Emergence of Clustering in an Acquaintance Model without Homophily

Uttam Bhat

Department of Physics, Boston University, Boston, MA 02215, USA
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA

P. L. Krapivsky

Department of Physics, Boston University, Boston, MA 02215, USA

S. Redner

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA
Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA

Abstract. We introduce an agent-based acquaintance model in which social links are created by processes in which there is no explicit homophily. In spite of this constraint, highly-clustered social networks can arise. The crucial feature of our model is that of variable transitive interactions. That is, when an agent introduces two unconnected friends, the rate at which a connection actually occurs between them is controllable. As this transitive interaction rate is varied, the social network undergoes a dramatic clustering transition and the network consists of a collection of well-defined communities close to the transition. As a function of time, the network can undergo an *incomplete* gelation transition, in which the gel, or giant cluster, does not constitute the entire network, even at infinite time. Some of the clustering properties of our model also arise, albeit less dramatically, in Facebook networks.

PACS numbers: 87.23.Ge, 05.65.+b, 89.75.Hc, 89.75.Da

1. Introduction to Acquaintance Modeling

An important feature of many complex networks is that they can be highly clustered. That is, such networks are comprised of well-connected modules, or communities, with weaker connections between them (see, e.g., [1–9]). Part of the motivation for focusing on communities is that unraveling this substructure may provide important clues about how such networks are organized, how they function, and how information is transmitted across them. While identifying communities has become a standard diagnostic of networks [10–13], and there has been much recent effort devoted to determine the community structure of complex networks, less is known about mechanisms that could lead to this heterogeneity; for contributions in this direction, see, e.g., [14–18]. Our goal is to develop a basic model for the formation of a social network in which highly-clustered substructures emerge spontaneously from homogeneous social interaction rules. We do not need to appeal to homophily (see, e.g., [19, 20]) or some other explicit source of heterogeneity to generate large-scale clustering.

In our modeling, the starting network consists of N complete strangers with no links between them. This might describe, for example, a set of entering students to a university in an unfamiliar location. We assume that the population remains constant over the time scale that social connections form. There are two distinct ways that connections are made:

- *Direct Linking:* An agent with either zero friends or one friend links to a randomly selected agent.
- *Transitive Linking:* An agent with two or more friends introduces two of them at random. These selected agents then create a link with a rate that is specified below.

These two mechanisms underlie the acquaintance model that was introduced by Davidsen et al. [21]; for related work see [22,23] and references therein. In [21], the rates of these two linking processes were fixed and a steady state was achieved by allowing any agent and all its attached links to disappear at a (small) rate and correspondingly adding a new agent to the network to keep the number of agents fixed. In this work, we impose a different, socially-motivated, mechanism that allows the network to reach a non-trivial long-time state.

A natural motivation for distinguishing between direct and transitive linking is that an individual with many friends typically has less impetus to initiate additional connections. Indeed, it has been previously noted that there seems to be a cognitively-limited upper limit—the Dunbar number of the order of a hundred—for the number of meaningful friends that any individual can sustain [24]. The threshold criterion for direct linking as defined above represents an extreme limit where an agent ceases to initiate new connections once he has two friends. Nevertheless, a popular person is still likely to make additional connections as a result of being introduced to someone new. It is worth emphasizing that the type of transitive linking employed in this work plays an essential role in many social networks, ranging from Granovetter's picture of the "strength of weak ties" [25] to Facebook, where users are invited to link to the friends of their Facebook friends [26–28].

The key feature of our acquaintance model is the imposition of distinct rates for direct and transitive linking that are determined by the current state of the network. There are two different mechanisms that we implement to control these rates:

(i) *Threshold-Controlled:* When two agents, α and β, are introduced by a common friend i, they connect if the number of their mutual friends $m_{\alpha\beta}$ (inside the oval in Fig. 1) to the total number of friends of either agent (the degrees k_α and k_β) equals or exceeds a specified friendship threshold F. That is, a connection occurs between α and β if $m_{\alpha\beta}/d_{\alpha\beta} > F$, where $d_{\alpha\beta} = \min(k_\alpha, k_\beta)$.

(ii) *Rate-Controlled:* The rates of transitive and direct linking are defined as R and 1 respectively.

The use of threshold-controlled transitive linking is motivated by the observation that one is more likely to become friends with a newly introduced person when the two of you already have many common friends. The degree of commonality can be a useful indicator how much two people have in common.

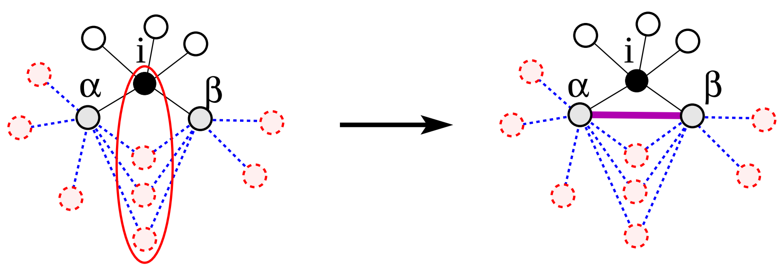
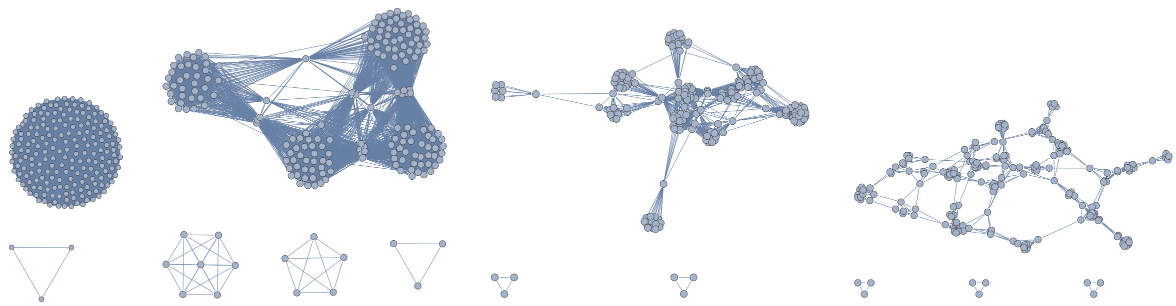


Figure 1. (color online) Illustration of threshold-controlled transitive linking. An agent i (solid) with five friends (open circles) is selected. The selected agent introduces two of them, α and β (shaded). These become friends (thick line) if the ratio of their mutual acquaintances (inside the oval) to the total acquaintances of either α or β (dashed circles) exceeds a specified threshold. Links outside this cluster are not shown.

