ARTICLE

OPEN

Check for updates

Heterogeneous node copying from hidden network structure

Max Falkenberg [1]✉

Node copying is an important mechanism for network formation, yet most models assume uniform copying rules. Motivated by observations of heterogeneous triadic closure in real networks, we introduce the concept of a hidden network model—a generative two-layer model in which an observed network evolves according to the structure of an underlying hidden layer—and apply the framework to a model of heterogeneous copying. Framed in a social context, these two layers represent a node's inner social circle, and wider social circle, such that the model can bias copying probabilities towards, or against, a node's inner circle of friends. Comparing the case of extreme inner circle bias to an equivalent model with uniform copying, we find that heterogeneous copying suppresses the power-law degree distributions commonly seen in copying models, and results in networks with much higher clustering than even the most optimum scenario for uniform copying. Similarly large clustering values are found in real collaboration networks, lending empirical support to the mechanism.

[1] Centre for Complexity Science, Imperial College London, London, UK. ✉email: max.falkenberg@protonmail.com

Node copying is an important network growth mechanism[1–7]. In social networks, copying is synonymous with triadic closure, playing an important role in the emergence of high clustering[8,9]. In biology, node copying encapsulates duplication and deletion, a key mechanism in the formation of protein-interaction networks[10–14].

Despite this range of applications, most node copying models assume uniform, or homogeneous copying, i.e., that the probability of copying any given neighbour of a node is equal. The exact formulation varies widely, but examples include "links are attached to neighbours of [node] j with probability p"[5], or "one node [is duplicated]... edges emanating from the newly generated [node] are removed with probability δ"[11]. Many other models use similar uniform copying rules[2,3,6–8,10,12,15–26].

Homogeneous copying is a sensible base assumption, often aiding a model's analytical tractability. However, especially in a social context, there are good reasons to believe that node copying may be heterogeneous. As an example, consider the social brain hypothesis, a theory which suggests that the average human has around 150 friends (Dunbar's number), encapsulating progressively smaller sub-groups of increasing social importance[27,28]. In contrast, large social networks often have an average degree far exceeding Dunbar's number[29], implying that most of these observed friends are only distant acquaintances. In this context, if individual A introduces individual B to one of their friends, C, (i.e., B is copying A's friend C), we may reasonably expect that C is more likely to be chosen from A's inner social circle, than A's wider social circle.

This is directly related to the principle of strong triadic closure: "If a node A has edges to nodes B and C, then the B−C edge is especially likely to form if A's edges to B and C are both strong ties"[30]. In weighted networks where tie strength can be equated to edge weight, empirical evidence for the strong triadic closure principle can be inferred by measuring the neighbourhood overlap between two nodes as a function of tie strength[30]; for example, using mobile communication networks[31], or using face to face proximity networks[32].

Unfortunately, for many networks tie strength data is unavailable or unknown. In these cases, evidence for asymmetric triadic closure may be inferred through proxy means. For instance, in academic collaboration networks, it has been shown that the ratio of triadic closure varies strongly with the number of shared collaborators between nodes[33]. Although the average triadic closure ratio is small (typically < 10%), the ratio rapidly increases with the number of shared collaborators. However, these aggregate measures are highly coarse-grained and likely only approximate real closure dynamics.

This motivates the study of simple heterogeneous copying models[4,5,9,34,35]. Typically these models fall into a small number of distinct categories. In the first, heterogeneity is introduced as a node intrinsic property (e.g., node fitness) in the absence of structural considerations[5]. In the second, heterogeneity is introduced via group homophily where the probability of triadic closure between nodes A and B is dependent on whether nodes A and B are in the same group or different groups (e.g., researchers from the same academic discipline, as opposed to different disciplines)[9,35]. However, intra-group copying is typically modelled uniformly. Finally, some models consider heterogeneous copying driven by the network structure around nodes A and B, without introducing node homophily[4].

Bhat et al.[4] define a threshold model where node A introduces node B to one of their friends C. An edge then forms between B and C if the fraction of neighbours common to B and C exceeds some threshold F. The model demonstrates a transition from a state where networks are almost complete for small F, to a state where networks are sparse but highly clustered as F increases past

a critical threshold. However, the model is limited in its tractability and has peculiarities such as the observation that fringe communities are almost always complete.

