (triangles) observed in the network out of the total number of possible triangles which may appear, namely

\[
C_3(i) = \frac{2t_i}{k_i(k_i - 1)},
\]

where \( t_i \) is the number of existing triangles containing node \( i \) and \( k_i \) is the number of neighbours of node \( i \), yielding a maximal number \( k_i(k_i - 1)/2 \) of triangles.

To access the cliquishness in bipartite networks a clustering coefficient \( C_4(i) \) has been proposed [21, 33, 34, 36], sometimes called the grid coefficient [36], defined as the quotient between the number of cycles of size \( s = 4 \) (squares) and the total number of possible squares. Explicitly, for a given node \( i \) with two neighbours, say \( m \) and \( n \), this coefficient yields [21]

\[
C_{4,mn}(i) = \frac{q_{mn}}{(k_m - \eta_{i,mn})(k_n - \eta_{i,mn}) + q_{i,mn}},
\]

where \( q_{mn} \) is the number of common neighbours between \( m \) and \( n \) (not counting \( i \)) and \( \eta_{i,mn} = 1 + q_{i,mn} + \theta_{mn} \) with \( \theta_{mn} = 1 \) if neighbours \( m \) and \( n \) are connected with each other and 0 otherwise.

After averaging over the nodes, the coefficients \( C_3 \) and \( C_4 \) characterize the contribution of the first and second neighbours, respectively, for the network cliquishness. In order to be a suitable quantity to measure the cliquishness of bipartite networks compared to their monopartite
counterparts, \( C_4 \) must behave the same way as \( C_3 \) when the network parameters are changed, as is indeed the case for \( \langle C_4 \rangle \) computed from equation (5). See [21] for details.

One should notice that in most \( m \)-partite networks, it is always possible to have cycles of size \( s = 4 \), indicating that \( C_4 \) is in some sense a more general clustering measure than \( C_3 \). However, it could be the case that for a larger number of partitions forming the network, the contribution of larger cycles increases. This is the case, for instance, of trophic relations in an ecological network of different individuals from different species, where large cycles tend to be abundant, namely the ones ranging from the higher predators to the plants at the lowest trophic level. In such cases, a general clustering coefficient counting the fraction of possible cycles of arbitrary size \( n \) may be needed. The generalization is straightforward yielding a clustering coefficient \( C_n = E_n/L_n \), where \( E_n \) is the number of existing cycles with size \( n \), \( L_n \) the maximal number of such cycles that it is possible to attain and \( n = 3, \ldots, N \) for a network of \( N \) nodes.

Having \( C_n \) for the required values of \( n \), one is able to introduce a general clustering measure of the network, given by the sum of all these contributions, namely

\[
C = \sum_{n=3}^{N} \alpha_n C_n = \sum_{n=3}^{N} \alpha_n \frac{E_n}{L_n},
\]

where \( \alpha_n \) is a coefficient that weights the contribution of each different clustering order \( n \) and obeys the normalization condition \( \sum_{n=3}^{N} \alpha_n = 1 \). A general expression for \( E_n \) and \( L_n \) in equation (6) may be easily derived using the correlation degree distribution \( q(k_1, k_2) \), which gives the fraction of connections linking a node with \( k_1 \) neighbours to a node with \( k_2 \) neighbours. In fact, since one expects to have \( NP(k_1) \) nodes with \( k_1 \) neighbours, the value \( NP(k_1)q(k_1, k_2) \) gives an estimate of the number of those nodes that are connected to nodes with \( k_2 \) neighbours. Summing over \( k_2 \) gives then the total number of connections connecting a node with \( k_1 \) neighbours. Assuming that these connections are part of at least one cycle composed by \( n \) edges, one considers a similar estimate for each one of those edges, yielding

\[
E_n = \sum_{k_1, \ldots, k_n} NP(k_1)q(k_1, k_2)NP(k_2)q(k_2, k_3) \ldots NP(k_n)q(k_n, k_1).
\]

where \( P(k) \) is the fraction of nodes with \( k \) neighbours. As for the total number of possible paths in the network one may simply estimate it as the total number of permutations of sets of \( n \) nodes out from the total network, yielding

\[
L_n = \frac{N!}{(N-n)!} = B^n_n n!,
\]

where \( B^n_n \) is the total number of combinations of \( n \) elements out of \( N \).