In each update step of our friendship model, an agent is selected at random. If the degree of this agent equals 0 or 1, the agent links to another randomly selected agent. If the degree of the initial agent is 2 or larger, a transitive link is created between two friends of the agent according to the rates given above. Notice that direct linking joins two clusters (here clusters are defined as the maximal disconnected components of a graph), while transitive linking merely "fills in" links within a cluster without altering its size. Updates continue until no more links can be created. Also notice that once every agent has at least two friends, cluster mergings no longer occur and links can be created only within a cluster. In threshold-controlled linking, the network reaches its final state when all the friends of any agent are either linked to each other or can no longer fulfill the threshold condition. In rate-controlled linking, the final state consists of a collection of complete subgraphs for any $R > 0$; henceforth, we term a complete subgraph of a network as a *clique*. The case $R = 0$ is unique, as will be discussed below. In either case, a final state is reached because geometrical constraints ultimately prevent the formation of additional links.

2. Threshold-Controlled Transitive Linking

The most prominent feature of threshold-controlled transitive linking is the emergence of highly-clustered substructures over a wide range of threshold value (Fig. 2).



(a) (b) (c) (d)

Figure 2. Typical networks of $N = 200$ agents for threshold values: (a) $F = 0.3$, (b) 0.35, (c) 0.4, and (d) 0.6.

For small F, a new friendship is created nearly every time two individuals are introduced by a mutual friend. In this regime, the resulting graph is nearly complete, but there also exists a small insular "fringe" population that is comprised of small disjoint cliques (Fig. 2). This fringe arises because once every agent in a cluster has at least two links, there is no mechanism for this cluster to merge with any other cluster. Thus even in the limiting case of $F = 0$, the final state typically consists of more than a single cluster, each of which is complete. Concomitantly, the largest cluster does not constitute the entire system for any value of F. For intermediate values of F, the networks in Figs. 2(b) and 2(c) are visually highly clustered, with the largest cluster comprised of a small number of well-connected modules.

To help quantify the network and its clustering, we study the dependences of the average degree and the distribution of community sizes as a function of the threshold F. The average degree exhibits a sharp change between a dense and a sparse regime for F in the range of 0.2–0.3 for network sizes between 1000 and 16000 (Fig. 3(a)). We define the location of the transition as the point where $\langle k \rangle / N$, the average degree divided by network size, equals $\frac{1}{2}$. According to this definition, the critical threshold value F_c decreases very slowly with N (Fig. 3(b)). From the data alone, it is not evident whether a transition exists at non-zero F_c for $N \to \infty$ or whether this transition is a very slow finite-size effect.

A more direct way to understand how the network structure depends on F is by studying the community-size distribution. Unlike real clusters, which are unambiguously defined, the notion of community is somewhat fuzzy [29]. Intuitively, a community is a

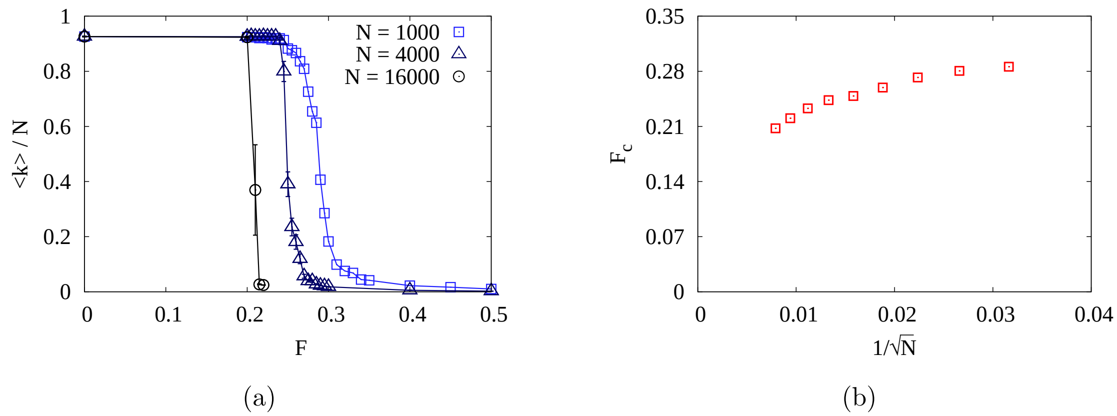


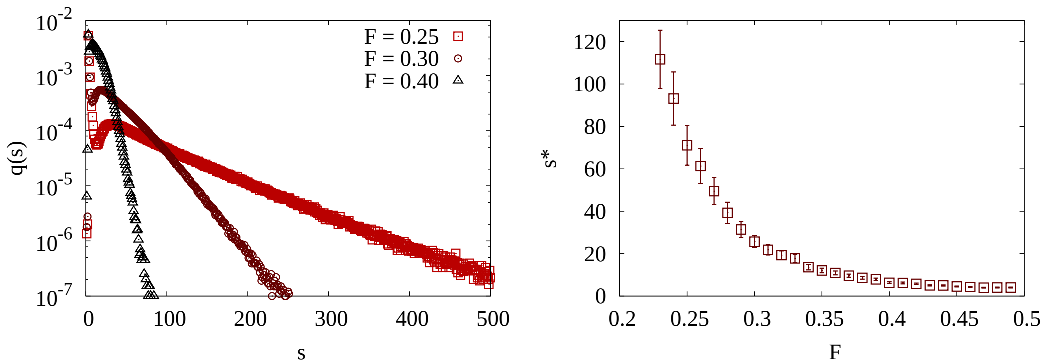
Figure 3. (a) Average degree divided by N versus threshold. (b) Critical threshold versus $1/\sqrt{N}$.