In the current work, our aim is to extend these ideas and introduce a more general framework for heterogeneous node copying based on the concept of hidden strong ties. To do so, we introduce the hidden network model, a framework based on multilayer networks[36] where layers have identical node structure but different edge structure. The framework lets us build models where local heterogeneity in the rules of network growth is a property of the hidden network structure and not arbitrarily encoded using node intrinsic properties or group homophily. The concept is closely related to other multilayer paradigms including the use of replica nodes to model heterogeneity[37], interdependent networks[38], and multilayer copying[21].

In the remainder of this paper, we define and analytically study the case of extreme heterogeneous copying, the correlated copying model (CCM). The CCM is an adaptation of the uniform copying model introduced by Lambiotte et al.[3], see Fig. 1a. In the UCM, a single node, α, is added to the network at time t_α, and connects to one target node, β, which is chosen uniformly at random. The formation of an edge between the new node and the target node puts the UCM in the class of corded copying models; Steinbock et al.[16] refer to the UCM as the corded node duplication model. We label each neighbour of β with the index γ_j where $j \in \{1, \cdots, k^\beta\}$, and k^β is the degree of node β. For each neighbour γ_j, the copied edge (α, γ_j) is added to the network independently with probability p.

Following the convention of previous copying models, the nodes α and β are sometimes referred to as the daughter and mother nodes respectively. The network is initialised at $t = 1$ with
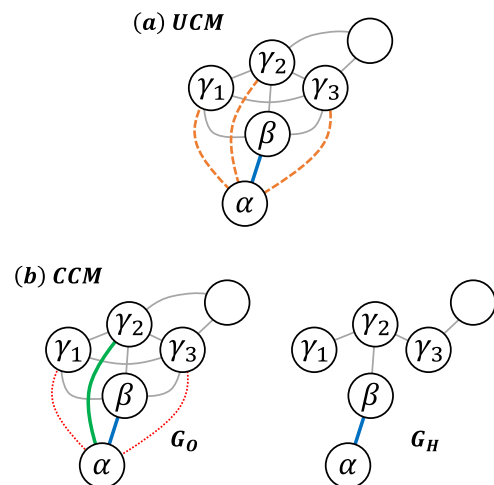

Fig. 1 Two models of network formation via node copying. a The uniform copying model (UCM), and **b** the correlated copying model (CCM). The UCM consists of a single layer. The CCM has an observed layer, in which copying takes place, G_O, and a hidden layer, G_H. For both models, a new node α (the daughter) is added to the existing network (nodes connected by grey edges) and forms a random link (blue) to a target node, β (the mother). **a** In the UCM, there is a uniform probability, p, of forming an edge to each of β's neighbours (γ_1, γ_2, γ_3; orange dashed edges). **b** In the CCM, copied edges are added to the observed network, G_O, deterministically. If an edge exists in the hidden network, G_H, between node β and node γ_j (e.g., the $\{\beta, \gamma_2\}$ edge), then node α copies that edge in G_O (e.g., forming the $\{\alpha, \gamma_2\}$ edge; solid green). If an edge does not exist in G_H (e.g., the $\{\beta, \gamma_1\}$ and $\{\beta, \gamma_3\}$ edges), the corresponding edges are not copied to G_O (red dotted lines). Copied edges are never added to G_H.

a single node. If $p = 0$, no edges are copied resulting in a random recursive tree. If $p = 1$, the UCM generates a complete graph.

In addition to the extreme copying case (the CCM), we numerically investigate a generalised form of the correlated copying model (GCCM) which interpolates between the UCM and CCM. The GCCM generates a diverse spectrum of network structures spanning both ergodic sparse and non-ergodic dense networks, with degree distributions ranging from exponential decay, through stretched-exponentials and power-laws, to extremely fat-tailed distributions with anomalous fluctuations. These networks exhibit a broad clustering spectrum from sparse networks with significantly higher clustering than their uniform equivalents to the unusual case where networks are almost complete, but with near-zero clustering. We comment on a selection of real collaboration networks, which, in line with the CCM, exhibit higher clustering than can be explained by uniform copying. This suggests that heterogeneous copying may be an important explanatory mechanism for social network formation.

