

# A comparative study of stochastic algorithmic models for social networks

Riitta Toivonen <sup>a,\*</sup>, Lauri Kovanen <sup>a</sup>, Mikko Kivelä <sup>a</sup>,  
Jukka-Pekka Onnela <sup>b,c,a</sup>, Jari Saramäki <sup>a</sup>, and Kimmo Kaski <sup>a</sup>

<sup>a</sup>*Department of Biomedical Engineering and Computational Science (BECS),  
Helsinki University of Technology, P.O. Box 9203, FIN-02015 HUT, Finland*

<sup>b</sup>*Physics Department, Clarendon Laboratory, Oxford University, Oxford OX1 3PU,  
United Kingdom*

<sup>c</sup>*Saïd Business School, Oxford University, Oxford OX1 1HP, United Kingdom*

---

## Abstract

This paper compares a selection of microscopic *stochastic algorithmic* (SA) models for social networks, and the basic mechanisms employed in them. The knowledge obtained from the comparison can be helpful in designing models with the desired properties. We study closely two (mutually non-exclusive) model types: those in which individuals are linked more likely if they lie close in a social or geographical space (*spatial models*, SM), and those where the addition of new links is dependent on the local network structure (*topological models*, TM). We fit representative models from each of these categories to two real world acquaintance networks with respect to basic network statistics. We then compare higher order structures in the resulting networks with those in the data, with the aim of determining which models produce the most realistic network structure. High clustering arises by design in topological models based on triadic and focal closure (TFC). We find that many of them also produce reasonably realistic degree distributions and clustering spectra, but not very high assortativity nor very clearly clustered structure. On the other hand, the spatial models successfully produce highly clustered and assortative networks and a structure of loosely connected, relatively dense clusters, but not very realistic degree distributions nor clustering spectra. None of the models are successful in all important respects, and a selection has to be made based on which qualities are judged most important. A comparison of the mechanisms employed in dynamical TFC models shows that details in the mechanisms produce important differences in the resulting network structure.

*Key words:* Social networks, Complex networks, Spatial models, Topological models

*PACS:* 64.60.aq, 89.65.Ef, 89.65.-s, 89.75.-k, 02.70.-c

---

# 1 Introduction

Modeling social networks serves at least two purposes. On the one hand, it helps us understand how social networks form and evolve. On the other, successful social network models can be used in studying social processes by simulation to specify the structure of interaction upon which the processes unfold. In this paper, we focus on the various models and the mechanisms within them responsible for generating social network structure, while keeping in mind the application to social dynamics. Comparing the outcome of the various models, we can make inferences about which microscopic mechanisms produce particular kinds of network structure observed in real world social networks. By attempting to match the models to real world data sets, we can also assess the adaptability of the various models.

A large variety of models have been proposed to explore how the microscopic mechanisms of network formation produce network structure. We will look at a category of network models in which links are formed according to rules imitating link formation processes in real social networks. As the rules typically include randomness, we label the models *stochastic algorithmic* (SA) models. Stochastic algorithmic models are based on a variety of assumptions about how social ties are formed. Given the large heterogeneity of SA models, no categorization could neatly label all models without ambiguity. However, we can identify a useful distinction between 1) models based on metric distance (*spatial models*), in which nodes are more likely to be linked if they lie close in a social or geographical space, and 2) models where the addition of new links is dependent on the (usually local) network structure (*topological models*, TM) (Fig. 1). An essential difference is that in the spatial models studied here, each node has unchanging intrinsic characteristics, based on which a distance can be calculated between two nodes.

Within the category of topological models, we can also make a distinction between two types of models. A) In *dynamical models*, the steps for adding and removing ties on a fixed set of nodes are repeated until the network no longer statistically changes, and the resulting network is then extracted. B) In *growing models*, links and nodes are simply added until the network has the desired number  $N$  nodes. A growing model may be appropriate for describing for example the development of a co-authorship network, where new links form but old ones remain, or an online social networking system where people rarely remove links, and new users keep joining the network. We point out that the growing models do not intend to simulate the evolution of a social network *ab initio*. However, the mechanisms are selected to imitate the way people might

---

\* Corresponding author.

*Email address:* Riitta.Toivonen@tkk.fi (Riitta Toivonen).

join an already established social network.

We note that the dynamical models are defined such that for each network realization, the algorithm is started from scratch and iterated until the stationary distributions are reached, where a single realization is then picked. The dynamical SA models bear some resemblance to another class of stochastic network models, the exponential random graph models (ERGM), recently reviewed by Robins et al. (2007). An essential difference is that the ERGM models are defined as a probability distribution of networks, defined by assumptions about dependencies among the random variables which correspond to links in the network. Sometimes a model defined in the ERGM form can also be interpreted as a stochastic algorithmic model. Whenever in an ERGM model the probability of each link existing is independent of other links, a realization from this distribution can equivalently be drawn starting from an empty network of  $N$  nodes, going through each potential link in the network, and creating it with the corresponding probability. Our study includes one such model (Wong et al., 2005) from which we have generated realizations in this manner. If the formulation of the ERGM model implies correlations between link probabilities however, such as when both the average number of triangles and the average number of links in the networks are specified in the distribution, designing an algorithmic model to produce realizations from the same distribution is not straightforward.

With respect to exponential random graph models, stochastic algorithmic models are typically computationally less expensive. The dynamical stochastic algorithmic models treated in this paper can easily produce networks of 10 000 nodes, and relatively easily up to 100 000 nodes. The growing models could produce networks with hundreds of millions of nodes. The networks that are typically studied with ERGM models consist of hundreds or at most a couple thousand nodes (Goodreau, 2007). The ability to generate large networks is advantageous in many respects. First of all, larger systems provide better statistics. Modern electronic communications provide very large systems and excellent statistics on social networks, and imitating the found structure requires generating large systems. Secondly, in simulation of social dynamics, the ability to generate large systems is required for studying the dependency of observed behavior with system size. Using a small system of, say, 100 nodes, we could not make inferences about how the same process would behave in a society-wide network. For computational reasons we cannot simulate systems of natural size, but the closer we get, the more realistic are the results.

Knowledge of the structure produced by various mechanisms is not only useful from the viewpoint of network evolution, but also simulation of social dynamics. The structure of social networks has been shown to influence many social processes such as opinion formation, the spread of information, and search in social networks, as reviewed by Castellano et al. (2007) from the point of

view of physicists. For example, the connectivity distribution of individuals in the network has turned out to be an important factor in spreading processes of information, fashions, or infective diseases. High connectors are effective in spreading information across the network (Moreno et al., 2004). Also correlations between the degrees of neighboring nodes have been observed to affect dynamics such as epidemics (Boguñá et al., 2002), which are analogous to the spread of information. Many other network features, such as community structure, have also been noted to affect social dynamics. For example, in dynamics simulating competition of different opinions in a population, fairly isolated clusters of individuals may hold on to a minority opinion (Castelló et al., 2007), and a study of recommendations between users of online shopping sites has shown that communities based on a particular common interest also shared recommendations about other topics (Leskovec et al., 2008). For realistic modeling of social dynamics, the network models used for simulation need to reproduce the essential features of real social networks. Observations on the ability of different models to produce the essential features for the dynamics in question help in selecting or designing an appropriate network model for simulation. This is our secondary motivation for the comparison of various models for social networks which we undertake in this paper.

Due to the very large number of models proposed for social networks, we cannot include them all in our study. Because of our focus on acquaintance networks, we exclude the many models designed to capture the features of affiliation networks. Several models were excluded either because the mechanism behind it was difficult to interpret in the context of social ties, or because some feature in the model would have ensured it will not look like our data, such as all nodes having an equal number of connections. We also left out a category of models combining the evolution of intrinsic node properties and network structure (reviewed in Castellano et al. (2007)), because we want to focus on simple tie formation mechanisms instead of more complex coevolutionary models. We were left with the above mentioned categories, which produce distinct kinds of networks. We note that some of the models were designed with a particular property in mind, such as a high average clustering coefficient. We will, nevertheless, assess their ability to reproduce many of the typical features of social networks, although these may not have been the focus of the authors.

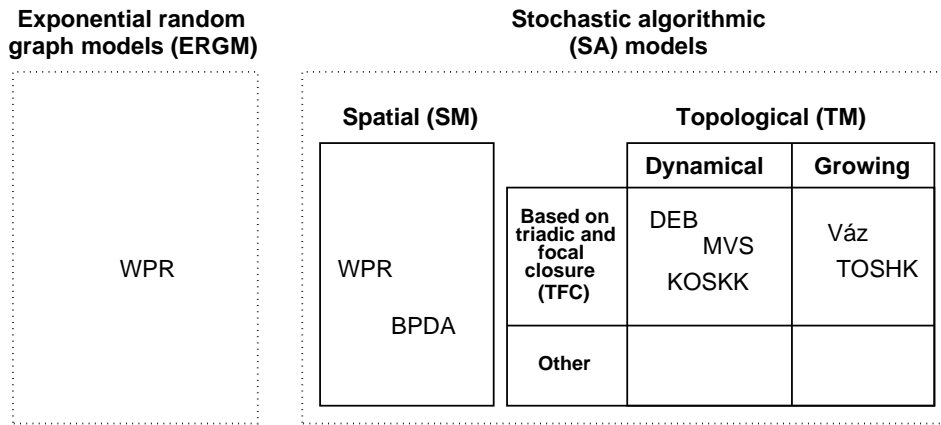


Fig. 1. In this paper, we study mechanisms employed in stochastic algorithmic (SA) models. The seven representative models displayed in the figure are discussed in Section 2. One of the spatial models, WPR, was originally presented as an exponential random graph model, but it can be equivalently formulated as a stochastic algorithmic model. Within the category of topological models, we focus on models based on triadic and focal closure (TFC), which are commonly used mechanisms imitating link formation in social networks.