community is a group of nodes that is densely interconnected and is sparsely connected to nodes external to the community. We adopt the definition given in Refs. [2, 3] in which the communities of a given network are defined by the partition that maximizes the modularity Q, defined by

$$Q = \sum_{uv} \left[\left(\frac{A_{uv}}{2L} \right) - \left(\frac{k_u}{2L} \right) \left(\frac{k_v}{2L} \right) \right] \delta \left(c_u, c_v \right) \; . \tag{1}$$

Here A_{uv} is the adjacency matrix, with $A_{uv} = 1$ if a link exists between nodes u and v and $A_{uv} = 0$ is no such link exists, k_u, k_v are the degrees of these nodes, L is the number of network links, and c_u, c_v label the communities that contain nodes u and v. The first term in Eq. (1) is the fraction of links within all communities. The second term gives the fractions of links that would exist within communities if all connections were randomly rewired subject to the constraint that all node degrees are preserved. Thus the modularity is the fraction of links within communities minus the fraction of links that would exist within communities by chance.

To find the maximizing partition into communities, we apply the Monte Carlo algorithm proposed by Blondel et al. [7]. In this algorithm, the network initially has the N nodes in N isolated communities. Then for each node i, we consider each of its neighbors j and evaluate the gain of modularity that would occur by placing node i in the community of node j. The node i is ultimately included in the community for which this gain is the largest (and positive). Once this step of assigning individual nodes to communities is done, the same fusion process is implemented on the current network of communities. This fusion of higher-order communities is repeated until no further gain in modularity is possible. At this point, the algorithm gives the community size distribution, $q(s)$. As shown in Fig. 4(a), the tail of this distribution, averaged over many realizations, decays as e^{-s/s^*}, with the characteristic community size s^* increasing as the threshold is decreased. The data also suggests that s^* diverges as F approaches F_c from above (Fig. 4(b)). For $F < F_c$, communities are all cliques, among which the



(a) (b)

Figure 4. (a) Community size distributions for various thresholds F for networks of 10^4 nodes averaged over 10^4 realizations each. (b) The slope of the exponential tail gives the characteristic community size.

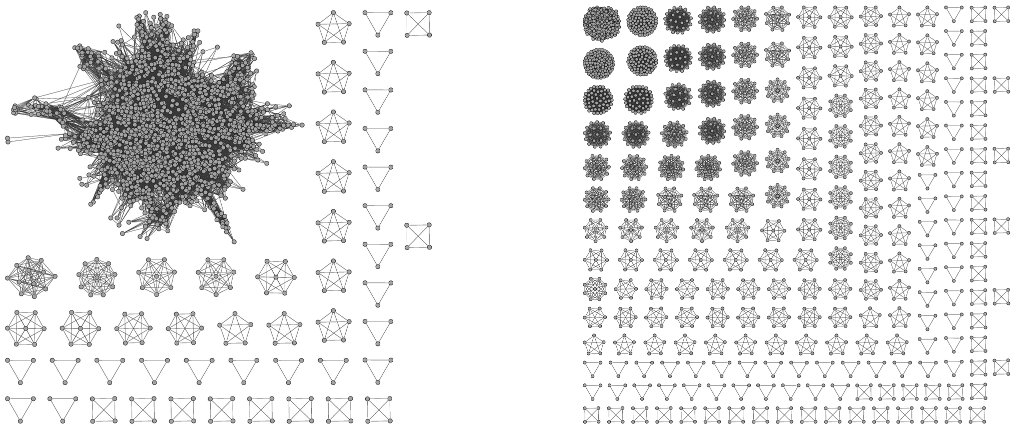
largest has a size that scales linearly with N. This apparent gelation phenomenon is best understood by investigating the rate-controlled version of our friendship model, to be discussed in the next section.

3. Rate-Controlled Transitive Linking

While the threshold-controlled friendship model leads to networks with visually striking community structure, many geometric and time-dependent properties are more readily understood within the framework of rate-controlled transitive linking. As outlined in section 1, starting with an initial state of N isolated nodes, nodes of degree 0 or 1 join to any other node in the network at rate 1, while a link between two mutual friends of a node of degree 2 or greater occurs with rate R. The process ends when the network is partitioned into a set of cliques. Figure 5 shows typical networks of $N = 2000$ agents at the instant when no nodes of degree 0 or 1 remain for: (a) $R = 2$, where the cluster-size distribution decays exponentially with size, and (b) $R \approx 15$, where the distribution has a power-law decay.