Results

Hidden network models. We define a hidden network model as the pair of single layer graphs $G = (G_O, G_H)$, comprising an observed network $G_O = (V, E_O)$ and a hidden network $G_H = (V, E_H)$, where V is the set of nodes for both networks and E_O and E_H are the set of edges for each network. The set V represents the same entities in both G_O and G_H, with differences lying exclusively in the edge structure between nodes. The key feature of a hidden network model is that the evolution of G_O is dependent on G_H (or vice versa). Mathematically, this is closely related to interdependent networks[39].

Correlated copying model. In the CCM, see Fig. 1b, the observed and hidden networks are initialised with a single node at $t = 1$. At $t = t_\alpha$, node α is added to both networks and a single target node, β, is chosen uniformly at random. We label the k_O^β neighbours of β in G_O with the index γ_j. Then, in the observed network only, the copied edge (α, γ_j) is formed with $p_{hid} = 1$ if the edge $(\beta, \gamma_j) \in E_H$, $p_{obs} = 0$ otherwise. The general case with intermediate copying probabilities is discussed in section "General correlated copying model". No copied edges are added to the hidden network G_H. The direct edge (α, β) is added to both G_O and G_H. The CCM therefore also falls into the class of corded node duplication models. Using the convention of referring to β as the mother node and α as the daughter node, we note that the hidden network consists exclusively of first-order relations (mother−daughter), whereas edges found only in the observed network correspond to second-order relations (sister−sister, or grandmother−granddaughter).

G_H evolves as a random recursive tree. Unlike the UCM, all copying in G_O is deterministic, with the only probabilistic element emerging in the choice of the target node β. For comparative purposes, we define the effective copying probability in the CCM as $p_{eff} = \langle k_H^\beta / k_O^\beta \rangle$, i.e., the fraction of the observed neighbours of node β which are copied by node α.

Framed in a social context, we might think of G_O as an observed social network where individuals have many friends, but the quality of those friendships is unknown, with most ties being weak. In contrast, underlying every social network is a hidden structure representing the inner social circle of individuals, where a node is only connected to their closest friends[28]. Copying in the CCM is biased to this inner circle.

Basic topological properties. The total number of edges in G_H scales as $E_H(t) \sim t$, with the average degree given by $\langle k_H \rangle = 2$.

Using the degree distribution of G_H, see below, $\langle k_H^2 \rangle = 6$. In the observed network, each time step a single edge is added by direct attachment, and one copied edge is added for each neighbour of the target node in G_H, k_H^β. The average change in the number of edges is therefore $\langle \Delta E_O(t) \rangle = 1 + \langle k_H^\beta \rangle = 1 + \langle k_H \rangle = 3$, such that $\langle E_O(t) \rangle \sim 3t$ and $\langle k_O \rangle = 6$.

As an alternative, note that the observed degree of node α can be written as

$$(k_O)_\alpha = \sum_{\beta=1}^{(k_H)_\alpha} (k_H)_{\alpha,\beta} \tag{1}$$

where the index α, β labels the $(k_H)_\alpha$ unique neighbours of α in G_H. Averaging both sides of Eq. (1) over all nodes we find,

$$\langle k_O \rangle = \frac{1}{t} \sum_{\alpha=1}^{t} \sum_{\beta=1}^{(k_H)_\alpha} (k_H)_{\alpha,\beta} = \frac{1}{t} \sum_{\ell=1}^{t} n_\ell \cdot (k_H)_\ell, \tag{2}$$

where n_ℓ is the number of times that the degree of node ℓ appears in the expanded sum. For any tree graph, node ℓ will appear exactly once in Eq. (2) for each of its $(k_H)_\ell$ neighbours. Hence, $n_\ell = (k_H)_\ell$ and $\langle k_O \rangle = \langle k_H^2 \rangle$. In Supplementary Note 1, Eq. (1) is used to derive $\langle k_O^2 \rangle \approx 62$.

We may naively expect that the effective copying probability is $p_{eff} = \langle k_H \rangle / \langle k_O \rangle = 1/3$. However, for the CCM, $p_{eff} = \langle k_H^\beta / k_O^\beta \rangle \neq \langle k_H \rangle / \langle k_O \rangle$. We have not found a route to calculating this exactly, but simulations suggest $p_{eff} \approx 0.374$.