## 2 Topological and spatial models

### 2.1 Topological models (TM) combining triadic and focal closure (TFC)

Network sociology identifies (a) *cyclic closure* and (b) *focal closure* as two fundamental mechanisms of tie formation (Kossinets and Watts, 2004). Cyclic closure refers to forming ties with one’s network neighbors, whereas focal closure refers to forming ties independently of the network distance, and is attributed to forming social ties through shared activities (hobbies etc.). Triadic closure is the most frequent type of cyclic closure. A combination of triadic and focal closure (TFC) forms the basis of a large number of stochastic algorithmic models for social networks. We will study a representative selection of such models, both dynamical and growing, and compare them with selected spatial models. Tables 1, 2 and 3 contain more detailed descriptions of the models and their parameters. As most authors did not name their models, we label the models using author initials.

#### 2.1.1 Dynamical TFC models

We will first look at three dynamical models that combine triadic and focal closure for creating new links (dynamical TFC): DEB (Davidsen et al. (2002)), MVS (Marsili et al. (2004)), and KOSKK (Kumpula et al. (2007)). Dynamical models in which new links are continuously added also need to have some mechanism for removing links, to avoid ending up with a fully connected net-

work. The different ways of implementing triadic closure and deletion of links are highlighted in Fig. 2. Davidsen et al. (2002) proposed a very simple dynamical model (DEB) where nodes repeatedly select a random pair from among their neighbors, and introduce them to each other if they are not already acquainted. This is labeled triad formation mechanism T1 in Fig. 2. In the DEB model, links are deleted when a node “leaves” the network and hence all of its links are cut. This method of removing links is called *node deletion* (ND). A new node then takes its place so that the number of nodes is kept constant. Due to its simplicity, the model is useful for understanding subtle differences resulting from details in model specification (Section 4.2).

Marsili et al. (2004) formulated their model (MVS) along similar lines, referring to the observation that personal acquaintances play a prominent role in individual search (Granovetter, 1973). In this dynamical model, new contacts are made through search via friends: A node “asks” one of its neighbors  $j$  to introduce it to one of  $j$ ’s neighbors. This mechanism is labeled T2. In the MVS model, each link has a probability  $\lambda$  of being cut at each time step (link deletion, LD), unlike in the DEB model where all links of a particular node were removed simultaneously. Marsili et al. (2004) did not mention which value of  $\lambda$  they used in the MVS model. We fixed  $\lambda = 0.001$  in our simulations, giving each tie an average lifetime of 1000 time steps. Tables 1, 2 and 3 list the parameters of each model (the fixed ones are in parentheses), and indicate the number of free parameters. The fixed parameters were selected according to the original authors’ choices wherever possible.

A weighted social network model by Kumpula et al. (2007) (KOSKK) uses similar mechanisms as the MVS model, with the notable exception that interaction strength is taken into account: new links are created preferably through strong ties, every interaction making them even stronger. In addition to satisfying the weak ties hypothesis (Granovetter, 1973), this mechanism is able to produce much clearer community structure (Kumpula et al., 2007). Another meaningful difference between the KOSKK and MVS models is the way links are deleted: MVS uses link deletion, whereas KOSKK uses node deletion (Fig. 2). Differences in network structure resulting from this choice are explored in Section 4.2.

When generating network realizations, the dynamical models MVS, DEB, and KOSKK are iterated until monitored distributions become stationary. Sometimes the authors do not state which particular criterion they used. We determined for each model how many iterations it takes until the average values of *degree*, squared degree, and *clustering coefficient* stabilize (please see Appendix A.1 for the definitions), and their distributions appear stationary. When generating networks, we used a larger number of iterations than this limit, to ensure that we pick a realization from the region of stationary distributions.

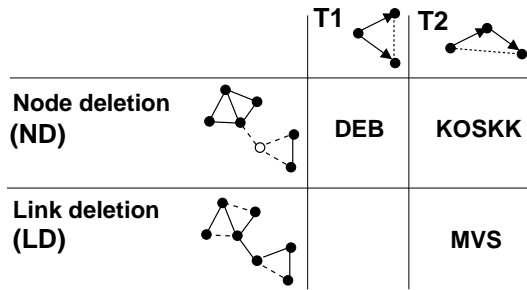


Fig. 2. The models DEB, KOSKK, and MVS classified according to the mechanisms for triadic closure and link deletion employed in them. (T1): introduce two neighbors, (T2): look for a neighbor of a neighbor. (ND): node deletion (all links of a node are deleted), (LD): link deletion (uncorrelated links are deleted). Note that there are many other differences between the models which are not indicated in the table, related to the way random contacts are added, connection strength, number of free parameters, etc. More detailed descriptions of the models are found in Table 1.

### 2.1.2 Growing TFC models

Vázquez (2003) proposed a growing model (Váz) inspired by the DEB model to enable analytical derivations of certain network statistics. In this model, a newcomer node first links to a random node  $i$  in the network, creating *potential edges* (Vázquez’s term) between itself and the neighbors of  $i$ . These ties may be realized later, generating triangles in the network. A growing model for social networks has also been proposed by Toivonen et al. (2006) (TOSHK). In this model, each new node links to one or more ‘initial contacts’, which in turn introduce the newcomer to some of their neighbors. The Váz and TOSHK models are similar in that triangles are generated only between the newcomer and the friends of its initial contact. In the Váz model, they can be created at a later time, while in TOSHK all links are created at once. As with all the models, we keep to the authors’ choices presented in the original paper. In the TOSHK model, we allowed one or two initial contacts, determined by the probability  $p$  (see Table 2), and picked the number of secondary contacts from the uniform distribution  $U[0, k]$ , although this clearly limits the adaptability of the model.

## 2.2 Spatial models

*Homophily* (McPherson et al., 2001), the tendency for like to interact with like, is also known to structure network ties of various types, including friendship, work, marriage, information transfer, and other forms of relationship. This is the starting point for several social network models: a link between two nodes is formed more likely if they are similar, i.e. if they are close to each other in some social characteristics or in geographical space. Hence, in these models nodes are located in a space which can be either one- or many-dimensional, depending

Table 1

CATEGORY: TOPOLOGICAL. DYNAMICAL TFC. Three representative dynamical models for social networks based on triadic and focal closure (TFC).

Parameters	Mechanisms. Number of nodes $N$ fixed; repeat steps for I) adding ties and II) deleting ties until stationary distributions reached
<b>DEB</b> (Davidsen et al., 2002)	
2 free $N, p$	I) Select a node $i$ randomly, and a) if $i$ has fewer than two ties, introduce it to a random node b) otherwise pick <i>two neighbors</i> of $i$ and introduce them if they are not already acquainted. II) Select a <i>random node</i> and with prob. $p$ <i>remove all of its ties</i> .
<b>MVS</b> (Marsili et al., 2004)	
3 free $N, \xi, \eta$ ( $\lambda=0.001$ )	I) Select a node $i$ randomly, and a) connect $i$ to another random node with prob. $\eta$ . b) select a <i>friend's friend</i> of $i$ with prob. $\xi$ and introduce $i$ to it if not already acquainted. II) Select a <i>random tie</i> and delete it with prob. $\lambda$ .
<b>KOSKK</b> (Kumpula et al., 2007)	
3 free $N, p_\Delta, p_r$ ( $w_0 = 1,$ $p_d = 0.001,$ $\delta = 0.5$ )	I) Select a node $i$ randomly, and a) select a <i>friend's friend</i> (by <i>weighted search</i> ) and introduce it to $i$ with prob. $p_\Delta$ (with initial tie strength $w_0$ ) if not already acquainted. <i>Increase tie strengths</i> along the search path by $\delta$ . b) additionally, with prob. $p_r$ (or with prob. 1 if $i$ has no connections), connect $i$ to a random node $j$ (with tie strength $w_0$ ). II) Select a <i>random node</i> and with prob. $p_d$ <i>remove all of its ties</i> .

Nodes represent individuals and links represent ties between them. Parameters that were fixed according to the original authors' choices are shown in parentheses.

on how many different traits are taken into account. The probability of a link existing between any two nodes is then defined as a function of the distance between them. We selected two spatial models in which nodes are uniformly distributed in the underlying social or geographical space, BPDA (Boguñá et al. (2004)) and WPR (Wong et al. (2005)). They differ in the dependence of link probability on distance and in the employed distance measure. While the authors mention that a social space of any dimension could be used, they analyse more closely the cases of 1D and 2D, respectively. We keep to their choices.

### 3 Fitting the models

In order to make networks generated by different models comparable, we need to unify some of their properties. To this end, we fit the models to two real-

Table 2

CATEGORY: TOPOLOGICAL. GROWING TFC. Two representative growing models proposed for social networks, where ties are based on triadic and focal closure (TFC).

Parameters	Mechanism. Repeat steps for a) adding nodes and ties b) adding ties only until network contains $N$ nodes.
<b>TOSHK</b> (Toivonen et al., 2006)	
3 free $N, p, k$ (simplified)	a) Add a new node $i$ to the network, connecting it to one random initial contact with probability $p$ , or two with probability $1 - p$ . b) for each random initial contact $j$ , draw a number $m_{sec}$ from the distribution $U[0, k]$ and connect $i$ to $m_{sec}$ neighbors of $j$ if possible.
<b>Váz</b> (Vázquez, 2003)	
2 free $N, u$	a) with probability $1 - u$ , add a new node to the network, connecting it to a random node $i$ . Potential edges are created between the newcomer $n$ and the neighbors $j$ of $i$ (a potential edge means that $n$ and $j$ have a common neighbor, $i$ , but no direct link between them). b) with probability $u$ , convert one of such potential edges generated on any previous time step to an edge. Potential edges generated by converting an edge are ignored.

world data sets with respect to as many of the most relevant network features as the model parameters allow. We selected two social network data sets with slightly different average properties in order to get a better picture of the adaptability of the models. In particular, we chose two data sets with different average degree, because we assumed that link density could be an important factor in the resulting network structure. These data sets display typical properties of large social networks: degree distributions that imply the presence of high degree nodes, high average clustering coefficients  $\langle c \rangle$ , decreasing clustering spectra  $c(k)$ , and positive degree-degree-correlations  $r$  (assortativity) (please see Appendix A.1 for the definitions). These features have also been observed in numerous other large scale social networks (Leskovec and Horvitz (2008), Onnela et. al (2007b), Gleiser and Danon (2003)).