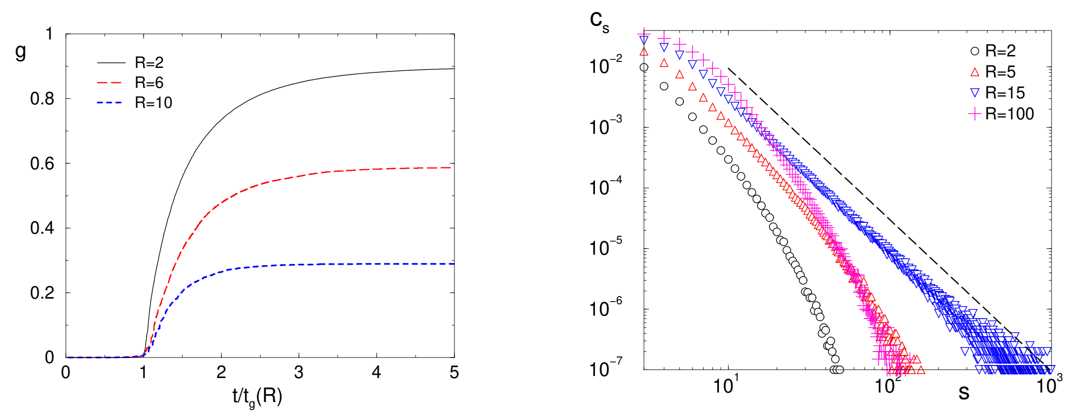
In the range $R < R_c \approx 15$, rate-controlled transitive linking leads to the emergence of a macroscopic cluster at a finite gelation time. However, this gelation phenomenon is *incomplete*, because the fraction of agents within this macroscopic cluster (also known as the gel fraction) saturates to a value that is *strictly* less than one as $t \to \infty$ (Fig. 6(a)). Thus in addition to the single macroscopic cluster, many small clusters persist forever. This behavior strikingly contrasts with classical gelation, where the gel encompasses the entire system when the reaction runs to completion [30, 31].

The incompleteness of the gelation transition arises because the reaction is controlled by *active nodes*—those of degree 0 and degree 1. When all these nodes have been used up by linking to other nodes, there is no possibility for additional cluster mergings. All that can occur is densification within each cluster. Thus if multiple



(a) (b)

Figure 5. Clusters at the instant when no nodes of degree 0 or 1 remain for a network of 2000 nodes for: (a) $R = 2$ (exponential cluster-size distribution) and (b) $R = 15$ (power-law distribution).



(a) (b)

Figure 6. Dependence of: (a) the fraction of agents in the largest cluster as a function of the normalized time $t/t_g(R)$ and (b) the cluster-size distribution c_s versus size s on a double logarithmic scale at infinite time, both for representative values of R values. Here $t_g(R)$ is the gelation time for a given R. In (b), linear behavior occurs only for $R = 15$. The dashed line has slope $-\frac{5}{2}$.

clusters happen to exist when active nodes are exhausted, these clusters will persist forever. In spite of this incompleteness feature, the gelation transition itself seems to conform to the classical mean-field description. As t approaches the gelation time $t_g(R)$ from below, the concentration of clusters of size s, c_s gradually broadens and changes from an exponential decay as a function of s to an algebraic decay (Fig. 6(b)). At the gelation time, $c_s \sim s^{-\alpha}$, with $\alpha \approx \frac{5}{2}$, as in classical gelation. The gelation time itself diverges for $R \geq R_c \approx 15$. In the regime where $R > R_c$, transitive linking events quickly use up all active nodes, which are the catalysts for cluster merging. Because the average cluster size is still small at the instant when active nodes are used up, gelation is suppressed for $R > R_c$. In this non-gelling regime, the cluster-size distribution decays exponentially with size at all times.

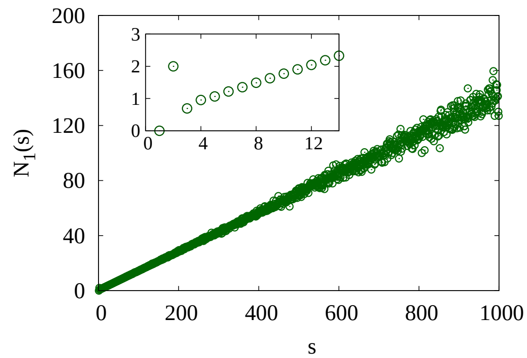


Figure 7. Average number of active nodes within clusters of size s at fixed time $t = 0.98$ for network of 10^6 nodes and for the case of $R = 2$. The data represents an average over 100 realizations. The inset shows the small-s behavior.

We can make a more quantitative correspondence between rate-controlled transitive linking and gelation by mapping the former onto a version of classical product kernel aggregation [30, 31]. This correspondence relies on our observation from simulations that the number of active nodes in a cluster of size s at any given time is proportional to s (Fig. 7). The proportionality constant is time dependent because the concentration of nodes of degree one (which we term leaf nodes) nodes change with time. As shown in the inset to Fig. 7, the contribution of clusters of size 1 and 2 deviates from the overall linear trend in the main figure. In particular, a monomer (a node of degree 0) has no leaf nodes while a dimer always has two leaf nodes (each of degree 1). For clusters of size $s > 2$, we assume that the average number of leaf nodes is given by $N_1(s,t) = \lambda(t)\, s$. Then the total number of leaf nodes, n_1, can be expressed through the cluster-size distribution: $n_1 = 0 \cdot c_1 + 2c_2 + \sum_{k\geq 3} \lambda\, k\, c_k$. Combining this equation with the conservation law $\sum_{k\geq 1} kc_k = 1$ we get $n_1 = 2c_2 + \lambda(1 - c_1 - 2c_2)$, from which

$$\lambda(t) = \frac{n_1 - 2c_2}{1 - c_1 - 2c_2} \tag{2}$$

We now now write the following product-kernel-like aggregation equations for the cluster-size distribution [30, 31], in which we separately account for the evolution of monomers and dimers:

$$\begin{aligned}
\dot{c}_1 &= -c_1 - (c_1 + n_1)c_1\,, \\
\dot{c}_2 &= c_1^2 - 2c_2 - 2c_1c_2 - 2n_1c_2\,, \\
\dot{c}_k &= c_1\big[c_{k-1}(k-1) - kc_k\big] + 2c_2\big[(k-2)c_{k-2} - kc_k\big] \\
&\quad + \sum_{j\geq 3} \lambda\, j\,(k-j)\, c_j c_{k-j} - \sum_{j\geq 3} \lambda\, j\, k\, c_j c_k - \lambda\, k\, c_k\,, \qquad k \geq 3,
\end{aligned} \tag{3}$$

