On the Formation of Community Structures from Homophilic Relationships

Katerine Guerrero and Jorge Finke

Abstract— **Many real-world networks consist of numerous interconnected groups which, as communities, display distinctive collective behavior. The division of a network into communities - groups of nodes with a high density of ties within but a low density of ties between groups - underlies the structure of social and technological networks. In human communities, for instance, individuals may group together according to special interest, occupation, intent, or belief, with tendency to establish stronger ties with individuals who are similar to themselves. Here, we introduce a formal framework for the formation of community structures from homophilic relationships between individuals. Stochastic modeling of local relationships allows us to identify a wide class of agent interactions which lead to the formation of communities and quantify the extent to which group size affects the resulting structure.**

I. INTRODUCTION

Network theory focuses on rendering quantitative relationships into abstract representations (graphs) that allow us to characterize to what extent an underlying set of relationships may contribute to both the evolution of the structure itself and the evolution of dynamics on large interconnected systems. Network models define a set of nodes (e.g., individuals as agents) with ties to other nodes (which may change over time) where the behavior of each agent on a network is conditioned (directly or indirectly) by the behavior of all other agents sharing the same connected component. The collective outcome generally depends on both the nature of the agents (i.e., their decision-making) and the strength of the ties between them (i.e., the degree to which an agent's decision-making affects other agents) [1]. Often, through a process of distributed decision-making, networks gradually develop a distinctive structural pattern, even without centralized information or coordination schemes. The lofty aim of network theory is to provide tools that enables us to define quantitative relationships from reams of amorphous data, enhancing our understanding of the emergence of collective outcomes not foreseen from individual behavior [2].

In trying to capture empirical patterns that reflect the structure of the relationships between agents, common characterizations of network models include: (i) Homophily: Agents tend to associate with others who are similar to themselves [1]. (ii) Transitivity: If agent A shares a tie with B and B with C, then there is an increased probability that A will also share a tie with C [3]. (iii) Community structure:

K. Guerrero is with Facultad de Ingeniería, Universidad Mariana, and Departamento de Ingeniería Electrónica, Universidad de Nariño, Pasto, Colombia, email: dkguerrero@javerianacali.edu.co.

J. Finke is with Departamento de Electrónica y Ciencias de la Computación, Pontificia Universidad Javeriana, Santiago de Cali, Colombia, email: finke@ieee.org.

Networks tend to divide naturally into groups (modules) with dense ties within but sparser ties between them [4].

Several studies identify homophily as a significant determinant for collective outcomes in social networks [5], including empirical patterns of transitivity [6], cultural evolution [7], and friendship formation [1]. For example, in relation to the size of racial groups, the authors of [1] observe that larger groups exhibit higher homophily indices than smaller ones. They suggest that the formation of ties depends on both the conditions of belonging to a group with distinct characteristics and the size of these groups.

Here, we present an agent-based model that allows us to capture the dynamics of the relationships between agents and the formation of community structures in a directed network. To our knowledge the proposed mechanisms is novel in that it generates network substrates with community structures (i) from homophilic interaction between agents (i.e., agents may obey one of two general rules which tend to establish or degrade ties to agents that are unlike themselves); and (ii) from a stochastic Markov process with transition probabilities that are a function of both the homophily index of each group and their relative sizes.

The remainder of this paper is organized as follows: Section II introduces a model of the interaction of agents of two different types. Section III presents analytical results which quantify the effect of group size on the interaction of agents (Lemma 1), showing how agents belonging to the majority drive homophily in network ties. Our main result (Theorem 1) establishes sufficient conditions for the formation of a community for the majority agents. Unlike them, the minority agents ultimately establish significant ties to agents of the majority group (Theorem 2). Section IV presents some numerical results and applies the proposed framework to generate networks with a desired number of communities. The proofs of the lemmas and theorems are available at http://jfinke.org/public_html/publications.html.

II. THE MODEL

The network of $n \in \mathbb{N}$ agents is composed of two groups. Each group represents a type of agent denoted by $\ell \in \{0, 1\}$ (e.g., agents of the same type may share traits, interests, schedules, etc.). Let $N = \{1, \dots, n\}$ be the set of all agents, $t_i : N \to \{0, 1\}$ the type associated to agent $i \in N$, $N_\ell = \{i \in N : t_i = \ell\}$ the set of agents of type ℓ, and $n_\ell = |N_\ell|$ the size of the set N_ℓ (suppose $n_0 \leq n_1$).