In this section, we discuss our choice of data sets and the targeted features used for fitting. Finally, we display the optimized parameters and basic properties of the networks generated with these parameters. Definitions of the common network measures we use, as well as a description of the optimization methods used, are included in the appendix.

Table 3

CATEGORY: SPATIAL. Two representative *spatial models* (SM) for social networks, in which link probability between two nodes depends on their distance in an underlying social or geographical space.

Parameters	Mechanism
<b>BPDA</b> (Boguñá et al., 2004)	
3 free $N, \alpha, b$	Distribute $N$ nodes uniformly in a (1-dimensional) social space (a segment of length $h_{max}$ ). Link nodes with prob. $p = 1/(1 + (d/b)^\alpha)$ , where $d$ is their distance in the social space. ( $h_{max}$ can be absorbed within $b$ ). If treated many-dimensionally, similarity along one of the social dimensions is sufficient for the nodes to be seen as similar.
<b>WPR</b> (Wong et al., 2005)	
4 free $N, H, p, p_b$	Distribute $N$ nodes uniformly in a (2-dimensional) social space of unit size. Link nodes with prob. $p + p_b$ if their distance is smaller than $H$ , and with prob. $p - p_\Delta$ otherwise (where $p_\Delta(p, p_b, H)$ is such that the total fraction $p$ of all possible links is generated.).

We selected the dimensionality (1D or 2D) based on how the authors treated the models when originally presented.

### 3.1 The friendship network at *www.last.fm* and the email network

Data of large acquaintance networks which reliably track social connections are not easily available. Affiliation networks, which link two individuals if they belong to the same affiliation (e.g. board) are more abundant and much researched, but they have particular features deriving from their construction, such as a very high number of cliques. We chose two acquaintance networks of suitable size, on the order of 1 000 to 10 000 individuals. Smaller networks would have provided less accurate distributions, and with much larger networks, finding the optimal parameters of the dynamical models would have been slow, because optimization requires generating very many instances of the networks with different parameters.

The larger of the two data sets against which we fit our models is a mutual friendship network collected from users of the web site *www.last.fm*, where people can share their musical tastes and designate other users as their friends. We used for this study the friendship information only, disregarding the musical preferences. Because there are several hundred thousand users on the site worldwide, we selected users in one country, Finland, to obtain a smaller network with 8003 individuals. The country labels were self-reported. This data (henceforth called *lastfm*) represents the largest connected component of Finnish users at this site. Individuals in the resulting network have on the average  $\langle k \rangle = 4.2$  friends, high clustering  $\langle c \rangle = 0.31$ , and the network is

highly assortative with  $r = 0.22$ , indicating that friends of those users who have many connections at the site are themselves well connected (please see Appendix A.1 for definitions). After designating someone as a friend, there is no cost to maintaining the tie, i.e. the link never expires. This means that the data may overestimate the number of active friendships within the last-fm web site. However, the degree distribution does not imply higher degree nodes than those observed in a network constructed from mobile phone calls (Onnela et al, 2007b), in which each contact has a real cost in time and money. Requiring ties to be reciprocated ensures that the users have at least both acknowledged one another.

The smaller data set is an acquaintance network collected by Guimerà et al. (2003), based on emails between members of the University Rovira i Virgili (Tarragona). In this network, two individuals are connected if each sent at least one email to the other during the study period. Bulk emails sent to more than 50 recipients were eliminated. Again, we use the largest connected component of the network. It consists of 1133 individuals, and it is a compact network with average shortest path length  $\langle l \rangle = 3.6$ , slightly larger average degree  $\langle k \rangle = 9.6$ , fairly high average clustering coefficient  $\langle c \rangle = 0.22$ , and fairly small assortativity  $r = 0.08$ .

Both of our real world networks are unweighted, meaning that tie strengths are not specified. All of the models studied here apart from KOSKK are unweighted as well. Averaged basic statistics of both data sets are displayed in Table 5. The degree distributions, clustering spectrum and degree-degree correlations of the *lastfm* and *email* networks are shown in Fig. 3, and more plots of their statistics are shown in Section 4.1 in connection with the fitted models.

### 3.2 Choice of target features for fitting the models

The most important features that we wish to align between the models and the data are the number of nodes and the number of links. Because both of our data sets are connected components of a larger network, we will focus on the properties of the largest connected component of the generated networks. Our first two fitting targets are largest connected component size  $N_{LC}$  and the average number of links per node, or average degree  $\langle k \rangle$ , within the largest component. They are already sufficient for fitting the DEB model, which has only two parameters.

A natural choice for the next target once the number of individuals and density of links are fixed is some measure related to triangles, since triangle density is known to be an important characteristic of social networks. One possible

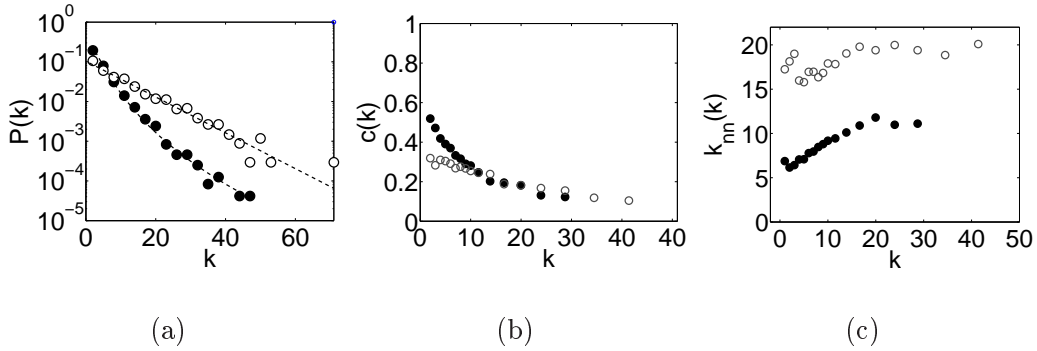


Fig. 3. Properties of the *lastfm* data set ( $\bullet$ ) and the *email* data ( $\circ$ ). a) degree distributions, with average degrees  $\langle k \rangle = 4.2$  and  $9.6$ , respectively. Guimerà et al. (2003) fitted to the *email* data an exponential distribution  $p(k) = e^{-k/k^*}$  with  $k^* = 9.2$ , which shows as a straight line in a semilogarithmic plot. The lognormal distribution fitted the *lastfm* data best of the different distributions we tried (exponential, Weibull, gamma, and lognormal), although not perfectly. b) Clustering  $c(k)$  decreases with degree  $k$  (average clustering  $\langle c \rangle = 0.31$  and  $0.22$ , respectively). c) Degree-degree correlations between nodes and their neighbors ( $k_{nn}$  signifies average nearest neighbor degree) show that both networks are assortative (with  $r = 0.22$  and  $r = 0.08$ , respectively).

choice would be the average number of triangles in the network. We will use the average clustering coefficient  $\langle c \rangle$  (see Appendix A.1 for the definition). These three features are sufficient for fitting the models TOSHK, BPDA, MVS, and KOSKK, if we fix some of the parameters according to the original authors' choices (see Table 1).

If matching  $N_{LC}$ ,  $\langle k \rangle$  and  $\langle c \rangle$  is not enough to fix all parameters of the model, we no longer have a straightforward choice. We considered using the assortativity coefficient and shortest path lengths (see A.1). In the WPR model, assortativity varies closely together with the average clustering coefficient as parameters are varied, so it could not be used as a fourth target feature. Instead, we used the average shortest path length to fit the fourth parameter. We also attempted using the assortativity coefficient for fitting the KOSKK model, when the weight increment parameter  $\Delta$  was allowed to vary, but ran into a different problem: attempting high assortativity forced the weight increment parameter to zero, thereby eliminating an important feature of the weighted model and weakening the community structure. Hence, we fixed the weight increment parameter  $\delta = 0.5$  in accordance with the authors' choice.

All of these measures - degree, clustering, assortativity, and path lengths - are important features of social networks and are likely to affect dynamics such as opinion formation or spreading of information in a network. Using averages, we obviously lose information compared to the distributions of each measure. However, typically each model produces similarly shaped distributions with

most choices of parameters, and the shape of the distribution cannot be controlled in any case.

Table 4 indicates which features were targeted when optimizing the parameters of each model, and displays the optimized parameters, and Table 5 displays properties of the networks generated with these parameters. Of course, due to the stochastic nature of the models, two network realizations generated with the same parameters are not likely to have exactly the same average properties. The plots and tables concerning the model networks in this paper always contain values averaged over 100 network realizations.

We observe that for all models and both data sets,  $\langle N_{LC} \rangle$  and  $\langle k \rangle$  fit closely. Note that for the models with two free parameters (DEB, Váz), once largest component size and average degree were fitted, we had no control over clustering, assortativity, path lengths, or any other network features. The TOSHK model, with its discrete parametrization of the number of triangles formed, was not able to exactly match the clustering values despite having three parameters. Average shortest path lengths matched closely for all but the spatial model treated in one dimension (BPDA), although the measure was only used for fitting the WPR model. The other features shown in the table,  $r$  and  $l_{max}$ , were not used for fitting but are shown because they are relevant for social networks.