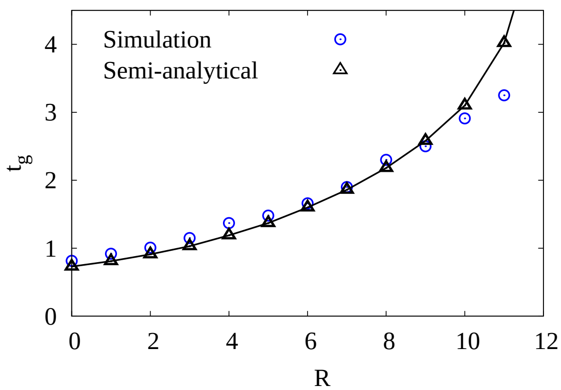


Figure 8. Comparison of gelation times, t_g obtained by semi-analytical calculation of M_2 and t_g obtained directly from simulations

where the dot denotes the time derivative. From these equations, the second moment of the cluster-size distribution evolves according to

$$\begin{aligned}\dot{M}_2 &= 16c_2^2 + 2c_1c_2 - n_1(c_1 + 8c_2) + (2c_1 + 8c_2)M_2(1-\lambda) \\ &\quad + \lambda(2M_2^2 + c_1 + 8c_2 - c_1^2 - 16c_2^2 - 10c_1c_2)\,.\end{aligned} \tag{4}$$

If gelation does occur, the second moment M_2 would diverge at a gelation time $t_g(R)$. However, we are unable to write a closed equation for the concentration of leaf nodes n_1 and thereby solve for M_2 and t_g. Thus to find t_g, we take the value n_1 from simulations and use it solve the equations for $\dot{c}_1$, $\dot{c}_2$ and $\dot{M}_2$ numerically. This approach gives good agreement with the value of $t_g(R)$ obtained by direct simulations (Fig. 8), as long as R is not close to R_c. We can also find R_c from our semi-numerical method by scanning across different values of R and finding the value of R where $M_2(R, t=\infty) = \infty$ and $M_2(R + \Delta R, t = \infty) < \infty$. This approach gives a lower value for $R_c \approx 12.1$ compared to $R_c \approx 15$ directly from simulations and thus gives a sense of the accuracy of our semi-numerical method.

4. Extremal Limits of Transitive Linking

To develop additional insights, we now investigate our friendship model in the extremal limits of no transitive linking or infinitely rapid transitive linking. The former case may be achieved in the threshold model with $F = \infty$ or in the rate model with $R = 0$. The latter is achieved in the rate model by setting $R = \infty$. In these limiting cases we can obtain useful insights about some basic network properties by analytical means.

4.1. No Transitive Linking

In the absence of transitive linking, the network evolves by a constrained aggregation process that is mediated only by active nodes—those of degrees 0 or 1 (Fig. 9). The network stops evolving when these active nodes no longer exist. By enumerating all the possible ways that an active node can interact (Fig. 9), the concentrations n_k of agents of degree k evolve according to

$$\begin{aligned}\dot{n}_0 &= -n_0(1+a)\\ \dot{n}_1 &= (n_0 - n_1)(1+a)\\ \dot{n}_2 &= n_1 + (n_1 - n_2)a\\ \dot{n}_k &= (n_{k-1} - n_k)a, \qquad k \geq 3\,.\end{aligned} \tag{5}$$

Here $a = n_0 + n_1$ is the concentration of active agents. One can verify that the conservation law $\sum_{k\geq 0} \dot{n}_k = 0$ is obeyed and that the mean degree grows according to $\frac{d}{dt}\langle k\rangle = \sum_{k\geq 0} k\dot{n}_k = 2a$.

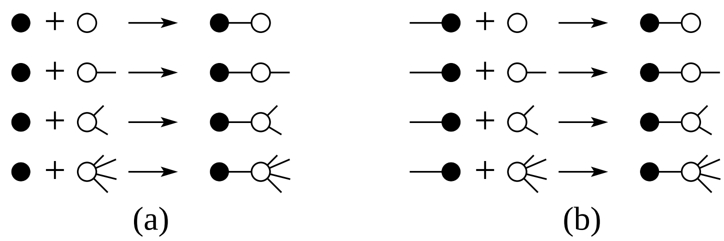


Figure 9. The elemental evolution steps without transitive linking: (a) reaction channels for a node of degree 0 and (b) for a node of degree 1. The solid circle denotes the initial node.

The rate equations (5) admit an exact, albeit implicit, solution. To simplify the first two lines of Eqs. (5), we introduce the time-like variable $dT = dt(1+a)$ to recast these equations as $n_0' = -n_0$ and $n_1' = n_0 - n_1$, where the prime denotes differentiation with respect to T. The solution is

$$n_0 = e^{-T}\,, \qquad\qquad n_1 = Te^{-T}\,. \tag{6}$$

Consequently the original and modified time variables are related by

$$t = \int_0^T \frac{dT'}{1 + (1+T')\,e^{-T'}} \tag{7}$$

Equations (6)–(7) provide the exact, but implicit solution for the densities of agents with degree 0 and degree 1.

To obtain more explicit results, we need to relate t and T. To this end, we write

$$\begin{aligned}T - t &= \int_0^T dx\left[1 - \frac{1}{1+(1+x)e^{-x}}\right]\\ &= \int_0^\infty dx\left[\frac{1+x}{1+x+e^x}\right] - \int_T^\infty dx\left[\frac{1+x}{1+x+e^x}\right]\\ &\equiv \alpha - \mathcal{O}(T\,e^{-T}) = \alpha - \mathcal{O}(t\,e^{-t})\,,\end{aligned} \tag{8}$$

with the value of α determined numerically to be $1.2802837\ldots$. Thus the densities of active agents asymptotically vary as

$$n_0 \to e^{-\alpha}\, e^{-t}, \qquad\qquad n_1 \to e^{-\alpha}\, t\, e^{-t}\,. \tag{9}$$

We estimate the time when active agents are exhausted by the criterion $n_1(t^*) = 1/N$; namely, a single node of degree 1 remains in a network of N nodes at the completion time t^*. From the above asymptotic dependence of n_1, the completion time is given by $t^* \simeq \ln N + \ln\ln N$.