From equation (7) one can assume approximately that \( E_n \sim \langle P \rangle \langle q \rangle N^n \) with \( \langle P \rangle \) and \( \langle q \rangle \) the average fractions of \( P(k) \) and \( q(k_1, k_2) \) respectively. Since \( L_n \) increases also as \( N^n \), a possible suitable choice for \( \alpha \) would be a constant, namely \( \alpha = 1/(N-2) \) obeying the normalization condition above. Having presented this general scenario, we now concentrate on the two first clustering coefficients, \( C_3 \) and \( C_4 \), to address the cycle size distribution.

We first show an estimate introduced in [37], which considers only the degree distribution \( P(k) \) and the distribution of the standard clustering coefficient \( C_3(k) \). One starts by considering the set of cycles with a central node, i.e. cycles with one node connected to all other nodes.
Figure 5. Illustrative examples of cycles (size $s = 6$) where the most connected node ($\circ$) is connected to (a) all the other nodes composing the cycle, forming four adjacent triangles. In (b) the most connected node is connected to all other nodes except one, forming two triangles and one sub-cycle of size $s = 4$, while in (c) the same cycle $s = 6$ encloses two sub-cycles of size $s = 4$ and no triangles (see text).

composing the cycle, as illustrated in figure 5 (a). The central node composes one triangle with each pair of connected neighbours. Due to this fact, the number of cycles with size $s$ can be easily estimated, since the number of different possible cycles that can occur is $n_0(s, k) = B_{s-1}^k \frac{(s-1)!}{2}$, for a central node with $k$ neighbours and the corresponding fraction of these cycles which is expected to occur is $p_0(s, k) = C_3(k)^{s-2}$, yielding a total number of $s$-cycles given by

$$N_s = N_{gs} \sum_{k=1}^{k_{\text{max}}} P(k)n_0(s, k)\ p_0(s, k),$$  \hspace{1cm} (9)$$

where $g_s$ is a factor which takes into account the number of cycles counted more than once.

The estimate in equation (9) is a lower bound for the total number of cycles since it considers only cycles with a central node. Further, this estimate only accounts for cycles up to size $s \leq k_{\text{max}} + 1$, with $k_{\text{max}}$ the maximal degree and is not suited for bipartite networks where $C_3(k) = 0$ for all $k$. Bipartite networks are typically composed of a set of nodes like that illustrated in figure 5(c), where no central node exists.

By using additionally the coefficient $C_4(k)$ in a similar estimate, one is now able to take into account several cycles without central nodes. One first considers the set of cycles of size $s$ with one node connected to all the others except one, as illustrated in figure 5(b). Assuming that this node has $k$ neighbours, $s - 2$ of them belonging to the cycle one is counting for, one has $n_1(s, k) = B_{s-2}^k(s-2)!/2$ different possible cycles of size $s$. The corresponding fraction of such cycles which is expected to occur is given by $p_1(s, k) = C_3(k)^{s-4}(c)_4(k)(1 - C_3(k))$. Writing an equation similar to equation (9), where instead of $n_0(s, k)$ and $p_0(s, k)$ one has $n_1(s, k)$ and $p_1(s, k)$ respectively and the sum starts at $s - 2$ instead of $s - 1$, one has an additional number $N_s'$ of estimated cycles which is not considered in estimate (9).

To improve the estimate further one repeats the same approach, taking out each time one connection to the initial central node, increasing by one the number of elementary cycles of size $s = 4$. Figure 5(c) illustrates a cycle of size $s = 6$ composed of two elementary cycles of size 4. In general, for cycles composed by $q$ sub-cycles of size 4 one finds $n_q(s, k) = \frac{(s-q-1)!}{2}B_{s-q-1}^k$ possible cycles of size $s$ looking from a node with $k$ neighbours and a fraction $p_q(s, k) = C_3(k)^{s-2q-2}(c)_4(k)^q(1 - C_3(k))^q$ of them which are expected to be observed.