## 4 Comparing higher order statistics

Having fitted the models according to average values of particular network characteristics, we will compare the distributions of these characteristics, as well as the community structure of the resulting networks. The first point we focus on is the distribution of connectivity in the network. Next, we address the clustering spectrum  $c(k)$ , which will turn out to be very different in the spatial models compared to the data. Path lengths as a measure of network compactness are also observed. Finally, we will observe how the networks are structured into clusters, a very important feature of social networks. In Section 4.1, we compare the model types described above, and in Section 4.2 we will combine and compare different network evolution mechanisms employed in the dynamical TFC models.

### 4.1 Properties of the spatial and topological models

The degree distribution of the *email* data set falls exponentially, while the *lastfm* data has a fatter tail (Fig. 3). In larger data sets based on one-to-

Table 4

Network features targeted when fitting each of the models, and the parameters that provided the best match to the *lastfm* and *email* data sets.

DEB	matched to $N_{LC}$ , $\langle k \rangle$ <i>lastfm</i> : $N = 8330, p = 0.203$ <i>email</i> : $N = 1138, p = 0.064$
MVS	matched to $N_{LC}$ , $\langle k \rangle$ , $\langle c \rangle$ <i>lastfm</i> : $N = 9300, \xi = 0.0022, \eta = 0.000368$ <i>email</i> : $N = 2270, \xi = 0.0062, \eta = 0.000071$
KOSKK	matched to $N_{LC}$ , $\langle k \rangle$ , $\langle c \rangle$ <i>lastfm</i> : $N = 8205, p_{\Delta} = 0.0029, p_r = 0.0008$ <i>email</i> : $N = 1135, p_{\Delta} = 0.0107, p_r = 0.0039$
TOSHK	matched to $N$ , $\langle k \rangle$ , $\langle c \rangle$ <i>lastfm</i> : $N = 8003, p = 0.60, k = 1$ <i>email</i> : $N = 1133, p = 0.06, k = 3$
Váz	matched to $N$ , $\langle k \rangle$ <i>lastfm</i> : $N = 8003, u = 0.524$ <i>email</i> : $N = 1133, u = 0.793$
BPDA	matched to $N_{LC}$ , $\langle k \rangle$ , $\langle c \rangle$ <i>lastfm</i> : $N = 8250, \alpha = 1.915, b = 1.51 \cdot 10^{-4}$ <i>email</i> : $N = 1133, \alpha = 1.565, b = 0.002032$
WPR	matched to $N_{LC}$ , $\langle k \rangle$ , $\langle c \rangle$ , $\langle l \rangle$ <i>lastfm</i> : $N = 8200, H = 0.0108, p = 0.000506, p_b = 0.9994$ <i>email</i> : $N = 1133, H = 0.040, p = 0.008498, p_b = 0.991$

$N_{LC}$ : Largest component size (number of nodes),  $\langle k \rangle$ : average degree,  $\langle c \rangle$ : average clustering coefficient,  $\langle l \rangle$ : average shortest path length.  $\langle k \rangle$ ,  $\langle c \rangle$ , and  $\langle l \rangle$  were calculated for the largest component of the network.

one communication, broader degree distributions have been observed, either power law (Lambiotte et al. (2008)) or power law with exponential cutoff (Onnela et. al (2007b)). We assume therefore that an exponential or broader degree distribution is a valid assumption for a large acquaintance network. The models show three kinds of connectivity distributions with very different behavior with respect to nodes with high connectivity (Fig. 4). The spatial models have Poisson distributions, with no nodes having very large degree. The Váz model has a power law distribution, which implies a few nodes with extremely many connections. The dynamical TFC models (DEB, MVS, and KOSKK) and the TOSHK model fall somewhere in between. The very different distributions in TOSHK and Váz, despite both being growing models, may be due to the limit on the number of links formed per time step in TOSHK. It appears that as models of large acquaintance networks, the TFC models produce the most realistic connectivity distributions.

The shape of the degree distribution of a spatial model depends on the underlying distribution of the nodes in space. The uniform distribution employed

Table 5

Basic statistics of the *lastfm* and *email* data sets and the models fitted to each.

<i>model / data</i>	$N_{LC}$	$L$	$\langle k \rangle$	$\langle c \rangle$	$r$	$\langle l \rangle$	$l_{max}$
Last-fm-fin	8003	16824	4.20	0.31	0.22	7.4	24
DEB	8009 ± 30	16858 ± 224	4.21 ± 0.05	0.38 ± 0.01	0.10 ± 0.01	7.0 ± 1.6	18.1 ± 1.4
MVS	7989 ± 38	16816 ± 153	4.21 ± 0.03	0.30 ± 0.01	0.02 ± 0.01	7.8 ± 1.6	17.4 ± 1.0
KOSKK	8006 ± 20	16849 ± 207	4.21 ± 0.05	0.31 ± 0.01	0.05 ± 0.01	7.2 ± 1.5	16.3 ± 0.9
TOSHK	8003	16791 ± 93	4.20 ± 0.02	0.34 ± 0.01	0.14 ± 0.01	6.6 ± 1.3	13.8 ± 0.6
Váz	8003	16801 ± 171	4.20 ± 0.04	0.29 ± 0.01	0.27 ± 0.02	8.3 ± 2.6	22.6 ± 1.5
BPDA	8005 ± 31	16794 ± 141	4.20 ± 0.03	0.29 ± 0.01	0.30 ± 0.02	23.9 ± 9.3	60.1 ± 8.0
WPR	8004 ± 19	16972 ± 150	4.24 ± 0.03	0.29 ± 0.01	0.30 ± 0.02	8.1 ± 1.6	18.2 ± 1.1

<i>model / data</i>	$N_{LC}$	$L$	$\langle k \rangle$	$\langle c \rangle$	$r$	$\langle l \rangle$	$l_{max}$
Email	1133	5451	9.62	0.22	0.08	3.6	7
DEB	1133 ± 3	5452 ± 249	9.62 ± 0.43	0.45 ± 0.01	0.06 ± 0.02	3.4 ± 0.9	7.7 ± 0.7
MVS	1113 ± 1	5282 ± 77	9.48 ± 0.14	0.23 ± 0.01	0.05 ± 0.04	3.8 ± 1.1	9.6 ± 0.6
KOSKK	1134 ± 2	5425 ± 193	9.57 ± 0.34	0.22 ± 0.01	0.06 ± 0.02	3.5 ± 0.9	7.5 ± 0.6
TOSHK	1133	5453 ± 52	9.63 ± 0.09	0.29 ± 0.01	0.09 ± 0.02	3.4 ± 0.8	6.1 ± 0.3
Váz	1133	5453 ± 136	9.63 ± 0.24	0.42 ± 0.02	0.12 ± 0.03	4.6 ± 1.7	13.6 ± 1.4
BPDA	1133 ± 1	5477 ± 172	9.67 ± 0.30	0.22 ± 0.01	0.22 ± 0.02	4.4 ± 0.8	8.4 ± 0.5
WPR	1133 ± 1	5448 ± 72	9.62 ± 0.13	0.21 ± 0.01	0.20 ± 0.03	3.6 ± 0.7	6.0 ± 0.2

All statistics are calculated for the largest component of each network.  $N_{LC}$ : Largest component size,  $L$ : number of links,  $\langle k \rangle$ : average degree,  $\langle c \rangle$ : average clustering coefficient,  $r$ : assortativity coefficient,  $\langle l \rangle$ : average shortest path length, and  $l_{max}$ : longest shortest path. The values are averaged over 100 realizations of each network model. The standard error of the averages is displayed whenever there was fluctuation in the values.

in the models treated here leads to a Poisson distribution of degrees, implying the absence of nodes with very high connectivity (Boguñá et al. (2004) derived the distribution for their model, and the Poisson distribution also fits the WPR model very well (Fig. 4)). The homophily principle does not always lead to a Poisson distribution, however. Masuda and Konno (2006) used an exponentially distributed fitness parameter as the basis for homophily, and obtained a flat degree distribution  $p(k)=\text{const}$ . As they observe, this is unrealistic and an indication that the homophily mechanism is insufficient in itself. The homophily principle, when combined with some other mechanism, can also lead to a broader degree distribution (Masuda and Konno, 2006).