A directed graph $G = (N, M)$ captures the interaction between agents. The nodes belonging to N represent the agents and the links associated with M represent some kind of tie between a pair of agents (e.g., affection, dependency,

support). The matrix $M = \{m_{ij}\}$, $m_{ij} \in \{0, 1\}$, quantifies the relationships between agents i and j. In particular, $\forall i, j \in N, i \neq j$, $m_{ij} = 1$ only if there exists an established relation from i to j (e.g, agent i chooses to befriend with agent j), and $m_{ij} = 0$ otherwise. Network relations may be reciprocal, but $\forall i \in N$, $m_{ii} = 0$.

Let $Q_i = \{j \in N : m_{ij} = 1\}$ represent the set of outgoing neighbors with whom agent i maintains some relation, and $Q'_i = \{j \in Q_i : t_i = t_j\}$ the outgoing neighbors with whom it shares the same type. Each agent $i \in N$ may have a variable number of incoming neighbors but has a fixed number of outgoing neighbors (e.g., because of limited resource to establish this links). Let $q = |Q_i|$, $q'_i = |Q'_i| \leq q$, and

$$x_i = \frac{q'_i}{q} \qquad (1)$$

represents the state of the agent $i \in N$, i.e., the proportion of neighbors of the same type. The vector $x(k) = [x_1(k), \ldots, x_n(k)]^\top$ denotes the state of the entire network at time k. To specify the assumptions on our model, we need to introduce the following definitions.

Definition 1: A network G is said to show total segregation if for all $i \in N$, $x_i = 1$.

In other words, a network shows total segregation if only if each agent of type ℓ connects selectively to agents of the same type $(\forall j \in N_\ell, \forall j \in Q_i, j \in Q'_i)$.

Definition 2: The homophily index of the group of type ℓ is given by

$$h_\ell = \frac{1}{n_\ell} \sum_{i \in N_\ell} x_i \qquad (2)$$

Note that the homophily index represents the average fraction of agents of the same type that share ties across the entire network.

Definition 3: A network G is said to be type-neutral if for all $i \in N$, $x_i = 0.5$.

Characterizing G as type-neutral implies that for each group of type ℓ the homophily index $h_\ell = 0.5$. The converse, however, is not necessarily true.

Definition 4: The set N_ℓ is said to show total homophily if $h_\ell = 1$ and total heterophily if $h_\ell = 0$.

Note that if both groups of the network show total homophily then G is totally segregated.

Definition 5: The set N_ℓ forms a community if $h_\ell > 0.5$.

In other words, a group is considered a community if the average number of ties to neighbors of the same type is greater than the average number of ties to neighbors of different type. The model rests upon three main assumptions: first, we need to describe some constraints on the structure of the network; second we specify two general conditions on the establishment of ties from homophilic relationships; and third we constrain the occurrence of events that lead to the formation of communities.

Assumption 1 (Network requirements): The network G satisfies:

    a. Each set N_ℓ can show total homophily $(q < \min\{n_\ell\})$.

    b. An agent can connect to both types of agents $(q > 2)$.

For a group of agents N_ℓ to show total homophily each agent $i \in N_\ell$ can establish at most $n_\ell - 1$ ties. Since agents of both types have the same number of outgoing neighbors we let $q \leq \min\{n_\ell\} - 1$ be bounded by the size of the minority group (a network showing total segregation may support at most $n(\min\{n_\ell\} - 1)$ ties).

Assumption 2 (Agent preferences): The decision to establish relationships between agents is stochastic but must obey one of the following two rules:

    a. Agent i tends to degrade ties to agents that are unlike itself by disconnecting from an agent $j \in Q_i$, $t_i \neq t_j$, and connecting to an agent $j' \notin Q_i$.

    b. Agent i tends to establish ties to agents that are unlike itself by connecting to an agent $j' \notin Q_i$, $t_i \neq t_{j'}$, and disconnecting from an agent $j \in Q_i$.

According to Assumption 2a, agent i may disconnect from a neighboring agent j of different type and connect to another agent $j' \notin Q_i$. According to Assumption 2b, agent i may connect to another agent $j' \notin Q_i$ of different type that is not currently a neighbor and disconnect from a neighboring agent j. Assumption 2 guarantees that the number of outgoing neighbors remains constant for all agents.

Let e_i represent the decision by agent i to reestablish its current relationships (according to Assumption 2a or 2b). At a fixed time, let $E_a = \{e_i : i \in N, t_i = \ell, t_i \neq t_j, x_i \leq h_\ell\}$ denote the set of all possible occurrences that satisfy Assumption 2a for any agent i of any type ℓ such that its state does not exceed the homophily index h_ℓ. Let $E_b = \{e_i : i \in N, t_i = \ell, t_i \neq t_{j'}, x_i \geq h_\ell\}$ denote the set of all possible occurrences that satisfy Assumption 2b for any agent i of any type ℓ such that its state exceeds the homophily index h_ℓ. An event e_i occurs at time k, if $e_i \in g(x)$, where $g(x)$ is a function that enables an event according to the following.