New Journal of Physics 9 (2007) 228 (http://www.njp.org/)
Figure 6. (a) The fraction \( N_s/N_{gs} \) of the number of cycles estimated from equations (9), dashed lines, and (10), solid lines, compared with (b) the exact number of cycles as a function of the size for the pseudo-fractal network [38]. From small to large curves one has pseudo-fractal networks with \( m = 2, 3, 4, 5 \) generations (see text). In (c) one sees the comparison between both estimates in a scale-free network with degree distribution \( P(k) = P_0 k^{-\gamma} \) with \( P_0 = 0.737 \) and \( \gamma = 2.5 \), and coefficient distributions \( C_{3,4}(k) = C_{3,4}^{(0)} k^{-\alpha} \) with \( C_{3}^{(0)} = 2, C_{4}^{(0)} = 0.33 \) and \( \alpha = 0.9 \).

Summing up over \( k \) and \( q \) yields our final expression

\[
N_s = N_{gs} \left[ \frac{s}{2} - 1 \right] \sum_{q=0}^{\left\lfloor \frac{s}{2} - 1 \right\rfloor} \sum_{k = s - q - 1}^{k_{\text{max}}} P(k) n_q(s, k) p_q(s, k). \tag{10}
\]

where \( \lfloor x \rfloor \) denotes the integer part of \( x \). In particular, the first term \( (q = 0) \) is the sum in equation (9) and the upper limit \( \left\lfloor \frac{s}{2} - 1 \right\rfloor \) of the first sum is obtained by forcing the exponent of \( C_3(k) \) in \( p_q(s, k) \) to be non-negative.

The estimate in equation (10) not only improves the estimated number computed from equation (9), but also enables the estimate of cycles up to a larger maximal size [21], namely up to \( s = 2k_{\text{max}} \) where \( k_{\text{max}} \) is the maximal number of neighbours in the network.

The estimate in equation (10) has also the advantage of being able to estimate cycles in bipartite networks. Since for bipartite networks \( C_3(k) = 0 \), all terms in equation (10) vanish except those for which the exponent of \( C_3(k) \) is zero, i.e. for \( s = 2(q + 1) \) with \( q \) an integer, which naturally shows the absence of cycles of odd size in such networks.

For highly connected networks, both estimates should nevertheless yield similar results, since in that case there is a very large number of both triangles and squares. For instance, the so-called pseudo-fractal network [38] is a deterministic scale-free network, constructed from three initial nodes connected with each other (generation \( m = 0 \)), and iteratively adding new generations of nodes such that in generation \( m + 1 \) one new node is added to each edge of generation \( m \) and is connected to the two nodes joined by that edge. For these networks, the exact number of cycles with size \( s \) can be written iteratively [39] and can be directly compared to the one obtained with the two estimates above. Figure 6(a) shows the two estimates, while in figure 6(b) the exact number is computed. We notice that both the real number \( N_s \) of cycles
and the normalized value \( N_s/(N g_s) \), though different, yield the same shape. Thus, although the estimates above are not able to explicit the geometrical factor \( g_s \), the corresponding normalized distributions agree very well with the real one. However, while in this simple situation both estimates are similar, in general they can deviate significantly, as illustrated in figure 6(c). In such cases, the estimate (10) is closer to the real distribution of cycle sizes [21].

4. Spreading phenomena in social networks

In the previous section, we show how the study of network structure can be addressed by using tools such as the clustering coefficient and first and second degree distributions. However, although the ability to communicate within a network of contacts is favoured by the network topology [40], to study dynamical phenomena occurring on the network other measures are necessary. Here, we focus on novel properties that help to ascertain the broadness and speed of propagating phenomena through the network. We will describe two helpful quantities to study propagation in a network. As we will see these tools are particularly suited for a simple model of gossip propagation, that yields a striking result: in real social systems it is possible to minimize the risk of being gossiped about, by choosing an optimal number of friendship acquaintances.

We start by introducing the additional quantities in the context of gossip propagation. As opposed to rumours, gossip always targets details about the behaviour or private life of a specific person. Some information of a specific piece of gossip is created at time \( t = 0 \) about the victim by one of his neighbours. Since typically the gossip tends to be of interest to only those who know the victim personally, we consider first that it only spreads at each time step from the vertices that know the gossip to all vertices that are connected to the victim and do not yet know the gossip. Our dynamics is therefore like a burning algorithm [41], starting at the originator and limited to sites that are neighbours of the victim. The gossip will spread until all reachable neighbours of the victim know it, yielding a spreading time \( \tau \).