Next, let us have a look at clustering in the networks, another important feature of social networks. The average clustering coefficient can be tuned in the models by varying the model parameters, but the shape of the  $\langle c(k) \rangle$  curve is likely to be invariable. For both of our data sets (Fig. 5) and many other acquaintance networks (e.g. Onnela et. al (2007b), Leskovec and Horvitz (2008)), clustering  $c(k)$  decreases with increasing connectivity  $k$  of a node. Many network models, including the TFC models studied here, display an inverse relation between node degree and clustering:  $c(k) \sim \frac{1}{k}^1$ . The homophily mechanism on which the spatial models are based, when the distribution of nodes in

<sup>1</sup> This follows naturally in any model where an increase in the number of links of a node goes hand in hand with an increase in the number of triangles around it. If on

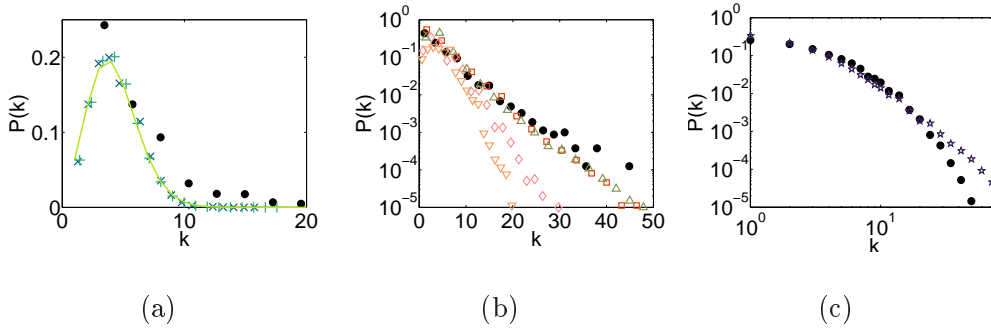


Fig. 4. Degree distributions  $P(k)$ . The exact shapes of the distributions are not as important as the markedly different probabilities of obtaining high degree nodes. The models were fitted to the *lastfm* data ( $\bullet$ ), shown in each panel for reference. a) The spatial models (+ BPDA,  $\times$  WPR) have Poisson degree distributions (the fitted line is a Poisson distribution). b) The tails of the degree distributions produced by the dynamical models ( $\nabla$  MVS,  $\square$  DEB, and  $\diamond$  KOSKK), and the growing TOSHK model ( $\triangle$ ) all decay clearly slower than the Poisson distribution produced by the spatial models, but faster than power law. Right: the Váz model degree distribution ( $\star$ ), shown in a double logarithmic plot, is very broad, implying that very high degree nodes are observed (Vázquez (2003) has shown the tail to decay as power law,  $P(k) \sim k^{-\gamma}$ ). All plots of properties of the model networks in Section 4 display averages over 100 network realizations.

the underlying social space is uniform, produces a flat clustering  $c(k) = \text{const.}$  Since this is strikingly different from observed data, it does not seem plausible that homophily (at least when implemented in this manner) could be the only mechanism at play in the formation of social networks, supporting the finding by Masuda and Konno (2006). The TFC models produce a plausible decreasing relation between node degree and its clustering coefficient.

An explanation of why  $c(k)$  is flat for the spatial models treated here can be given as follows. The clustering coefficient  $c_i$  of node  $i$  is defined as the fraction of pairs of neighbors of node  $i$  which are directly connected. Thus, in order to determine  $c(k)$ , we need to know how the probability of a link existing between two neighbors of a node  $i$  depends on the degree  $k_i$ . If we assume that the underlying distribution of nodes is homogeneous and the nodes are independently and identically distributed in space, it follows that regardless of node degree, its neighbors are similarly distributed in any disc centered at the node. Hence, picking randomly two of its neighbors, the distance between them does not depend on how many other nodes there happen to be within the disc, implying that the probability of a link between them does not depend on the degree  $k$  of  $i$ . This is seen to hold whether the link probability is a step

---

average increasing the degree  $k$  of a node by one is accompanied by an increase of the number  $N_\Delta$  of triangles around the node by  $a$ , the resulting clustering coefficient for a node of degree  $k$  will be on average  $c(k) = \frac{N_\Delta}{k(k-1)/2} = \frac{ak}{k(k-1)/2} \approx \frac{2ak}{k^2} = \frac{2a}{k}$ .

function (as in WPR), or depends linearly on distance (not shown), or is some other monotonously decreasing function of distance (as in BPDA).

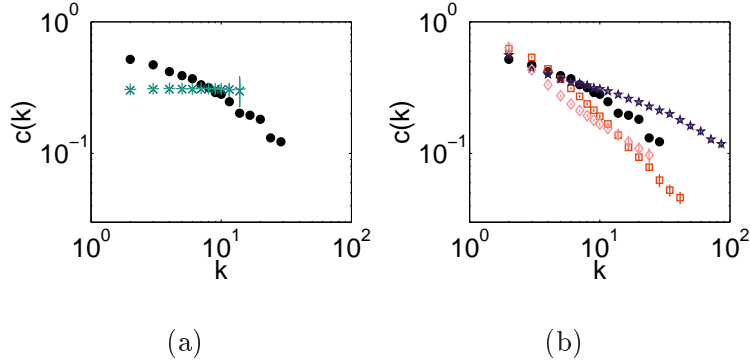


Fig. 5. Clustering spectra  $c(k)$  in the *lastfm* data ( $\bullet$ ) and the models fitted to it. a) It is notable that the curve is flat for the spatial models (+ BPDA,  $\times$  WPR), implying that the neighbors of high connectors have the same probability of being directly connected as the neighbors of a small degree node, unlike in the data. b) In both the TFC models and the data, clustering decreases with degree  $k$ . Representative TFC models  $\square$  DEB,  $\diamond$  KOSKK, and  $\star$  Váz are shown. Error bars denote standard deviation of the averages of each network in each bin. Bins containing fewer than 20 elements were omitted because they would not display reliable averages and standard deviation.

The compactness of a network is described by the distribution of *path lengths* (Fig. 6, right). All of the models in this study include a mechanism for generating random links across the network, which create shortcuts and greatly reduce path lengths. It follows that, apart from the spatial model treated one-dimensionally (BPDA), in which path lengths are strikingly long compared to the data, all networks display reasonable path lengths. The median path lengths in the spatial model treated in 2D are also slightly larger than in the data. The dynamical TFC models and TOSHK are perhaps slightly too compact, with largest path lengths falling below those in the data. The differences are not very important, however, apart from the 1-dimensional case. For reference, even in an extremely large acquaintance network of several million individuals worldwide (Leskovec and Horvitz, 2008), the average distance between two individuals was 6.6, and path lengths up to 29 were found.

We would naturally also like to assess the community structure, or clusters within the networks. Perhaps the simplest possible measure of community structure is the number of cliques, or fully connected subgraphs, of different sizes in the networks (Fig. 7). Because the number of nodes and links in the networks is the same, the different numbers of cliques are an indicator of the distribution of links in the network and not simply of global link density. The spatial models fare best, producing a clique size distribution comparable to the data sets in both fits. The dynamical TFC models have trouble producing

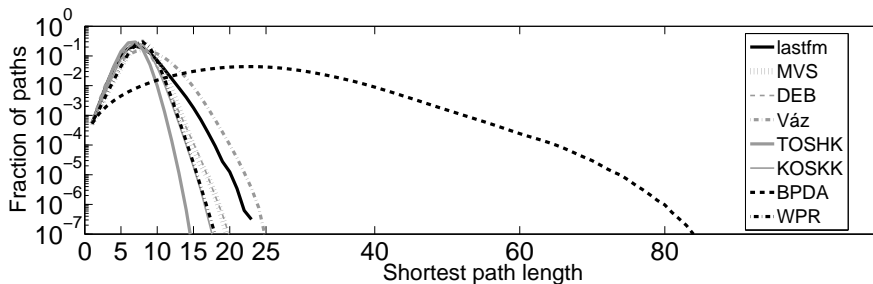


Fig. 6. Shortest path lengths in the *lastfm* data (thick black line) compared with the models fitted to it. The spatial model treated as 1-dimensional (BPDA) naturally has the longest shortest paths. Otherwise there are no great differences. The Váz model, surprisingly, has rather long shortest paths despite its broad degree distribution. Generally, high degree nodes decrease path lengths across the network, but the high assortativity of the Váz networks seems to counter the effect. The other models are slightly too compact, and the spatial model treated two-dimensionally, WPR, has slightly too long median path lengths.

large enough cliques when link density is low (the *lastfm* fits). MVS produced the fewest large cliques, followed by KOSKK and DEB. A possible explanation of why the MVS model produces the fewest cliques is indicated by the comparison of Section 4.2, where node deletion is seen to preserve more cliques than link deletion. In the email fit, with higher link density, the TFC models produce more reasonable numbers of cliques of different sizes. Of the two growing models, Váz produces far too many cliques in the *email* fit, while TOSHK produces too few in both fits. The parametrization of the TOSHK model, requiring that the number of secondary contacts be drawn from a uniform distribution, severely limits the number of coincident triangles and hence cliques which can be formed.

To obtain further understanding of the community structure of the networks, we will employ an edge characteristic called the *overlap*  $O_{ij}$  (Onnela et. al, 2007b), which bears resemblance to the measure of edgewise shared partners. The overlap indicates, for the end nodes  $i$  and  $j$  of the edge, which fraction of their neighbors are common to both. The measure varies between 0 and 1 and is defined as

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}}, \quad (1)$$

where  $n_{ij}$  is the number of neighbors common to both nodes  $i$  and  $j$ , and  $k_i$  and  $k_j$  are their degrees. Overlap is defined for edges with at least one end having degree larger than one. Because we are dealing with the largest connected components of the networks, overlap is defined for all edges. Within a cluster, adjacent nodes tend to share many neighbors, and thus overlap is high, while edges between communities will often have low or zero overlap values. The measure can discern clusters, or communities, in which the end nodes of each link share a larger fraction of their neighbors than the end nodes of links connecting the cluster to the rest of the network.