Degree distribution. The hidden network evolves as a random recursive tree which has a limiting degree distribution given by

$$p_H(k_H) = 2^{-k_H}, \text{ for } k_H > 1. \tag{3}$$

In Supplementary Note 2, we show that the degree distribution for the observed network can be written as the recurrence

$$p_O(k_O) = \frac{\pi_O(k_O - 1) \cdot p_O(k_O - 1) + 2^{1-k_O}}{1 + \pi_O(k_O)}, \text{ for } k \geq 2, \tag{4}$$

where the final term is the probability that at time t the newly added node has initial degree k_O and

$$\pi_O(k_O) = 1 + \langle k_H | k_O \rangle, \tag{5}$$

with $\langle k_H | k_O \rangle$ as the average degree of nodes in the hidden network with observed degree k_O. Here, the 1 corresponds to edges that are gained from direct attachment, whereas $\langle k_H | k_O \rangle$ corresponds to edges gained from copying. Although we have not found an exact expression for $\langle k_H | k_O \rangle$, we can make progress by considering the evolution of individual nodes.

Consider node α added to the network at t_α. The initial conditions for node α are

$$(k_H(t_\alpha))_\alpha = 1, \tag{6a}$$

$$\langle k_O(t_\alpha) \rangle_\alpha = 1 + \langle k_H(t_\alpha - 1) \rangle_\beta, \tag{6b}$$

where the final term is the average hidden degree of the target node β. In G_H, node α gains edges from direct attachment only. Hence, at $t > t_\alpha$,

$$\langle k_H(t) \rangle_\alpha = 1 + \sum_{j=t_\alpha}^{t-1} \frac{1}{j} = 1 + H_{t-1} - H_{t_\alpha-1}, \tag{7}$$

where H_n is the nth harmonic number. In G_O, either node α is targeted via direct attachment, or a copied edge is formed from the new node to node α via any of the $(k_H(t))_\alpha$ neighbours of node α. Hence,

$$\langle k_O(t) \rangle_\alpha = \langle k_O(t_\alpha) \rangle_\alpha + \sum_{j=t_\alpha}^{t-1} \frac{1 + \langle k_H(j) \rangle_\alpha}{j}$$

$$= \langle k_O(t_\alpha) \rangle_\alpha + \sum_{j=t_\alpha}^{t-1} \frac{2 + H_j - H_{t_\alpha-1} - 1/j}{j}, \quad (8)$$

where we have subbed in Eq. (7) and $H_{j-1} = H_j - 1/j$. Evaluating this sum, see Supplementary Note 2, we find

$$\langle k_O(t) \rangle_\alpha = \langle k_O(t_\alpha) \rangle_\alpha + \frac{1}{2} \Big[(4 + H_{t-1} - H_{t_\alpha-1}) \\ \times (H_{t-1} - H_{t_\alpha-1}) - H_{t-1}^{(2)} + H_{t_\alpha-1}^{(2)} \Big], \quad (9)$$

where $H_n^{(m)}$ is the nth generalised Harmonic number of order m. For $t \to \infty$, $H_t^{(2)} \to \pi^2/6$. Hence, for large t we can drop the final two terms and substitute in Eq. (7) to give

$$\langle k_O(t) \rangle_\alpha \approx \langle k_O(t_\alpha) \rangle_\alpha + \frac{1}{2} \big(\langle k_H(t) \rangle_\alpha + 3 \big) \big(\langle k_H(t) \rangle_\alpha - 1 \big). \quad (10)$$

Noting, that Eq. (10) is a monotonically increasing function of k_H for $k_H > 1$, we assume that we can drop the index α and the time dependence giving the average observed degree of nodes with specific hidden degree as

$$\langle k_O|k_H \rangle \approx \langle \tilde{k}_O|k_O \rangle + \frac{1}{2} (k_H + 3)(k_H - 1), \quad (11)$$

where $\langle \tilde{k}_O|k_O \rangle$ denotes the average initial observed degree of nodes with current degree k_O. Finally, we make the approximation that $\langle k_H|k_O \rangle \approx \langle k_O|k_H \rangle^{-1}$ where the exponent denotes the inverse function. This gives

$$\pi_O(k_O) = 1 + \langle k_H|k_O \rangle \approx \sqrt{2(k_O + 2 - \langle \tilde{k}_O|k_O \rangle)}. \quad (12)$$