To measure how effectively the gossip spreads or more generally the amount of information attained by the neighbours of the starting node (victim), we define the spreading factor \( f \) given by

\[
    f = n_f/k,
\]

where \( n_f \) is the total number of the \( k \) neighbours who eventually hear the gossip in a network with \( N \) vertices (individuals). Although similar in particular cases, the spreading factor \( f \) and the clustering coefficient are, in general, different because the latter one only measures the number of bonds between neighbours giving no insight about how they are connected.

In figure 7 one sees how the spreading time \( \tau \) depends on the degree \( k \) of the starting node. The Apollonian (APL) network [42] is illustrated in figure 7(a), while the case of Barabási–Albert (BA) networks is given in figure 7(b). In both cases \( \tau \) clearly grows logarithmically,

\[
    \tau = A + B \log k,
\]

for large \( k \). In the case of the Apollonian network, one can even derive this behaviour analytically as follows. In order to communicate between two vertices of the \( n \)-th generation, one needs up to \( n \) steps, which leads to \( \tau \propto n \). Since for the Apollonian network one has [42] \( k = 3 \times 2^{n-1} \), one immediately obtains that \( \tau \propto \log k \).

New Journal of Physics 9 (2007) 228 (http://www.njp.org/)
Figure 7. Semi-logarithmic plot of the spreading time $\tau$ as a function of the degree $k$ for (a) the Apollonian ($n = 9$ generations) and (b) the BA network with $N = 10^4$ nodes for $m = 3$ (circles), 5 (squares) and 7 (triangles), where $m$ is the number of edges of a new site, and averaged over 100 realizations. In the inset of (a) we show a schematic design of the Apollonian lattice for $n = 3$ generations. Fitting equation (12) to these data, we have $B = 1.1$ in (a) and $B = 5.6$ for large $k$ in (b).

For the Apollonian network all neighbours of a given victim are connected in a closed path surrounding the victim, as can be seen from the inset of figure 7(a), yielding $f = 1$. This stresses the fact that the spread factor $f$ is rather different from the clustering coefficient which in this case is $C = 0.828$ [42].

Next, we will show that for these two features to appear one needs the existence of degree correlations between connected nodes, as usually observed in real empirical networks. In figure 8 we plot the results of gossip spreading on an empirical set of networks extracted from survey data$^5$ in 84 US schools. Here, the logarithmic growth of $\tau$ with $k$, shown in figure 8(a), follows the same dependence of the average degree $k_{nn}$ of the nearest neighbours [43], as illustrated in the inset. As in the case of the BA networks, we also find for the schools a characteristic degree $k_0$ for which $f$ and therefore the gossip spreading is smallest. The inset of figure 8(b), however, gives clear evidence that the school networks are not scale-free. Since the same optimal degree appears in BA networks, one argues that the existence of this optimal number is not necessarily related to the degree distribution of the network, but rather to the degree correlations.

However, the relation between degree correlations, measured by $k_{nn}$, and the logarithmic behaviour of the spreading time is not straightforward. While in the empirical network we found the same distribution for both $k_{nn}$ and $\tau$, in BA and APL networks $k_{nn}$ follows a power-law with $k$ (not shown). As for the spread factor $f$, a mean field approach can be derived, yielding an $f$-rate equation which depends in general on $P(k)$ and two and three-point correlations of the degree. In the case of uncorrelated networks, two- and three-point correlations reduce to

$^5$ Add Health program designed by J R Udry, P S Bearman and K M Harris funded by National Institute of Child and Human Development (PO1-HD31921).
Figure 8. Gossip propagation on a real friendship network of American students (Add Health programme) averaged over 84 schools. In (a) we show the spreading time $\tau$ and, in the inset, the average degree of neighbours of nodes with degree $k$. In (b) the spread factor $f$ is given as a function of degree $k$. In the inset of (b) we see the degree distribution $P(k)$.

Figure 9. Distribution $P(\tau)$ of spreading times $\tau$ for (a) the APL network of eight generations, and (b) the real school network (circles) and the BA network with $m = 9$ and $N = 1000$ (solid line).

simple expressions of the moments of the degree distribution. Therefore, $f$ is independent of the degree, similarly to what is observed for the density of particles as derived by Catanzaro et al [44] in diffusion–annihilation processes on complex networks. For correlated networks, such as the empirical network here studied, the analytical approach is not straightforward and will be presented elsewhere.