To determine the density of agents with two or more friends, we define a second time-like variable $d\tau = a\, dt$ to recast the last line of Eq. (5) as

$$\frac{dn_k}{d\tau} = n_{k-1} - n_k \qquad k \geq 3\,. \tag{10}$$

Assuming that we know n_2, we use the Laplace transform method to solve (10) and then invert the Laplace transform to give the recursive solution

$$n_k(\tau) = \frac{1}{(k-3)!}\int_0^\tau d\tau'\,(\tau-\tau')^{k-3} e^{\tau'-\tau}\, n_2(\tau') \tag{11}$$

for $k \geq 3$. To determine n_2 we rewrite its evolution equation, the third line of (5), as

$$\frac{dn_2}{dT} + n_2\,\frac{d\tau}{dT} = n_1\,. \tag{12}$$

Integrating this equation and making use of $\tau = T - t$ gives the implicit solution

$$n_2(T) = e^{-\tau}\int_0^T dT'\, T' e^{-t(T')}\,. \tag{13}$$

In the $t \to \infty$ limit, the density of nodes of degree 2 is

$$n_2(\infty) = e^{-\alpha}\int_0^\infty dT\, T\, \exp\left[-\int_0^T \frac{dT'}{1+(1+T')e^{-T'}}\right] = 0.6018583\ldots$$

We now exploit this result to determine the density of nodes of arbitrary degree in the limit $t \to \infty$. First, notice that, by definition, the rescaled time $\tau = T - t$. Thus from Eq. (8), $\tau \to \alpha$ as $t \to \infty$, so that the fraction of nodes of degree $k > 2$ at infinite time is given by

$$n_k(\infty) = \frac{1}{(k-3)!}\int_0^\alpha d\tau\,(\alpha-\tau)^{k-3} e^{\tau-\alpha}\, n_2(\tau)\,. \tag{14}$$

The large-k asymptotic is simpler to determine, since the small τ limit of $n_2(\tau)$ makes the dominant contribution to the above integral. For early times, one gets $\tau \simeq t$ and $n_2 \simeq 2t^2$ as $t \to 0$. Substituting $n_2(\tau) \simeq 2\tau^2$ into (14) gives, for $k \gg 1$,

$$n_k(\infty) \simeq 4\, e^{-\alpha}\, \frac{\alpha^k}{k!}\,. \tag{15}$$

We may similarly obtain the average degree of the final network. We first rewrite the evolution equation for the mean degree as $\frac{d\langle k\rangle}{d\tau} = 2$, so that $\langle k\rangle = 2\tau$. The mean degree therefore starts at zero and increases to $2\alpha = 2.560567483\ldots$ as $t \to \infty$. Thus when transitive linking is not allowed, the final network is sparse and only slightly more dense than a tree. [For a tree of N nodes, the average degree is $2(1-\frac{1}{N})$.]

4.2. Infinite Transitive Linking Rate

The complementary situation where transitive linking is infinitely rapid is also tractable because each cluster is always a clique and may thus be fully characterized by its size. The update steps for this limiting case are summarized by (see Fig. 10):

(i) Select an active agent that connects to another agent in a cluster of size k. If the initial agent has degree 0, the cluster size increases from k to $k+1$, If the initial agent has degree 1, the cluster size increases to $k+2$.

(ii) After each growth event, all possible links within the enlarged cluster are immediately filled in so that the resulting cluster remains complete.

Once all agents of degree 0 or degree 1 are used up, the network has reached a final state that consists of a collection of cliques.

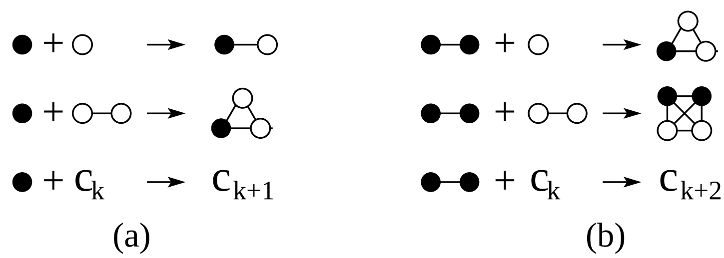


Figure 10. Elemental processes when transitive linking occurs with infinite rate for a node of degree: (a) 0 and (b) 1. The initial cluster is shown solid.

Using a rate equation approach, we can determine basic properties of this final clique-size distribution. From the reaction steps outline above, the concentrations of clusters of size k, c_k, evolve according to

$$\begin{aligned}
\dot{c}_1 &= -c_1 - c_1^2 - 2c_1c_2 \\
\dot{c}_2 &= c_1^2 - 2c_1c_2 - 2c_2 - 4c_2^2 \\
\dot{c}_k &= c_1\big[(k-1)c_{k-1} - kc_k\big] + 2c_2\big[(k-2)c_{k-2} - kc_k\big], \qquad k \geq 2\,.
\end{aligned} \tag{16}$$

Let us first determine the concentration of active clusters—monomers of size 1 or dimers of size 2. Keeping the two largest terms in equations for $\dot{c}_1$ and $\dot{c}_2$, one finds the following long-time behaviors for the monomers and dimer concentrations:

$$c_1 \simeq (2e^t - 1)^{-1} \to \tfrac{1}{2}\, e^{-t}, \qquad c_2 \simeq \tfrac{t}{4}\, e^{-2t}\,. \tag{17}$$

As one might anticipate, the concentrations c_1 and c_2 asymptotically decay exponentially with time. Since these are the catalysts for the reactions of larger clusters, the network quickly reaches a final static state. By numerically integrating the master equations (16), the cluster-size distribution evolves to a final, time-independent form $c_s \sim e^{-s/s^*}$ for large s, with $s^* \approx 3.21$ (Fig. 11.).