To proceed, let us solve the degree distribution at $k_O = 2$. Although the average initial condition $\langle \tilde{k}_O \rangle = 1 + \langle k_H \rangle = 3$, in this case $\langle \tilde{k}_O|2 \rangle = 2$. Therefore

$$p_O(2) = -\pi_O(2) \cdot p_O(2) + 2^{-1} = -p_O(2) \cdot \sqrt{2(2)} + 2^{-1}, \quad (13)$$

giving $p_O(2) = 1/6$. Since $\langle \tilde{k}_O|k_O \rangle$ has an almost negligible effect on $\pi_O(k_O)$ for $k_O > 2$, for simplicity we set $\langle \tilde{k}_O|k_O \rangle = 2$. We can now rewrite Eq. (4) as

$$p_O(k_O) = \frac{p_O(k_O - 1)\sqrt{2(k_O - 1)} + 2^{1-k_O}}{1 + \sqrt{2k_O}}, \text{ for } k_O > 2. \quad (14)$$

Although computing this recurrence shows good agreement with simulations, see Fig. 2, we have not found a closed-form solution to Eq. (14).

As an approximation, we return to Eq. (9) and note that $H_{t-1} - H_{t_\alpha-1} \approx \ln(t/t_\alpha)$. Substituting this into Eq. (9) and dropping small terms

$$\langle k_O(t > t_\alpha) \rangle_\alpha \approx 2\ln(t/t_\alpha) + \frac{\ln^2(t/t_\alpha)}{2}, \quad (15)$$

which inverted gives

$$\ln(t/t_\alpha) \approx -2 + \sqrt{2(k_O + 2)} \approx \sqrt{2k_O}, \text{ for } k \gg 2. \quad (16)$$

We have dropped the expectation value and define t_α as the time a node was created such that its degree at time t is approximately k_O. Exponentiating each side and taking the reciprocal,

$$\frac{t_\alpha}{t} \approx e^{-\sqrt{2k_O}}. \quad (17)$$

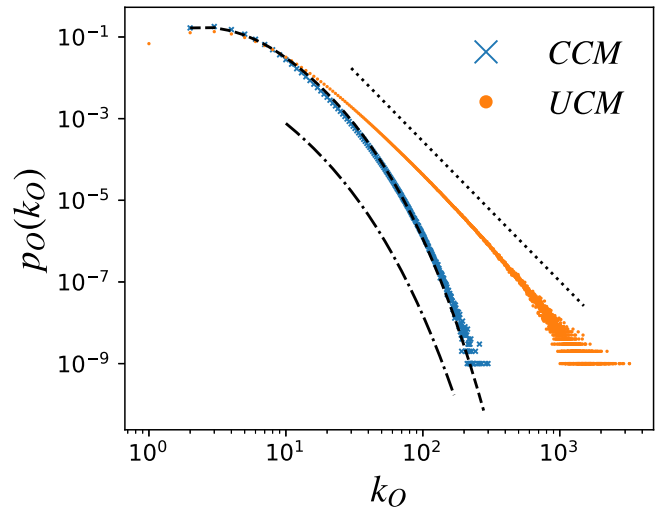Finally, by substituting this approximation into the cumulative

Fig. 2 The degree distributions for the correlated copying model (CCM) and uniform copying model (UCM). Degree probability, $p_O(k_O)$, plotted as a function of the observed degree, k_O. UCM initialised with copying probability $p = 0.374$ (equal to the CCM's effective copying probability). Networks grown to $t = 10^7$, averaged over 100 networks. Error bars omitted for clarity. Dashed line: analytical expression for CCM in Eq. (14). Dot-dashed: stretched exponential approximation. Dotted: power-law scaling.

degree distribution we find

$$\tilde{p}_O(k_O) = \sum_{k'_O=2}^{k_O} p_O(k'_O) \approx 1 - \frac{t_\alpha}{t} \approx 1 - e^{-\sqrt{2k_O}}, \quad (18)$$

which corresponds to a Weibull (stretched exponential) distribution, suppressing the power-law scaling observed in the UCM, see Fig. 2.

The approximation for the cumulative degree distribution stems from the observation that, on average, nodes with $k' > k_O$ were added to the network at $t' < t_\alpha$, whereas nodes with $k' < k_O$ were added to the network at $t' > t_\alpha$. Note that Eq. (18) is close to the scaling expected from sub-linear preferential attachment[40] with an exponent 1/2.