Assumption 3 (Event trajectories): The occurrence of event e_i is stochastic but satisfies the following conditions:

    a. The homophily index of the group N_ℓ ranges between $0 < h_\ell < 1$.

    b. Agents are equally likely to establish or degrade ties to agents that are unlike themselves (both kinds of events occur with equal probability).

    c. The agents j' and j to whom agent i connects and disconnects, respectively, are selected randomly from an uniform distribution.

When $e_i \in g(x)$ occurs, the next state of the entire network $x(k+1)$ is defined by $x(k+1) = f_e(x(k))$ where f_e is an operator that defines the state transitions. If at time k' the network satisfies $h_0(k'), h_1(k') \in \{0, 1\}$, then $E_a = E_b = \emptyset$ and there is no enabled event. We model the deadlock of the network by defining e_0 such that $\forall k \geq k'$, $e_0 \in g(x(k))$, $x_i(k+1) = x_i(k')$ and $h_\ell(k+1) = h_\ell(k')$. If $e_i \in g(x(k))$, $e_i \neq e_0$, then $f_e(x(k)) = x(k+1)$ is given by

$$x_i(k+1) = x_i(k) + \mu_i(k) \qquad (3)$$

where

$$\mu_i(k) = \begin{cases} 0, & \text{if } t_i \neq t_j,\ t_i \neq t_{j'}; \\ 1/q, & \text{if } t_i \neq t_j,\ t_i = t_{j'}; \\ -1/q, & \text{if } t_i = t_j,\ t_i \neq t_{j'}. \end{cases}$$

The state of agent i at time $k + 1$, $x_i(k+1)$, is the state of agent i at time k plus a factor $\mu_i(k)$ that quantifies the increase (decrease) in the agent's state from establishing a new tie to an agent of the same type (to an agent of different type, respectively). The enable function $g(x)$ together with the state transitions $f_e(x)$ define the evolution of the structure of the network.

III. ANALYTIC RESULTS

Assumptions 1-3 specify minimal requirement that will allow us to introduce a local mechanism to generate communities structures without any form of centralized coordination. Here, we first analyze the evolution of ties driven by the probability of occurrence of both rules defined above. We then determine the transition probabilities of the type ℓ homophily index and identify conditions which lead to an aggregate behavior of agents of type ℓ.

A. Probabilities of Rules

If $e_i \in g(x)$, $e_i \in E_a$ and $t_i = \ell$ $(x_i \leq (q-1)/q)$, agent i disconnects from agent $j \in Q_i$, $t_i \neq t_j$, connecting to agent $j' \notin Q_i$. It must be the case that

(i) $t_i = t_{j'}$ with probability

$$P[t_i = t_{j'}] = \frac{n_\ell - q_i' - 1}{n - q - 1} = \frac{n_\ell - qx_i - 1}{n - q - 1} \quad (4)$$

(ii) $t_i \neq t_{j'}$, $t_{j'} = z$ with probability

$$P[t_i \neq t_{j'}] = 1 - P[t_i = t_{j'}] = \frac{n_z - (q - q_i')}{n - q - 1} \quad (5)$$

where $n - q - 1$ is total number of agents that agent i may establish ties to, $n_\ell - q_i' - 1$ is the number of same-type agents, and $n_z - (q - q_i')$ is the number of different-type agents that are not currently neighbors of agent i. Equations (4) and (5) suggest that group size affects the likelihood that agent i shares the same type with a randomly selected agent j'.

On the other hand, group size does not affect how agent i establishes ties according to Assumption 2b. If $e_i \in g(x)$, $e_i \in E_b$, and $t_i = \ell$ $(1/q \leq x_i)$, agent i connects to agent $j' \notin Q_i$, $t_i \neq t_{j'}$, disconnecting from agent $j \in Q_i$. It must be the case that

(i) $t_i = t_j$ with probability

$$P[t_i = t_j] = x_i \quad (6)$$

(ii) $t_i \neq t_j$ with probability

$$P[t_i \neq t_j] = 1 - P[t_i = t_j] = \frac{q - q_i'}{q} = 1 - x_i \quad (7)$$

where $q - q_i'$ is the number of neighbors of different type. Note that (6) and (7) are a function of the state of the agent i, but do not depend on group size n_ℓ.