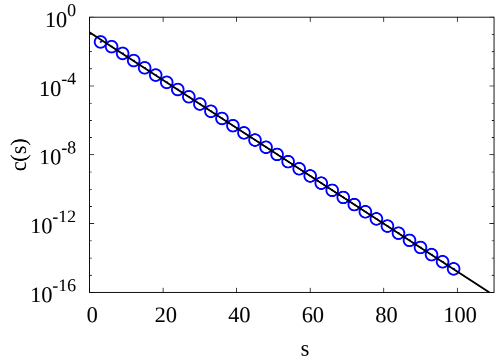


Figure 11. Numerical solution to Eqs. (16) (circles) and an exponential fit to this data (line). **replace c(s) by** c_s

From the rate equations, the first three integer moments of the cluster-size distribution, $M_n \equiv \sum k^n c_k$, evolve according to

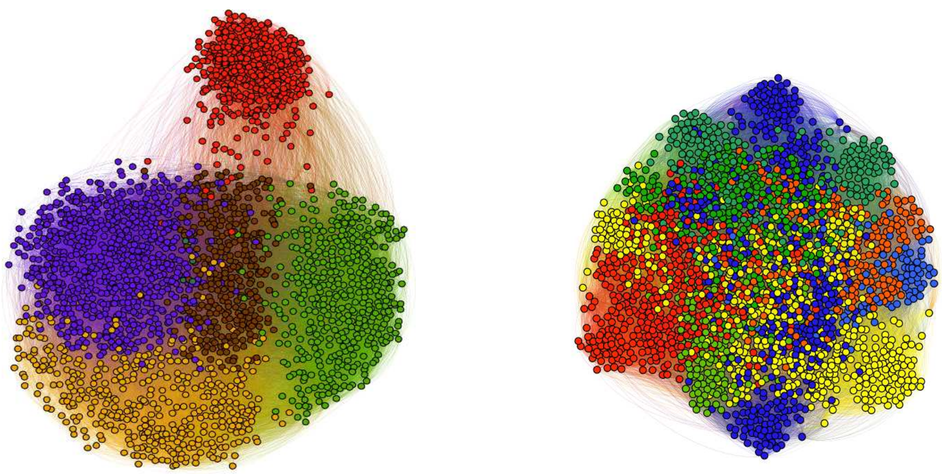
$$\dot{M}_0 = -c_1 - 2c_2, \qquad \dot{M}_1 = 0, \qquad \dot{M}_2 = 2M_2(c_1 + 4c_2)\,.$$

Using the above asymptotic behaviors of c_1 and c_2, we see that the cluster density M_0 approaches a non-zero asymptotic value exponentially quickly in time. Numerically, the final concentration of clusters is given by $M_0(\infty) = 0.1666474164\ldots$. Similarly, the average cluster size saturates to a finite value as $t \to \infty$.

5. Outlook

We introduced a class of acquaintance models in which macroscopic clustering emerges naturally from the underlying social dynamics rather than heterogeneity being one of the building blocks of the model. We showed that our models lead naturally to large-scale community structure. Our approach is based on distinguishing (i) direct linking, where agents with few friends initiate connections with other agents, and (ii) transitive linking, in which two agents become friends as a result of being introduced by a common friend. By controlling the relative rates of direct and transitive linking, we can generate networks that range from nearly complete, with a tiny component of isolated cliques, to highly clustered, that are comprised of well-defined and well-connected communities.

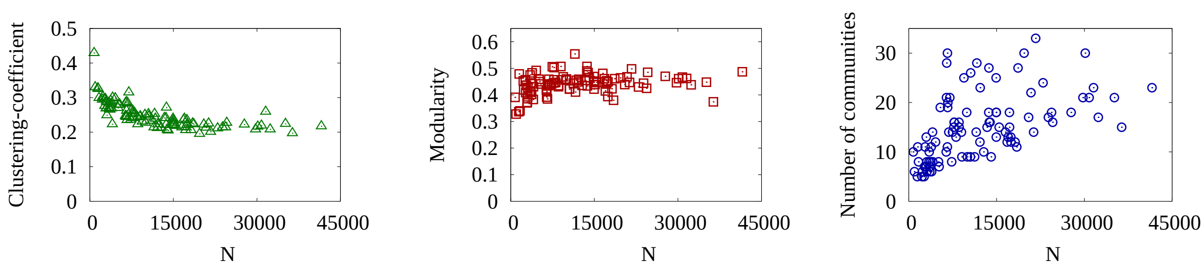
It is natural to inquire whether our acquaintance model can account for the observed features of real networks. A useful empirical dataset with which we can make such a comparison are anonymized Facebook networks from 100 well-known universities in the US. Many of these networks exhibit significant clustering, although not to the same degree as in our acquaintance model. One possible reason for this difference in the degree of clustering is the sharp cutoff between direct and transitive linking. In our model, once an agent has two friends, he no longer makes additional friends directly, a social interaction that would join two communities. Thus the sharp cutoff between direct and transitive linking enhances insularity. However in the real social interactions, individuals with many friends can still engage in direct linking, a mechanism that decreases the modularity and clustering coefficient of the resulting network compared to the case of a sharp cutoff.