Clique distribution. In a simple undirected graph, a clique of size n is a subgraph of n nodes that is complete. A clique of size $n = 2$ is an edge, whereas $n = 3$ is a triangle. Here we calculate the exact scaling for the number of n cliques, $Q_n(t)$, in G_O.

Let us first consider the case of triangles. At $t = t_\alpha$, there are two mechanisms by which a new triangle forms:

1. **Direct triangles.** The new node, α, forms a direct edge to the target node, β, and forms copied edges to each of the k_H^β neighbours of node β, labelled with the index γ_j. The combination of the direct edge (α, β), the copied edge (α, γ_j), and the existing edge (β, γ_j) creates one triangle, $(\alpha, \beta, \gamma_j)$, for each of the k_H^β neighbours.
2. **Induced triangles.** If node α forms copied edges to both node γ_j, and to node $\gamma_{j'}$, $j \neq j'$, the triangle $(\alpha, \gamma_j, \gamma_{j'})$ is formed if $(\gamma_j, \gamma_{j'}) \in E_O$.

Combining these mechanisms, the change in the number of triangles can be written as

$$\Delta Q_3(t_\alpha) = \Delta Q_3^D(t_\alpha) + \Delta Q_3^I(t_\alpha), \quad (19)$$

where the first and second terms on the right correspond to direct and induced triangles respectively. One new direct triangle is formed for each of the k_H^β neighbours of node β, $\Delta Q_3^D = k_H^\beta$. For induced triangles, the copied edge (α, γ_j) is only formed if $(\beta, \gamma_j) \in E_H$.

Additionally, all pairs of nodes which are next-nearest neighbours in G_H must be nearest neighbours in G_O. Hence, the edge $(\gamma_j, \gamma_{j'})$ must exist in the observed network if both γ_j and $\gamma_{j'}$ are copied. As a result, one induced triangle is formed for each pair of copied edges (α, γ_j) and $(\alpha, \gamma_{j'})$ such that

$$\Delta Q_3^I = \begin{pmatrix} k_H^\beta \\ 2 \end{pmatrix} = \frac{(k_H^\beta)^2 - k_H^\beta}{2}. \qquad (20)$$

A visual example of the combinatorics for $k_H^\beta = 3$ is shown in Fig. 3.

Extending the triangle argument to general n we can write

$$\Delta Q_n(t_\alpha) = \Delta Q_n^D(t_\alpha) + \Delta Q_n^I(t_\alpha), \qquad (21)$$

where direct cliques are those which include the edge (α, β). For a clique of size n, the number of direct cliques is given by the number of ways in which $n - 2$ nodes can be chosen from k_H^β nodes,

$$\Delta Q_n^D(t_\alpha) = \begin{pmatrix} k_H^\beta \\ n-2 \end{pmatrix}, \qquad (22)$$

whereas the number of induced cliques is given by the number of ways in which $n - 1$ nodes can be chosen,

$$\Delta Q_n^I(t_\alpha) = \begin{pmatrix} k_H^\beta \\ n-1 \end{pmatrix}. \qquad (23)$$

As $t \to \infty$, the average change in clique number is

$$\langle \Delta Q_n(t) \rangle = \sum_{k_H=1}^{\infty} p_H(k_H) \left[\begin{pmatrix} k_H \\ n-2 \end{pmatrix} + \begin{pmatrix} k_H \\ n-1 \end{pmatrix} \right], \qquad (24)$$

where $p_H(k_H)$ is the probability that the randomly chosen target node $k_H^\beta = k_H$. To avoid ill-defined binomials, we rewrite Eq. (24) as

$$\langle \Delta Q_n(t) \rangle = p_H(n-2) + \sum_{k_H=n-1}^{\infty} p_H(k_H) \cdot \begin{pmatrix} k_H + 1 \\ n-1 \end{pmatrix}, \qquad (25)$$

where we have combined the two terms into a single binomial. After subbing in $p_H(k_H)$ and solving the sum,

$$\langle \Delta Q_n(t) \rangle = 2^{2-n} + \sum_{k_H=n-1}^{\infty} 2^{-k_H} \cdot \begin{pmatrix} k_H + 1 \\ n-1 \end{pmatrix} = 4. \qquad (26)$$