The following lemma identifies the conditions on the size of the groups that guarantee that the probability that agent i with $t_i = 1$ disconnects from a same-type agent j is less (more for $t_i = 0$) than the probability that agent i connects to a same-type agent j'.

Lemma 1 (Group size effect on agent preferences):
Suppose Assumptions 1-3 hold and $n_1 > (q-1)n_0 + 1$. For every agent $i \in N$, $0 < x_i < 1$, disconnecting from an agent j to connect to an agent j'

$$P[t_i = t_j] < P[t_i = t_{j'}], \qquad \text{if } t_i = 1$$

and

$$P[t_i = t_j] > P[t_i = t_{j'}], \qquad \text{if } t_i = 0.$$

B. Homophily Index Transitions

Next, we apply Lemma 1 to determine the transition probabilities between homophily indices. First, using (2) and (3), if $e_i \in g(x)$ and $t_i = \ell$, the homophily index of type ℓ at time $k + 1$ is given by

$$h_\ell(k+1) = h_\ell(k) + \frac{\mu_i(k)}{n_\ell}$$

Variations in the homophily indices $\Delta h_\ell = \mu_i / n_\ell$ depend on the size of the group ℓ. Let $m > 0$ represent the mth element of the set $H_\ell = \{b/(qn_\ell) : b = 0, \ldots, qn_\ell\}$ be the set of all possible values of h_ℓ. For convenience, let $c_\ell = |H_\ell| = qn_\ell + 1$. The *transition probabilities* between homophily indices are defined as

$$p_{mv} = P[h_\ell(k+1) = w_v | h_\ell(k) = w_m]$$

Using (2), the transition probabilities N_ℓ are

$$p_{mv} = \begin{cases} 0, & \text{if } m \in \{1, c_\ell\} \\ & \quad \text{and } v \neq m; \\ \frac{n_\ell - qw_m - 1}{2(n-q-1)}, & \text{if } m \in \{2, \ldots, c_\ell - 1\} \\ & \quad \text{and } v = m + 1; \\ \frac{w_m}{2}, & \text{if } m \in \{2, \ldots, c_\ell - 1\} \\ & \quad \text{and } v = m - 1; \\ 1 - \frac{n_\ell - qw_m - 1}{2(n-q-1)} - \frac{w_m}{2}, & \text{if } m \in \{2, \ldots, c_\ell - 1\} \\ & \quad \text{and } v = m; \\ 1, & \text{if } m \in \{1, c_\ell\} \\ & \quad \text{and } v = m. \end{cases} \quad (8)$$

Taking into account that the events defined according to Assumption *2a* and *2b* occur with equal likelihood, note the following. First, the homophily index of N_ℓ may change only if $w_m \notin \{0, 1\}$ $(m \notin \{1, c_\ell\})$. Thus, $p_{mv} = 0$ for $v \neq m$ and $p_{mv} = 1$ for $v = m$. If $m \in \{2, \ldots, c_\ell - 1\}$ and $v = m + 1$, the

second term in (8) results from

$$
\begin{aligned}
p_{mv} &= \frac{1}{2n_\ell} \sum_{i \in N_\ell} P[t_i = t_{j'}] \\
&= \frac{1}{2(n-q-1)} \left(n_\ell - 1 - \frac{q}{n_\ell} \sum_{i \in N_\ell} x_i \right) \\
&= \frac{n_\ell - qw_m - 1}{2(n-q-1)}
\end{aligned}
$$

Moreover, if $m \in \{2, \ldots, c_\ell - 1\}$ and $v = m - 1$, the third term results from

$$
p_{mv} = \frac{1}{2n_\ell} \sum_{i \in N_\ell} P[t_i = t_j] = \frac{1}{2n_\ell} \sum_{i \in N_\ell} x_i = \frac{w_m}{2}
$$

And similarly, if $m \in \{2, \ldots, c_\ell - 1\}$ and $v = m$, the fourth term results from

$$
\begin{aligned}
p_{mv} &= \frac{1}{2n_\ell} \left(\sum_{i \in N_\ell} P[t_i \neq t_{j'}] + \sum_{i \in N_\ell} P[t_i \neq t_j] \right) \\
&= \frac{1}{2n_\ell} \left(2n_\ell - \sum_{i \in N_\ell} P[t_i = t_{j'}] - \sum_{i \in N_\ell} P[t_i = t_j] \right) \\
&= 1 - \frac{n_\ell - qw_m - 1}{2(n-q-1)} - \frac{w_m}{2}
\end{aligned}
$$

Note that the transition probabilities defined in (8) satisfy $0 \leq p_{mv} \leq 1$, and for any m, $\sum_{\text{all } v} p_{mv} = 1$. Finally note that, the current value of the homophily index of the group N_ℓ at time index k depends on its value at $k-1$ but not on values further in the past. In other words, a *homogeneous discrete-time Markov chain* characterizes the evolution of the structure of the network.