Another quantity of interest is the distribution $P(\tau)$ of spreading times, which clearly decays exponentially for the Apollonian network, as illustrated in figure 9(a). This behaviour can also be obtained analytically by considering that $P(\tau)d\tau = P(k)dk$ and using equation (12) together...
with the degree distribution, \( P(k) \propto k^{-\gamma} \), to obtain

\[
P(\tau) \propto e^{\tau(1-\gamma)/B},
\]

for large \( k \). The slope in figure 9(a) is precisely \( (1 - \gamma)/B = -0.17 \) using \( B \) from figure 7(a) and \( \gamma = 2.58 \) from [42]. For the school network, \( P(\tau) \) follows also an exponential decay for large \( \tau \), but with a 3.5 times smaller characteristic decay time, and has a maximum for small \( \tau \), as seen in figure 9(b) (circles). Compared to the \( P(\tau) \) of the BA network with \( m = 9 \) (solid line), the shapes are similar but the BA case is slightly shifted to the right, due to the larger minimal number of connections.

Many other regimes of gossip and of propagation phenomena can also be addressed with these two quantities. Namely, a more realistic scenario could be addressed by enabling each node to transfer information with a probability \( 0 \leq p \leq 1 \). Further, the assumption that the person to which a gossip did not spread at the first attempt will never get it, yields a regime similar to percolation conditional to the neighbourhood of the victim. Differently, if at each time-step the neighbours which already know the gossip repeatedly try to spread it to the common friends, one observes the same value of \( f \) measured for \( q = 1 \), and the spreading time scales as \( \tau' \sim \tau/q \), where \( \tau \) is measured for \( q = 1 \). Finally, other possible regimes comprehend the situation where the gossip spreads over strangers, i.e. over nodes which are not directly connected to the victim. Such cases are being studied in detail and results will be presented elsewhere [45].

5. Discussion and conclusions

In this paper, we presented and developed recent achievements in social network research, concerning the modelling of empirical networks, and specific mathematical tools to address their structure and dynamical processes on them.

Concerning the modelling of empirical networks, we described briefly a recent approach based on a system of mobile agents. Further developments were given, namely in what concerns the analytical expression which fits the typical degree distributions observed in empirical social networks. We gave evidence that such distributions follow a Brody distribution which depends on a single parameter that scales with the average degree of the network. A question which now remains to be answered is how to derive such a distribution from an analytical and meaningful approach.

Showing that the usual clustering coefficient is, in general, inappropriate when addressing the clustering properties of social networks, we described a suitable measure to access these properties and presented its additional applications for estimating the distribution of cycles of higher order. This additional clustering coefficient was also put in a general framework with other different higher-order coefficients, that could be useful for particular situations of multipartite networks. An expansion combining all possible coefficients was also proposed, motivated by previous works [4], which depends only on the degree distribution and degree–degree correlations. However, computational effort to compute such coefficients increases exponentially with their order and therefore it is not yet clear how useful such an expansion may be.

Finally, to study dynamical processes in social networks, in particular the propagation of information, two simple measures were introduced. Namely, a spread factor, which measures
the maximal relative size of the neighbourhood reached, when the information starts from a local source (node), and a spreading time, which gives the number of sufficient steps to reach such maximal size. These two measures gave rise to introduction of a minimal model for gossip propagation, which can be seen as a particular model of opinions. Within this specific model, the spread factor was found to be minimized by a particular nontrivial degree of the source, which is related to the degree–degree correlations arising in the network. Whether such a possibility of minimizing the danger of being gossiped can be tested in a real situation and which other implications these findings have in other situations—e.g. in internet virus propagation—remain open questions for forthcoming studies.

Acknowledgments

We thank M C González, J S Andrade Jr, L da Silva and O Durán for useful discussions. We thank the Deutsche Forschungsgemeinschaft and the Max Planck Prize.