(a) (b)

Figure 12. Comparison between (a) the Facebook network of a well-known US university, with clustering coefficient $C = 0.29$ and modularity $Q = 0.45$, and (b) a realization of our soft-cutoff acquaintance model of identical size to (a) in which the rate λ that an agent makes a direct link is an exponentially decaying function of its degree, $\lambda = e^{-k/k^*}$. In (b), the choice $k^* = 7.3$, $F = 0.17$ leads to $C = 0.32$ and $Q = 0.45$. Communities are indicated by the different colors.

We are therefore led to define what we believe is a more realistic "soft-cutoff" model, in which the rate λ at which an agent engages in direct linking is a decreasing function of number of his friends, $\lambda = e^{-k/k^*}$. Figure 12 shows a comparison between an anonymized Facebook network with $N = 2252$ nodes and $L = 84387$ links from a well-known US university and a realization of our soft-cutoff acquaintance model with the same number of nodes and links. For the latter, the values of $k^* = 7.3$ and $F = 0.17$ give modularity and clustering coefficients close to those of the Facebook example.



(a) (b) (c)

Figure 13. (a) Clustering coefficient C, (b) modularity Q, and (c) number of communities for the Facebook networks of 100 well-known US universities.

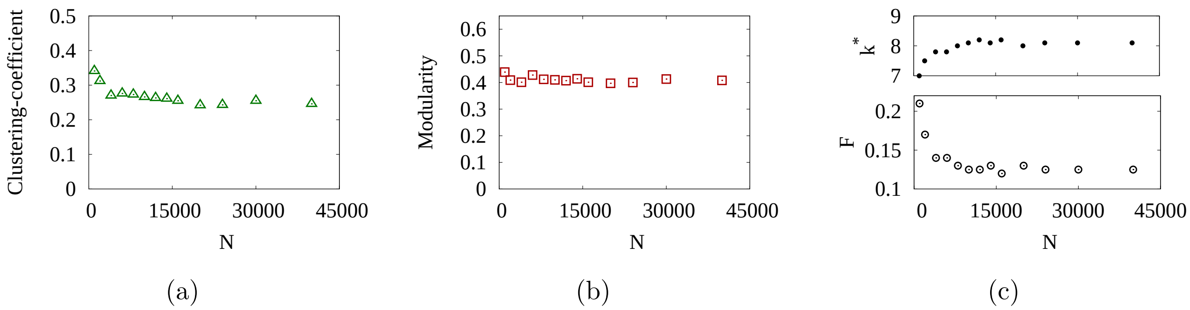


Figure 14. (a) Clustering coefficient and (b) modularity from the soft-cutoff model. (c) The corresponding parameters k^* and F in the soft-cutoff model.

This ability to match our soft-cutoff acquaintance model to Facebook networks extends to the 100 examples in our dataset. Figure 13 shows the clustering coefficient C, modularity Q, and the number of communities of Facebook networks as a function of the number of nodes. As a comparison, Fig. 14 shows the values of C and and Q from our soft-cutoff model as a function of N. For the best match between the Facebook networks and the model, it is necessary to choose k^* weakly increasing with N and F weakly decreasing with N, as given in Fig. 14(c). This comparison shows that our soft-cutoff acquaintance model can match the clustering properties of real social networks with reasonable and slowly varying parameter values.

We thank Mason Porter for providing the Facebook data used in this study. This research was partially supported by the AFOSR and DARPA under grant #FA9550-12-1-0391 and by NSF grant No. DMR-1205797.

6. References

[1] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[2] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[3] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[4] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. USA **101**, 2658 (2004).
[5] G. Palla, I. Derenyi, and I. Farkas, and T. Vicsek, Nature **435**, 814 (2005).
[6] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
[7] V. D. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. P10008, (2008).
[8] J. Copic, M. O. Jackson, and A. Kirman, J. Theor. Econ. **9**, Article 30 (2009).
[9] S. Fortunato, Phys. Repts. **486**, 75 (2010).
[10] A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimera, Eur. Phys. J. B **38**, 373 (2004).
[11] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, Physica A **352**, 669 (2005).
[12] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Math. **6**, 29 (2009).
[13] A. Lancichinetti, S. Fortunato, Phys. Rev. E **84**, 066122 (2011).
[14] M. Boguna, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).
[15] C. Li and P. K. Maini, J. Phys. A: Math. Gen. **38**, 9741 (2005).
[16] R. Toivonen, J.-P. Onnela, J. Saramäki, J. Hyvönen, and K. Kaski, Physica A, **371**, 851 (2006).
[17] R. Lambiotte and M. Ausloos, J. Stat. Mech. (2007) P08026.
[18] R. Lambiotte, M. Ausloos, and J. Holyst, Phys. Rev. E **75**, 030101(R) (2007).
[19] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. San Miguel, J. Conflict Res. **51**, 905 (2007).
[20] B. Golub and M. O. Jackson, Quarterly J. Econ. **127**, 1287 (2012)
[21] J. Davidsen, H. Ebel, and S. Bornholdt, Phys. Rev. Lett. **88**, 128701 (2002).
[22] M. Marsili, F. Vega-Redondo, and F. Slanina, Proc. Natl. Acad. Sci. USA **101**, 1439 (2004).
[23] P. Klimek and S. Thurner, New J. Phys. **15**, 063008 (2013).
[24] R. I. M. Dunbar, J. Human Evol. **22**, 469 (1992).
[25] M. S. Granovetter, Am. J. Sociol. **78**, 1360 (1973).
[26] M. O. Jackson and B. W. Rogers, Amer. Econ. Rev. **7**, 890 (2007).
[27] M. O. Jackson, *Social and Economic Networks* (Princeton University Press, Princeton, NJ, 2008).
[28] A. L. Traud, P. J. Mucha, and M. A. Porter, Physica A **391**, 4165 (2012).
[29] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).
[30] F. Leyvraz, Phys. Repts. **383**, 95 (2003).
[31] P. L. Krapvisky, S. Redner and E. Ben-Naim, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge, 2010).