When $m \in \{1, c_\ell\}$, $p_{mv} = 0$ for all $v \neq m$; therefore $\{w_1\}$ and $\{w_{c_\ell}\}$ are closed subsets of H_ℓ with no transition to any homophily index outside $\{w_1\}$ or $\{w_{c_\ell}\}$. For $m \in \{2, \ldots, c_\ell - 1\}$, $w_m \in H_\ell$ cannot be reached from $\{w_1\}$ or $\{w_{c_\ell}\}$, i.e., the chain is *reducible*. Because $p_{11} = p_{c_\ell c_\ell} = 1$, w_1 and w_{c_ℓ} are *absorbing*. Next, let $T_\ell = \{w_2, \ldots, w_{c_\ell - 1}\}$, $T_\ell \subset H_\ell$, denote the set of all possible homophily indices that are not absorbing.

We denote the transitions probabilities from the homophily index $w_m \in T_\ell$ to the *adjacent* homophily indices by

$$
\begin{aligned}
a_m &= P[h_\ell(k+1) = w_{m+1} | h_\ell(k) = w_m] = p_{m(m+1)} \\
b_m &= P[h_\ell(k+1) = w_{m-1} | h_\ell(k) = w_m] = p_{m(m-1)}
\end{aligned}
$$

Using (8) note that for $m \in \{2, \ldots, c_\ell - 1\}$ the probability $p_{mm} = 1 - a_m - b_m$. For $m = 1$, $a_1 = p_{12} = 0$ and $p_{11} = 1 - a_1 = 1$. Moreover, for $m = c_\ell$, $b_{c_\ell} = p_{c_\ell(c_\ell - 1)} = 0$ and $p_{c_\ell c_\ell} = 1 - b_{c_\ell} = 1$. The transition diagram for the homophily indices is shown in Fig. 1.

Let r_ℓ be the difference among a_m and a_{m+1}, $m \in \{2, \ldots, c_\ell - 2\}$. Using (8) we obtain

$$
\begin{aligned}
r_\ell &= a_m - a_{m+1} \\
&= \frac{n_\ell - qw_m - 1}{2(n-q-1)} - \frac{n_\ell - qw_{m+1} - 1}{2(n-q-1)} \\
&= \frac{1}{2n_\ell(n-q-1)}
\end{aligned} \tag{9}
$$

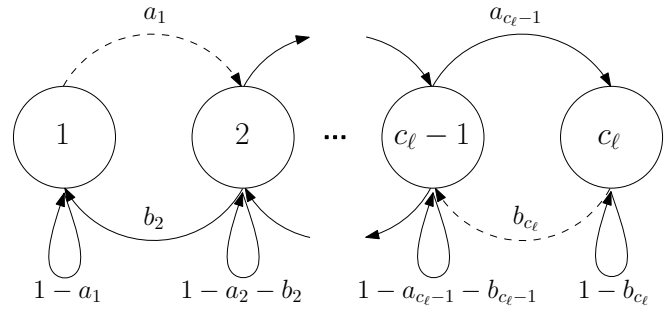

Fig. 1. Transition diagram for the homophily indices for group N_ℓ under Assumptions 1-3, $a_1 = b_{c_\ell} = 0$ (only under a relaxed set of event trajectories, see Assumption 3† and Assumption 3*, may $a_1 > 0$ and/or $b_{c_\ell} > 0$).

Note that since $q < n-1$, $r_\ell > 0$, and $r_\ell = 0$ as $n \to \infty$, we know that $a_m \geq a_{m+1}$. Next, using (8) and (9) and the definitions of a_2 and w_2 yields

$$
a_2 = p_{23} = \frac{n_\ell - qw_2 - 1}{2(n-q-1)} = (n_\ell^2 - n_\ell - 1)r_\ell \tag{10}
$$

Applying the definition of r_ℓ for $m \in \{3, \ldots, c_\ell - 1\}$

$$
a_m = a_{m-1} - r_\ell = a_2 - (m-2)r_\ell \tag{11}
$$

Moreover, for $m \in \{3, \ldots, c_\ell - 1\}$, the difference between b_m and b_{m-1} yields

$$
b_m - b_{m-1} = \frac{w_m - w_{m-1}}{2} = \frac{1}{2qn_\ell} = \frac{w_2}{2} = p_{21} = b_2
$$

Note that $b_2 > 0$ and $b_{m-1} < b_m$. Using (8) and the definition of b_2 yields

$$
b_m = b_{m-1} + b_2 = (m-1)b_2 \tag{12}
$$

From a_2 and b_2 we know the probabilities a_m and b_m for any $m \in \{3, \ldots, c_\ell - 1\}$. These transition probabilities will determine the formation of community behavior of both groups.