References

[1] Ball P 2002 Physica A 314 1–14
[2] Ball P 2003 Complexus 1 190–206
[3] Albert R and Barabási A-L 2002 Rev. Mod. Phys. 74 47–97
[4] Newman M E J 2003 SIAM Rev. 45 167–256
[5] Dorogovtsev S N and Mendes J F F 2002 Adv. Phys. 51 1079–187
[6] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 Phys. Rep. 424 175–308
[7] Bollobás B 1998 Modern Graph Theory (New York: Springer)
[8] Newman M E J 2001 Proc. Natl Acad. Sci. 98 404
[9] Newman M E J, Watts D J and Strogatz S H 2002 Proc. Natl Acad. Sci. USA 99 2566–72
[10] Krapivsky P L and Redner S 2003 Phys. Rev. Lett. 90 238701
[11] Rogers A 2003 Phys. Rev. Lett. 90 158103
[12] Freeman L C 2004 The Development of Social Network Analysis (Vancouver: Empirical)
[13] González M C, Lind P G and Herrmann H J 2006 Phys. Rev. Lett. 96 088702
[14] González M C, Lind P G and Herrmann H J 2006 Eur. Phys. J. B 49 371–6
[15] Newman M E J and Park J 2003 Phys. Rev. E 68 036122
[16] Watts D J and Strogatz S H 1998 Nature 393 440–2
[17] Daley D J and Kendall D G 1964 Nature 204 1118
[18] Grönlund A and Holme P 2004 Phys. Rev. E 70 036108
[19] Toivonen R, Onnela J-P, Saramäki J, Hyvönen J and Kaski K 2006 Preprint physics/0601114
[20] Jin E M, Girvan M and Newman M E J 2001 Phys. Rev. E 64 046132
[21] Lind P G, González M C and Herrmann H J 2005 Phys. Rev. E 72 056127 cond-mat/0504241
[22] Davidsen J, Ebel H and Bornholdt S 2002 Phys. Rev. Lett. 88 128701
[23] Eisenberg E and Levanon E Y 2003 Phys. Rev. Lett. 91 138701
[24] Brody T A 1973 Lett. Nuovo Cimento 7 482
[25] Amaral L A N, Scala A, Barthélémy M and Stanley H E 2000 Proc. Natl Acad. Sci. USA 21 11149
[26] Dorogovtsev S N and Mendes J F F 2000 Phys. Rev. E 62 1842–5
[27] González M C, Lind P G and Herrmann H J 2006 Physica D 224 137
[28] Boguñá M, Pastor-Satorras R, Díaz-Guilera A and Arenas A 2004 Phys. Rev. E 70 056122
[29] Watts D J, Dodds P S and Newman M E J 2002 Science 296 1302–5
[30] McGraw P N and Menzinger M 2005 Preprint cond-mat/0501663
[31] Stadler P F, Wagner A and Fell D A 2001 Adv. Complex Syst. 4 207–26

New Journal of Physics 9 (2007) 228 (http://www.njp.org/)
[32] Newman M E J 2003 Social Netw. 25 83–95
[33] Holme P, Edling C R and Liljeros F 2004 Social Netw. 26 155
[34] Holme P, Liljeros F, Edling C R and Kim B J 2003 Phys. Rev. E 68 056107
[35] Guimerà R, Guardiola X, Arenas A, Díaz-Guilera A, Streib D and Amaral L A N Quantifying the creation of social capital in a digital community, private communication
[36] Caldarelli G, Pastor-Satorras R and Vespignani A 2004 Eur. Phys. J. B 38 183–6
[37] Vázquez A, Oliveira J G and Barabási A-L 2005 Phys. Rev. E 71 025103(R)
[38] Dorogovtsev S N, Goltsev A V and Mendes J F F 2002 Phys. Rev. E 65 066122
[39] Klemm K and Stadler P F 2006 Phys. Rev. E 73 025101(R) (Preprint cond-mat/0506493)
[40] Trusina A, Rosvall M and Sneppen K 2005 Phys. Rev. Lett. 94 238701
[41] Herrmann H J, Hong D C and Stanley H E 1984 J. Phys. A: Math. Gen. 17 L261
[42] Andrade J S Jr, Herrmann H J, Andrade R F S, da Silva L R 2005 Phys. Rev. Lett. 94 018702
[43] Cantazaro M, Boguña M and Pastor-Satorras R 2005 Phys. Rev. E 71 027103
[44] Cantazaro M, Boguña M and Pastor-Satorras R 2005 Phys. Rev. E 71 056104
[45] Lind P G, Andrade J S Jr, da Silva L R, Herrmann H J 2007 Europhys. Lett. 78 68005