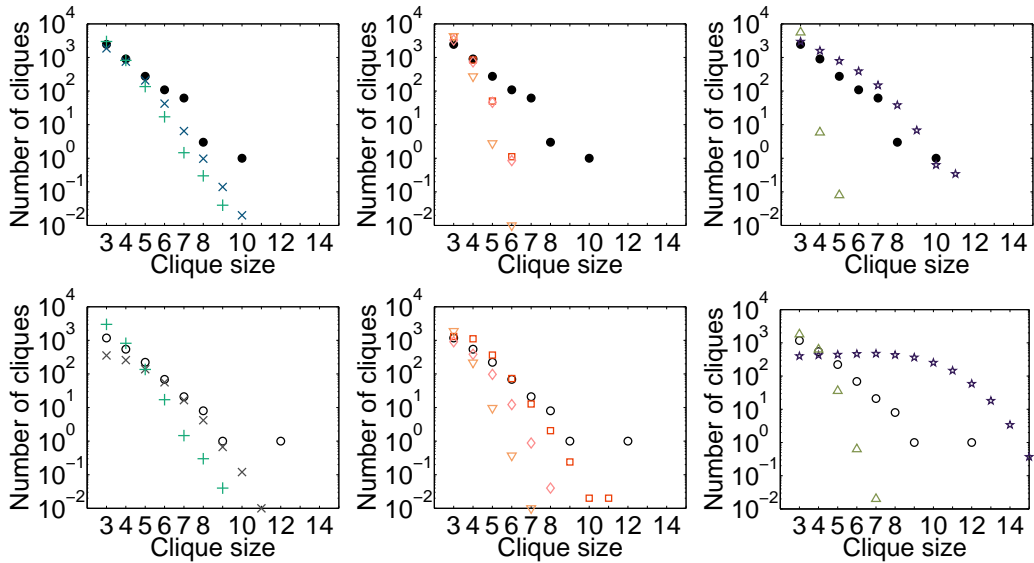


Fig. 7. Assessing the community structure of the networks: Cliques in the model networks fitted to the *lastfm* ( $\bullet$ ) data (top row) and the *email* ( $\circ$ ) data (bottom row). From left to right: spatial models (+ BPDA,  $\times$  WPR), dynamical TFC models ( $\square$  DEB,  $\diamond$  KOSKK,  $\nabla$  MVS), and growing TFC models (TOSHK  $\triangle$ , Váz  $\star$ ). Cliques within larger cliques, such as triangles within a 4-clique, are not counted.

In order to separate such clusters and observe their sizes, we now remove low overlap links from each network. Removing links up to a certain overlap value is called *thresholding* by overlap. Thresholding breaks the network into components (connected subsets of nodes). We monitor both the size of the largest component and the densities of the smaller components that break apart. If all components remaining after thresholding are small, the network consists of small clusters held together by low overlap links. Such structure is depicted in Fig. 8(a). If a large component remains, on the other hand, the network has a core which does not consist of such loosely connected clusters, Fig. 8(b). We note that the implicit definition of communities entailed by this approach is useful when we are dealing with networks with very clear and loosely connected clusters. We do not propose it as a general definition of community structure, nor present thresholding by overlap as a general method of community detection suitable for all types of networks.

Fig. 9(a) displays the relative sizes of the largest component after removing zero overlap links for the *lastfm* data and the models fitted to it. The spatial models break down to small clusters, whereas in the other models a large core remains. This seems to indicate that the spatial models have the structure depicted in Fig. 8(a), while the others do not. However, we know that the spatial models have a larger fraction of zero overlap links than most of the others (Table 6). Did they break down while the others didn't simply because a larger number of links was removed from them than from the other models? We can test this by removing an equal fraction of the lowest overlap links from

all networks (41%, the maximum fraction of links removed from any network when only zero overlap links were removed). The result is shown in Fig. 9(b). Again, a core remains intact in most of the TFC models, an exception being the KOSKK model. Panel b) displays component link densities  $d = 2l/s(s-1)$ , where  $s$  is the number of nodes in the component and  $l$  the number of links, in the case where zero overlap links have been removed. The densities are also observed to be higher in the spatial models than in the others (Fig. 9), despite the fact that more links were removed from the spatial models due to their larger fraction of zero overlap links.

We can conclude that spatial models do indeed consist of denser clusters more loosely connected with one another than most of the TFC models and the *lastfm* data. The KOSKK model, which is known to be able to produce very clear community structure (Kumpula et al., 2007), bears resemblance to the spatial models in that it breaks down when small overlap links are removed. In the *email* fits, link density in the network is higher, and it takes slightly larger overlap links to decompose the network to small clusters (not shown), but the difference between the spatial and TFC models remains.

As a complementary observation to the role of low overlap links as bridges between communities, we can ask how many zero overlap links there are in the network, and where they are located. They could reside mostly between clusters, as depicted in Fig. 8(a), or largely in leaves and chains on the 'boundary' of the network, as in Fig. 8(b). Table 6 displays the statistics for both data sets and all models fitted to them. It seems plausible that in a growing network nodes that recently joined the network would produce a lot of leaves on the boundary of the network. This is indeed the case with the Váz model, which has the highest fraction of nodes and links belonging to leaves and chains, and to a large extent also in the data sets. The data sets can also be thought of as growing networks because links and nodes in them accumulate over time and are rarely removed. In the spatial models, a much smaller fraction of zero overlap links and also a smaller fraction of nodes resides in leaves and chains, particularly in the *email* fits.

#### 4.2 Comparison of mechanisms for triadic closure and tie removal

In order to gain better insight on the different mechanisms employed in the dynamical models based on triadic and focal closure, we will combine some of them and compare the resulting networks. We take as a starting point the simplest of the dynamical models, the 2-parameter DEB model, which employs triad formation mechanism T1 and node deletion. Remember that node deletion refers to deleting all the links of a node, whereas link deletion refers to deleting randomly selected links. Both are used as a mechanism of

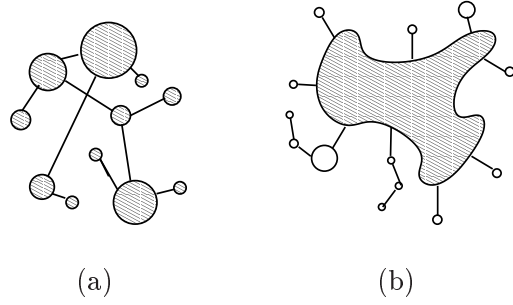


Fig. 8. Schematic depiction of the structural differences related to links with low overlap (links whose end nodes have very few common neighbors). a) Low overlap links connect small, relatively tightly bound clusters together. The spatial models display this type of structure. b) The network contains a core that does not disintegrate when low overlap links are removed, although small clusters and a large number of single nodes do break apart. A large fraction of both nodes and zero overlap links belong to leaves of chains. Most of the TFC models have this structure.

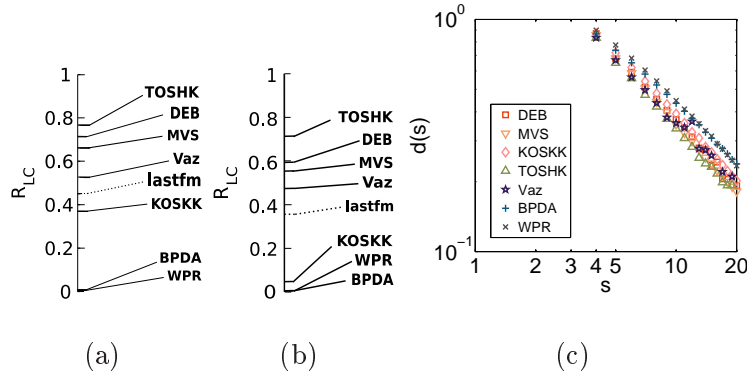


Fig. 9. Panels a) and b) show the relative size of the largest connected component  $R_{LC}$  after removing links, relative to the largest component in the original network, in the *lastfm* data and the models fitted to it. a)  $R_{LC}$  in the models fitted to the *lastfm* data after removing links with overlap  $O = 0$ . The spatial models break down to small components, while in the other models and the data a relatively large core remains intact. b) To show that the breakdown of the spatial models was not simply due to a larger number of links removed, we now remove the same fraction of the lowest overlap links from all models and data (41%, the maximum fraction removed in Fig. 9(a)). Again, a core remains intact in most of the TFC models and the data, while the spatial models and the KOSKK model break down to small clusters. c) Link densities  $d(s)$  in clusters of size  $s$  in the models fitted to the *lastfm* data, after removing links with overlap  $O = 0$ . The largest component is not included. The spatial models show highest cluster density, followed by the KOSKK model and then the other TFC models.

reducing links in order to keep the network from becoming fully connected. By varying the mechanisms of triad formation and deletion of ties, we obtain four different models (see Fig. 10, a). Note that these mechanisms do not yet

Table 6  
 Statistics of zero overlap links.

<i>model / data</i>	% $l_0$	% $l_c$	% $n_c$	<i>model / data</i>	% $l_0$	% $l_c$	% $n_c$
<i>lastfm</i>	31.2	14.3	30.0	<i>email</i>	22.4	2.8	13.7
DEB	20.1 ± 0.5	7.7 ± 0.3	16.1 ± 0.4	DEB	4.1 ± 0.5	1.2 ± 0.2	5.8 ± 0.8
MVS	32.0 ± 0.6	4.8 ± 0.2	10.0 ± 0.3	MVS	17.2 ± 1.9	0.3 ± 0.1	1.6 ± 0.4
KOSKK	34.5 ± 0.5	8.1 ± 0.3	17.0 ± 0.6	KOSKK	33.1 ± 0.9	1.1 ± 0.2	5.4 ± 0.7
TOSHK	22.4 ± 0.4	6.1 ± 0.2	12.8 ± 0.4	TOSHK	4.7 ± 0.3	0.1 ± 0.1	0.4 ± 0.2
Vâz	21.2 ± 0.4	20.6 ± 0.4	43.3 ± 0.6	Vâz	5.2 ± 0.4	5.1 ± 0.4	24.5 ± 1.9
BPDA	35.5 ± 0.6	3.7 ± 0.2	7.8 ± 0.4	BPDA	26.7 ± 0.8	0.01 ± 0.01	0.05 ± 0.07
WPR	40.9 ± 0.5	3.5 ± 0.2	7.5 ± 0.4	WPR	41.2 ± 0.8	0.01 ± 0.01	0.06 ± 0.07

Percentage of links having zero overlap (%  $l_0$ ), and percentages of links (%  $l_c$ ) and nodes (%  $n_c$ ) belonging to leaves and chains as depicted in Fig. 8(b). ( $l_c$  is a subset of  $l_0$ .)

fully determine the models. We will follow the original DEB model in that a node first connects to two randomly chosen nodes before it starts forming triads, after which it no longer makes random connections.