C. Reducible Markov Chains Analysis

Given that the Markov chain is reducible, we now compute the probability that the chain enters one closed set before the other. Let q_{m1} be the probability that the chain starting at $w_m \in T_\ell$ visits w_1 before it visits w_{c_ℓ}, defined by

$$
q_{m1} = P[k_{m1} < k_{mc_\ell}] \tag{13}
$$

where $k_{m1} = \min\{k \geq 0 : h_\ell(0) = w_m, h_\ell(k) = w_1\}$ and $k_{mc_\ell} = \min\{k \geq 0 : h_\ell(0) = w_m, h_\ell(k) = w_{c_\ell}\}$. Note that $q_{11} = 1$ and $q_{c_\ell 1} = 0$.

Lemma 2: The probability that the group N_ℓ starting at $0 < h_\ell(0) < 1$ shows total heterophily before total homophily is

$$
q_{m1} = \frac{\gamma_m}{1 + \gamma_m} q_{(m-1)1} \tag{14}
$$

where

$$
\gamma_m = \sum_{i=m}^{c_\ell - 1} \prod_{j=m}^{i} \frac{b_j}{a_j} \tag{15}
$$

We apply Lemma 2 to both types of agents in order to reach the following result.

Lemma 3: Suppose the network $G(0)$ is type-neutral and Assumptions 1-3 hold. Furthermore, suppose the ratio $n_0/n_1 < 1/(q-1)$ as $n_0 \to \infty$. Then the majority group N_1 shows total homophily before total heterophily with probability 1. The minority group N_0 shows total heterophily before total homophily with probability 1.

As a corollary of Lemma 3 it can be show which of the two absorbing homophily index each group N_ℓ visits before the other, starting at a homophily index $w_m \in T_\ell$ such that $m \in \{2, \ldots, c_\ell - 1\}$ for $\ell = 1$ and $m \in \{2, \ldots, (c_\ell + 1)/2\}$ for $\ell = 0$. We cannot however, predict which absorbing homophily index the minority group N_0 visits if it starts at any homophily index w_m, $m \in \{(c_0 + 3)/2, \ldots, c_0 - 1\}$. To compute the probability that the reducible Markov chain enters a closed set $(\{w_1\}$ or $\{w_{c_\ell}\})$, let ρ_m be the probability that the group N_ℓ starting from the current homophily index $w_m \in T_\ell$, first revisits w_m within a finite number of steps,

$$\rho_m = P[\min\{k > 0 : h_\ell(0) = w_m, h_\ell(k) = w_m\} < \infty]$$

A homophily index $w_m \in T_\ell$ is said to be *transient* if $\rho_m < 1$, and it is said to be *recurrent* if $\rho_m = 1$. In other words, a transient homophily index may be visited again, but with the positive probability $1 - \rho_m$ it will not; on the other hand, a recurrent homophily index will definitely be visited again [8].

Lemma 4: Suppose Assumptions 1-3 hold. Furthermore, suppose the ratio between groups $n_0/n_1 < 1/(q-1)$ as $n_0 \to \infty$. All homophily indices $0 < h_\ell < 1$, $\ell \in \{0,1\}$, are transient.

As in [8], to compute the probability that the reducible Markov chain enters the closed set $\{w_1\}$, let ρ_{m1} be the probability that the type ℓ chain starting at the transient homophily index $w_m \in T_\ell$

$$\begin{aligned}\rho_{m1} &= P[h_\ell(k) = w_1 \text{ for some } k > 0 | h_\ell(0) = w_m] \\ &= p_{m1} + \sum_{r : w_r \in T_\ell, |r - m| \leq 1} \rho_{r1} p_{mr} \end{aligned} \quad (16)$$

Note that (16) takes into account that either the chain enters w_1 at $k = 1$, or that it enters a transient homophily index at $k = 1$ and eventually reaches w_1. The summation term in (16) captures the homophily index w_m and all transient homophily indices which are adjacent to w_m. Let $\beta_j = b_j/a_j$ where b_j (a_j) is the probability of transitioning from the current homophily index to the lower (higher, respectively) homophily index.

Lemma 5: The probability ρ_{m1} for the group N_ℓ starting at $0 < h_\ell(0) < 1$ is

$$\rho_{m1} = \begin{cases} \rho_{21}\left(1 + \displaystyle\sum_{i=2}^{m-1}\prod_{j=2}^{i}\beta_j\right) - \displaystyle\sum_{i=2}^{m-1}\prod_{j=2}^{i}\beta_j, \\ \qquad \text{if } m \in \{3, \ldots, c_\ell - 2\}; \quad (17a) \\ \dfrac{\beta_{c_\ell - 1}}{1 + \beta_{c_\ell - 1}}\rho_{(c_\ell - 2)1}, \\ \qquad \text{if } m = c_\ell - 1. \quad (17b) \end{cases}$$

Using Lemma 5 we can now state our main results.

Theorem 1: Suppose Assumptions 1-3 hold. Furthermore, suppose the ratio $n_0/n_1 < 1/(q-1)$ as $n_0 \to \infty$. The majority group N_1 starting at $h_1(0) > 0$ shows total homophily with probability 1 (i.e., the probability that the majority group N_1 starting at any $0 < h_1(0) < 1$ enters the absorbing homophily index $h_1 = 0$ is $\lim_{n_0 \to \infty} \rho_{m1} = 0$).

Theorem 2: Suppose Assumptions 1-3 hold. Furthermore, suppose the ratio $n_0/n_1 < 1/(q-1)$ as $n_0 \to \infty$. The minority group N_0 starting at $h_0(0) < 1 - 1/(qn_0)$ shows total heterophily with probability 1 (i.e., the probability that the minority group N_0 starting at any $0 < h_0(0) < 1 - 1/(qn_0)$ enters the absorbing homophily index $h_0 = 0$ is $\lim_{n_0 \to \infty} \rho_{m1} = 1$).

According to Theorems 1 and 2, agents belonging to N_1 ultimately connect selectively to agents of the same type, and agents belonging to N_0 ultimately connects selectively to agents of a different type. Note that the network G does not show total segregation.

IV. SIMULATION RESULTS

We implement the proposed model and illustrate the effect of different assumptions on the resulting network. First, we relax Assumption 3 in order to get a network G where the majority (minority) group cannot reach total homophily (total heterophily, respectively).

Assumption 3†: The occurrence of event e_i is stochastic but satisfies the following conditions:

a. The homophily index of the group N_ℓ ranges between $0 \leq h_\ell \leq 1$.
b. Agents are equally likely to establish or degrade ties to agents that are unlike themselves.
c. The agents j' and j to whom agent i connects and disconnects, respectively, are selected randomly from an uniform distribution.

Under Assumption 3†, the homophily indices representing total heterophily and total homophily are no longer absorbing. Let $G(0)$ be type-neutral and suppose Assumptions 1-3† hold. Let $q = 4$ and the size of the majority group be $n_1 = 25$. Fig. 2 shows the mean of the average homophily index for both the majority group N_1 and minority group N_0, varying the size of the minority group, n_0, from 5 to n_1. Each data point corresponds to thirty simulation runs. The standard deviation of average homophily index is close to zero, which suggests a remarkable influence of group size over the behavior of the agents and the emergence of structured communities. Here, for values of $n_0/n_1 \leq 16/25$ the group N_1 behaves as community according to Definition 5 (page 2), whereas for any value n_0/n_1 the minority group N_0 does not reach a homophily index above 0.5 (i.e., the minority group does not behave as community). The simulation results are useful to identify design characteristics that the network G must satisfy to reach a desired homophily index for the majority group.

Next, we relax Assumption 3 in order to get a network G where the minority group N_0 cannot reach total heterophily.
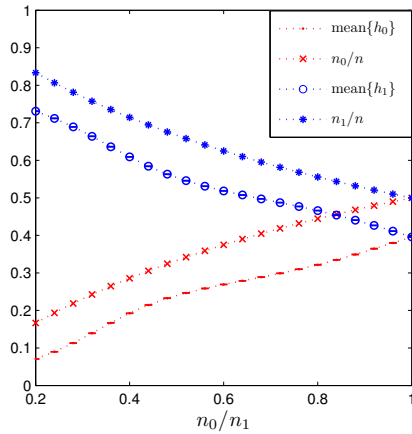
Fig. 2. Mean of the average of the type ℓ homophily index and fraction of type ℓ agents inside of the network G.
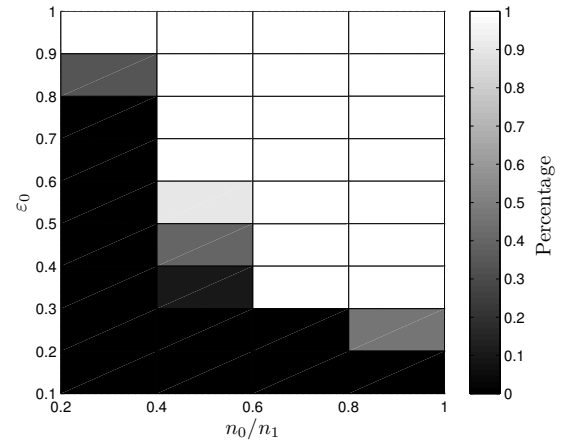

Fig. 3. Percentage of times that the minority group N_0 shows total homophily ($h_{c_0} = 1$).