Two findings speak in favor of using the node deletion mechanism: The degree-degree correlation plot (Fig. 10, b) shows that node deletion produces more assortative networks than the link deletion mechanism. In fact, the combination of link deletion with T1 produces dissortative networks. As social networks are typically assortative, it might be recommendable to use node deletion in dynamical models of social networks. Node deletion also preserves more cliques in the network, a desirable feature for social networks, than link deletion (Fig. 10, c). The larger number of cliques preserved by node deletion is not explained by the clustering coefficients, which turned out to be similar in all networks. The parameters were selected such that  $N_{LC}$  and  $\langle k \rangle$  matched the *lastfm* data.

The choice of triangle generation mechanism, on the other hand, is seen to affect the degree distribution. Networks generated with the T1 mechanism have higher degree nodes than those using the T2 mechanism (Fig. 10, d). The explanation arises from the fact that following a link is more likely to lead to a highly connected node than picking a node randomly. Because in T1 both of the nodes gaining a link in the triad formation step are chosen by following a link, high degree nodes obtain more additional links than when the T2 mechanism is used, in which one of the nodes is chosen randomly. The shape of the degree distribution is the same in both cases, however, so it is not likely to have a qualitative effect on processes taking place on the network. The choice of T1 or T2 does not seem to have an effect on the number or size of cliques generated, nor on degree-degree correlations.

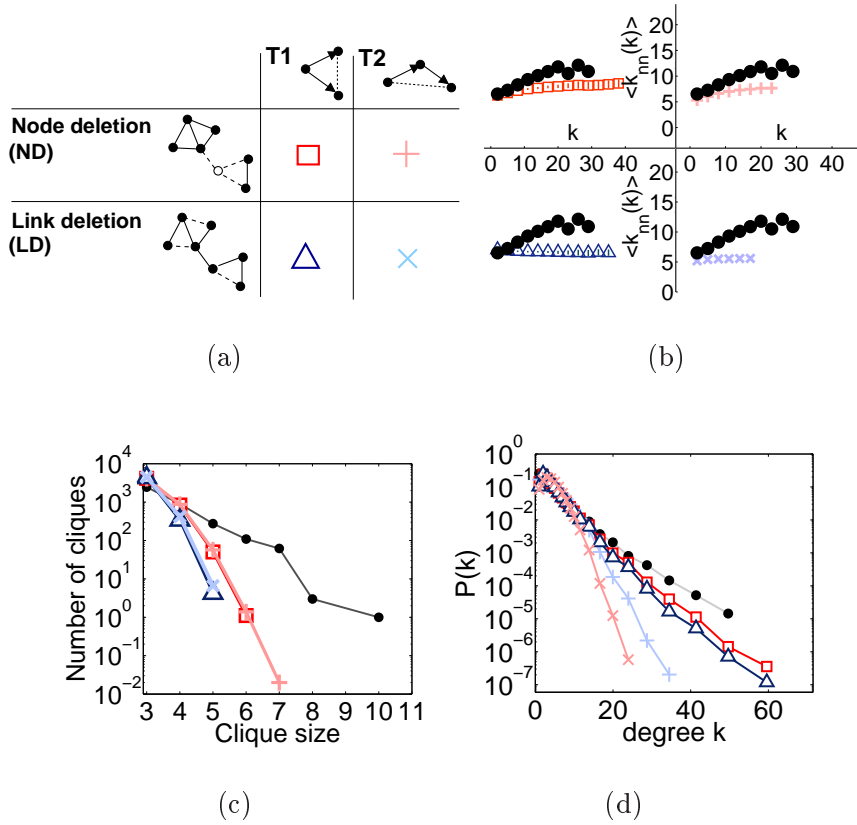


Fig. 10. Comparing different mechanisms employed in dynamical stochastic algorithmic models. a) Two ways of triad formation (T1 and T2) are combined with two ways of deleting links (node deletion refers to deleting all links of a node, and link deletion refers to deleting randomly selected links). b) Average nearest neighbor degree  $\langle k_{nn} \rangle$  with respect to node degree  $k$ , variants arranged as in the schematic figure. The *lastfm* data is also shown in each panel. The model variants using T1 show a clearly assortative relation, suitable for social network models, whereas the T2 networks are disassortative or very weakly assortative. c) Cliques. Model variants using node deletion produce more and larger cliques than those using link deletion, but all of these very simple models fall short of the data. d) Degree distribution. Model variants using T1 produce nodes with higher connectivities than those using T2.

## 5 Discussion

We have studied a set of stochastic algorithmic (SA) models for social networks belonging to two basic categories, one based on intrinsic features of the nodes and homophily (spatial models, SM) and the other on local network topology (topological models, TM). In the category of topological models, we focus on models that use mechanisms of triadic and focal closure (TFC). No categorization could neatly label all existing models, but the distinction we made was useful for the purposes of the comparison. We selected representa-

tive models of each category, and compared the various models with the aim of assessing their ability to reproduce essential features of real world social networks. We determined for each model a set of parameter values that produced the best match to the data, and observed the networks obtained using those parameters. The particular selection of data sets to which we fit the models is not of crucial importance, because they resemble other social networks in the features that we are interested in: they have high clustering, degree distributions that imply the presence of high degree nodes, positive degree-degree correlations, a decreasing clustering spectrum, and clustered structure.

Due to the different numbers of parameters in the models, and their different adaptability, some models could match more data features than others. Naturally, the models with only two parameters allowed only two average quantities to be fitted. Therefore, once largest component size and average degree were matched, we had no control over clustering, assortativity, path lengths, or any other network features. In general, the more parameters, the more flexibility in fitting, but the mechanisms employed in the models greatly affect adaptability. Having many parameters did not ensure that particular network features could be simultaneously obtained. One could argue that the selection of a mechanism is in essence a parameter as well.

The homophily principle employed in the spatial models was seen to be sufficient for producing strong positive degree-degree correlations. When link probability is based on distance, high connectors arise in locations with a dense population of nodes, so that many of their neighbors will also be high connectors. The spatial models also produced a large number of cliques, and consisted of clusters loosely connected with low overlap links. Their clustered structure was more pronounced than in the data. Spatial models in which the nodes are located with uniform probability in the underlying social space and links are based solely on homophily, produce a clustering spectrum  $c(k)$  strikingly different from observed data, indicating that it is not a sufficient description of the mechanisms at play in the formation of social networks. The spatial models based on homogeneously distributed node characteristics also produce peaked degree distributions without very high degree nodes which, based on our data and previously published data, does not seem appropriate for large social networks.

The topological models based on triadic and focal closure (TFC) produced reasonably realistic degree distributions and clustering with respect to degree. The dynamical TFC did not produce very high assortativity, unlike the spatial models and the growing TFC, although we need to keep in mind that high assortativity was specifically attempted only in fitting the KOSKK model. Many of the TFC models produced a relatively small number of cliques, fewer than were observed in the data. Most of the TFC models, in accordance with the data, contained a large core that did not break apart when low overlap

links were removed. Only the KOSKK model broke down with removal of low overlap links. With respect to thresholding by overlap, the KOSKK model seemed to display the clearest clustered structure of all the TFC models. The weights employed in tie formation in the KOSKK model apparently play an important role in the formation of community structure in TFC models, as the authors observed (Kumpula et al., 2007). The TFC models also had a large fraction of their nodes and zero overlap links residing in leaves or chains, as in the data, whereas in the spatial models the zero overlap links mostly resided between clusters.

In order to observe the effects of the different mechanisms employed in the dynamical TFC models in a more controlled way, we combined them in four simple models and compared the outcomes. The triad formation mechanism in which two neighbors of a node were linked (T1), was observed to generate higher degree nodes than the mechanism linking a node to a neighbor of a neighbor (T2). Deleting links in connection with an individual, as opposed to deleting random links, increased assortativity and the density of cliques in the network, both desirable features for social networks.

Many aspects of social networks were attained even with very simple mechanisms. However, neither the spatial models based on homophily, nor the topological models based on triadic and focal closure, were able to reproduce all important features of social networks. It seems obvious that any model for large scale social networks ought to take into account geographical factors, and perhaps homophily based on social characteristics as well, but that the role of triadic and focal closure in the formation of links should not be neglected. As both mechanisms obviously are present in the evolution social networks, a combination of the model types could yield more realistic networks.

## **Acknowledgements**

We acknowledge the Academy of Finland, the Finnish Center of Excellence Program 2006-2011, Project No. 213470. R.T. is supported by the ComMIT graduate school.

## A Appendix

### A.1 Basic network measures

The network representation of social contacts consists of *nodes* representing the individuals, and *links* representing the ties between them. We denote by  $N$  the number of nodes in a network, i.e. network *size*. A component of a network is a connected subset of nodes. In this paper, we study the *largest component* LC of each network. Unless the network is very fragmented, the largest component contains a large fraction of the nodes. We denote the size of the largest component by  $N_{LC}$ .

The number of network neighbors of a node is called its *degree*  $k$ . An isolated node has degree zero. The brackets  $\langle \rangle$  denote averaging over all nodes (or links) within the network. Averaging over several networks is not explicitly denoted in the plots and tables in this paper, but they always contain values averaged over 100 network realizations.