Assumption 3*: The occurrence of event e_i is stochastic but satisfies the following conditions:

a. The homophily index of the group N_ℓ ranges between $0 \leq h_\ell < 1$.
b. Agents of the minority (majority) group degrade or establish ties to agents that are unlike them with probability ε_0 (ε_1) and $1 - \varepsilon_0$ ($1 - \varepsilon_1$, respectively), respectively.
c. The agents j' and j to whom agent i connects and disconnects, respectively, are selected randomly from an uniform distribution.

Under Assumption 3* the homophily index representing total heterophily is no longer absorbing. Assumption 3*b introduces parameters ε_0 and ε_1 (both in the range of zero to one) which modify the likelihood of occurrence of both rules. Again let network $G(0)$ be type-neutral and suppose Assumptions 1-3* hold. For the agents of the majority group N_1 let the parameter $\varepsilon_1 = 0.5$. For an agent $i \in N_0$ we vary the parameter ε_0. Fig. 3 suggests that the smaller n_0/n_1, the greater should be ε_0 in order that the minority group N_0 shows total homophily and behaves as a community.

Finally, we add a new group to the network. Each group represents a type of agent denoted by $\ell \in \{0, 1, 2\}$ and $t_i : N \rightarrow \{0, 1, 2\}$ is the type associated to agent $i \in N$. We let agents from different groups act as majority or minority agents depending on the relative size of the group they belong to (i.e., allowing for overlapping communities to emerge). Suppose Assumptions 1-3† hold. The transition probabilities for the group of type ℓ are still governed by the size group (n_ℓ), the total number of agents (n) and the total number of relationships of each agent (q) in the network. Fig. 4 shows the network generated by three groups of 50, 15 and 5 agents. In its interaction with agents of different type, the group of intermediate size (with 15 members) behaves as a minority group relative to the combined size of the remaining agents (with 55 members). It is no surprise that the smallest and largest groups present the lowest and highest homophily index, respectively.
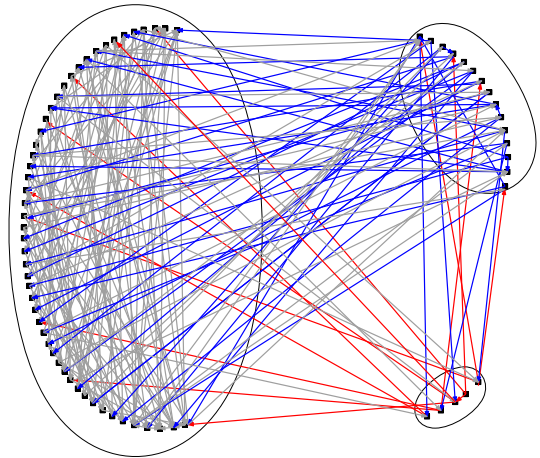

Fig. 4. Network with three groups.

REFERENCES

[1] S. Currarini, M. O. Jackson, and P. Pin, "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, vol. 77, no. 4, pp. 1003–1045, 2009.
[2] J. W. Forrester, "Counterintuitive behavior of social systems," *Theory and Decision*, vol. 2, no. 2, pp. 109–140, 1971.
[3] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, vol. 31, no. 3, pp. 204–213, 2009.
[4] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Physical Review Letters*, vol. 100, no. 11, p. 118703, 2008.
[5] M. Pearson, C. Steglich, and T. Snijders, "Homophily and assimilation among sport-active adolescent substance users," *Connections*, vol. 27, no. 1, pp. 51–67, 2006.
[6] P. Hoff, "Modeling homophily and stochastic equivalence in symmetric relational data," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, vol. 20, pp. 657–664.
[7] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. S. Miguel, "Homophily, cultural drift, and the co-evolution of cultural groups," *Journal of Conflict Resolution*, vol. 51, no. 6, pp. 905–929, 2007.
[8] C. Cassandras and S. Lafortune, *Introduction to discrete event systems*, 2nd ed. Springer, 2008.