A measure of local triangle density, the *clustering coefficient*, describes the extent to which the neighbors of node  $i$  are acquainted with one another: if none of them know each other,  $c_i$  is zero, while if all of them are acquainted,  $c_i = 1$ . For a node  $i$  with degree  $k_i$  and belonging to  $T_i$  triangles, the clustering coefficient is defined as

$$c_i = \frac{T_i}{k_i(k_i - 1)/2}, \quad (\text{A.1})$$

where the denominator  $k_i(k_i - 1)/2$  expresses the maximum possible number of triangles  $i$  could belong to given its degree. The clustering coefficient is not defined for nodes with degree  $k < 2$ . The average clustering coefficient, averaged over all nodes with  $k \geq 2$  in the network, is denoted  $\langle c \rangle$ .  $c(k)$  denotes the average clustering coefficient of nodes having degree  $k$ . The curve  $c(k)$  is called the *clustering spectrum*.

It has been observed that social networks are typically assortative, 'popular people know other popular people'. One way of quantifying this effect is using linear correlation, or the Pearson correlation coefficient, between the degrees  $k_i$  and  $k_j$  of pairs of connected nodes, also called the *assortativity coefficient*  $r$  (Newman, 2002):

$$r = \frac{\sum_e k_i k_j / E - [\sum_e \frac{1}{2}(k_i + k_j)]^2 / E^2}{\sum_e \frac{1}{2}(k_i^2 + k_j^2) / E - [\sum_e \frac{1}{2}(k_i + k_j)]^2 / E^2},$$

where  $E$  is the total number of links in the network. A positive value of the assortativity coefficient signifies that the nodes with a large number of ties are connected to one another more likely than would be expected by chance, and nodes with a small number of ties are connected more likely with one another. A negative value signifies that mostly nodes with small degree are connected to the large connectors, which are not directly linked between themselves. Of course, this measure only captures the linear part of the correlation. Assortativity can also be quantified using the

measure *average nearest neighbor degree*  $\langle k_{nn}(k) \rangle$ , found by taking all nodes with degree  $k$ , and averaging the degrees of their neighbors. If the curve  $\langle k_{nn}(k) \rangle$  plotted against  $k$  has a positive trend, nodes with high degree typically also have high-degree neighbors, hence the network is assortative.

The shortest path length  $l_{ij}$  between nodes  $i$  and  $j$  in a network means the minimum number of links that need to be traversed in order to get from  $i$  to  $j$ . The average length  $\langle l \rangle$  of shortest paths between nodes describes the compactness of the network.

## A.2 Determining optimal network parameters

We attempt to find the best network parameters (optimal points in the *parameter space*) which produce network realizations whose features match our data as closely as possible. We attempt to minimize the relative error in each chosen feature. For the average degree  $\langle k \rangle$  in a model with given parameters  $\mathbf{p}$ , for example, being fitted to a data set with average degree  $\langle k \rangle^{target}$ , the relative error would be

$$|\epsilon_{\langle k \rangle(\mathbf{p})}| = \left| \frac{\langle k \rangle(\mathbf{p}) - \langle k \rangle^{target}}{\langle k \rangle^{target}} \right|. \quad (\text{A.2})$$

The errors for each feature are combined in the *error function*  $f$ , whose norm  $|f|$  is then minimized. For example, when fitting to  $N_{LC}$ ,  $\langle k \rangle$  and  $\langle c \rangle$ , the error function would take the shape

$$f(\mathbf{p}) = [w_{N_{LC}}\epsilon_{N_{LC}} \quad w_{\langle k \rangle}\epsilon_{\langle k \rangle} \quad w_{\langle c \rangle}\epsilon_{\langle c \rangle}], \quad (\text{A.3})$$

and its norm, which we attempt to minimize, would be

$$|f(\mathbf{p})| = \sqrt{w_{N_{LC}}^2\epsilon_{N_{LC}}^2 + w_{\langle k \rangle}^2\epsilon_{\langle k \rangle}^2 + w_{\langle c \rangle}^2\epsilon_{\langle c \rangle}^2}. \quad (\text{A.4})$$

Whenever the model is able to match the target values closely, the result is not sensitive to the weights  $w$ . If all values cannot be exactly matched despite a large enough number of parameters, the weights determine which parameters are matched more closely. In such a case (WPR, *lastfm* fit), we put more emphasis on matching the number of nodes and links, less on clustering and least on average shortest path lengths. The values of the weights matter only in relation to one another, and hence the weights can be scaled such that  $w_{N_{LC}}^2 + w_{\langle k \rangle}^2 + w_{\langle c \rangle}^2 = 1$ . The error function should have equally many components as there are network parameters.

We attempt to find the network parameters  $\mathbf{p}$  that produce the minimum value of the norm (A.4). It is important to know how the error function depends on the parameters - is it smooth, does it have many local minima, etc. Visualization helps to make sure that the function is smooth enough and that the found optimum is a true optimum and not just a local minimum or due to random fluctuation. The simple graphical method of visualization could also be used for locating the optimal parameter values. However, as it takes a lot of computation time to evaluate the

function at a large number of points, most of which are far from the optimum, it is not an efficient optimization method.

Several more sophisticated methods are available for optimizing stochastic functions. One of the methods we used consists of using a linear approximation for the components of the error function, and iteratively refining the approximation close to the optimum. We also used the well established *Nelder-Mead method* (Nelder and Mead, 1965), which involves calculating values of the error function at the corners of a simplex (a triangle in 2-dimensional space, a tetrahedron in 3D). The optimal value of the error function is iteratively approached by rolling one corner of the simplex over the others such that the object moves towards the region where the function gets optimal values. The diameter of the simplex is adjusted during iteration to increase accuracy. Sometimes analytical estimates derived by the authors of the models could be used as a starting point in optimization (BPDA, Váz). These initial estimates were then refined by optimization.

## References

- Amaral, L.A.N, Scala, A., Barthélémy, M., and Stanley, H.E., 2000. Classes of small-world networks. *Proc. Natl. Acad. Sci. (USA)* 97 (21) 11149-11152.
- Boguñá, M., Pastor-Satorras, R., 2002. Epidemic spreading in complex correlated networks. *Phys. Rev. E* 66, 047104.
- Boguñá, M., Pastor-Satorras, R., Díaz-Guilera, A., and Arenas, A., 2004. Models of social networks based on social distance attachment. *Phys. Rev. E* 70, 056122.
- Castellano, C., Fortunato, S., and Loreto, V., 2007. Statistical physics of social dynamics, arXiv:0710.3256.
- X. Castelló, X., Toivonen, R., Eguíluz, V.M., Saramäki, J. Kaski, K., and San Miguel, M., 2007. Anomalous lifetime distributions and topological traps in ordering dynamics. *EPL* 79, 66006.
- Davidson, J., Ebel, H., and Bornholdt, S., 2002. Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks. *Phys. Rev. Lett.* 88 (12) 128701-1.
- Gleiser, P. and Danon, L, 2003. Community Structure in Jazz. *Advances in Complex Systems*, Vol. 6, No. 4 (2003) 565-573
- Goodreau, S. M., 2007. Advances in exponential random graph ( $p^*$ ) models applied to a large social network. *Social Networks* 29, 231-248.
- Granovetter, M., 1973. The Strength of Weak Ties. *Am. J. Soc.*, 78, 1360-1380.
- Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., and Arenas, A., 2003. Self-similar community structure in a network of human interactions, *Physical Review E*, vol. 68, 065103.
- Jin, E.M., Girvan, M., and Newman, M.E.J., 2001. Structure of growing social networks, *PRE* 64, 046132.

- Kossinets, G. and Watts, D.J., 2006. Empirical Analysis of an Evolving Social Network. *Science* 311, 88-90.
- Kumpula, J., Onnela, J.P., Saramäki, J., Kaski, K., and Kertész, J., 2007. Emergence of communities in weighted networks. *Phys. Rev. Lett.* 99, 228701.
- Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P., 2008. A gravity model for the geographical dispersal of mobile communication networks. arXiv:0802.2178v1.
- Leskovec, J., and Horvitz, E., 2008. Planetary-Scale Views on an Instant-Messaging Network, arXiv:0803.0939.
- Leskovec, J., Adamic, L. A., and Huberman, B. A., 2007. The dynamics of viral marketing. *ACM Transactions on the Web*, 1, 1.
- Marsili, M., Vega-Redondo, F., and Slanina, F., 2004. The rise and fall of a networked society: A formal model. *PNAS* 101 (6), 1439-1442
- Masuda, N. and Konno, N., 2006. VIP-club phenomenon: Emergence of elites and masterminds in social networks. *Social Networks* 28, 297-309.
- McPherson, M., Smith-Lovin, L., and Cook, J.M., 2001. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.* 27, 415-444.
- Moreno, Y., Nekovee, M., and Pacheco, A. F., 2004. Dynamics of rumor spreading in complex networks. *Phys. Rev. E* 69, 066130.
- Nelder, J.A. and Mead, R., 1965. A simplex algorithm for function minimization. *The Computer Journal* 7, pp. 308-313.
- Newman, M.E.J., 2002. Assortative Mixing in Networks *Phys. Rev. Lett.* 89(20):208701
- Onnela, J.-P., Saramäki, J., Hyvönen, Szabó, G., Argollo de Menezes, M., Kaski, K., Barabási, A.-L., and Kertész, J., 2007. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics* 9, 179.
- Onnela, J.-P., Saramäki, J., Hyvönen, Szabó, G., Lazer, D., Kaski, K., Kertész, K., and Barabási, A.-L., 2007. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. (USA)* 104, 7332
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D., 2007. An introduction to exponential random graph ( $p^*$ ) models for social networks. *Social Networks* 29, 173-191.
- Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J., and Kaski, K., 2006. A model for social networks. *Physica A* 371 (2) 851-860
- Vázquez, A., 2003. Growing networks with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E* 67, 056104
- Wong, L.H., Pattison, P. and Robins, G., 2005. A spatial model for social networks. *Physica A* 360, 99-120.