Consequently, for large t we find the curious result that the number of n cliques scales as

$$Q_n(t) \sim 4t, \text{ for } n > 2, \qquad (27)$$

independent of the clique size. In practice, this result only applies for $t \to \infty$. To see this, note that the largest clique in G_O at time t is always directly related to the largest degree node in G_H,

$$\text{Max}(n, t) = \text{Max}(k_H, t) + 1, \qquad (28)$$

with the largest hidden degree at time t scaling as approximately

$$\text{Max}(k_H, t) \sim \ln(t). \qquad (29)$$

We can invert this and ask how large the network is if we observe that the largest observed clique is n. This gives

$$t_n \sim e^n. \qquad (30)$$

Hence, the scaling relation in Eq. (27), is only valid for cliques of size n when $t \gg t_n$. In Supplementary Note 3, we plot the number of cliques in simulations of the CCM as a function of t. For small clique sizes, the scaling in Eq. (27) is clearly apparent early in the evolution of the CCM. However, for moderate and large cliques, the standard deviation in the number of cliques is significantly larger than the average number of cliques, obscuring a clear trend.
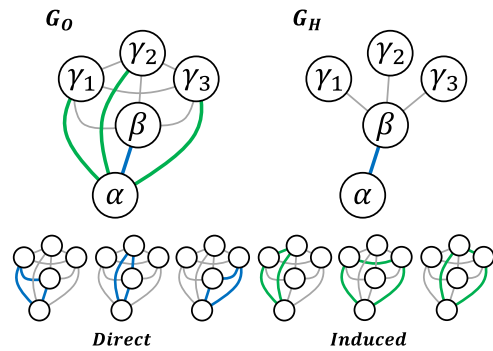

Fig. 3 A schematic illustrating the number of triangles formed in a single time step of the correlated copying model. G_O: the observed network, G_H: the hidden network. The new node, α, forms a direct edge (blue) to a node β which has three existing hidden neighbours ($\gamma_1, \gamma_2, \gamma_3$). The copying process forms three new edges (green) in G_O. The copying process results in three new direct triangles (outlined in blue), involving the edge $\{\alpha, \beta\}$, and three new induced triangles (outlined in green), excluding the edge $\{\alpha, \beta\}$. Triangles are formed in the observed network only; the hidden network remains a random tree. Note that in G_O and G_H, the grey edges represent the existing network, while in the small graphs at the bottom they represent edges that are not part of the outlined triangles.

Clustering. Transitivity is a global clustering measure defined as

$$\tau_{G_O} = 3 \times \frac{\#(\text{triangles in } G_O)}{\#(\text{twigs in } G_O)}, \qquad (31)$$

where a twig is any three nodes connected by two edges. The number of twigs is equivalent to the number of star graphs of size 2, S_2, where a star graph of size n is a subgraph with 1 central node and n connected neighbours. The number of subgraphs of size 2 is related to the degree distribution by

$$S_2(t) = t \sum_{k_O \geq 2} \begin{pmatrix} k_O \\ 2 \end{pmatrix} \cdot p_O(k_O) = t \cdot \frac{\langle k_O^2 \rangle - \langle k_O \rangle}{2}, \qquad (32)$$

where we have used the property that $p_O(k_O < 2) = 0$. Recalling that $\langle k_O \rangle = 6$ and $\langle k_O^2 \rangle \approx 62$, the number of twigs scales as $S_2 \sim 28t$, such that

$$\tau_{G_O} = \frac{3 Q_3}{S_2} \sim \frac{3 \cdot 4t}{28t} = \frac{3}{7}. \qquad (33)$$

The observed network can be recovered from the hidden network by converting every wedge in G_H into a triangle. This can be thought of as complete triadic closure where every possible triangle which can be closed, from the addition of a single edge to the hidden network, is closed. This implies that the CCM has the largest possible transitivity from a single iteration of triadic closure on a random recursive tree.

The local clustering coefficient, $cc(\alpha)$, is defined as the number of edges between the $(k_O)_\alpha$ neighbours of α, normalised by the the number of edges in a complete subgraph of size $(k_O)_\alpha$. For the CCM,

$$cc(\alpha) = \frac{\begin{pmatrix} (k_H)_\alpha \\ 2 \end{pmatrix} + \sum_{\beta=1}^{(k_H)_\alpha} \begin{pmatrix} (k_H)_{\alpha,\beta} \\ 2 \end{pmatrix}}{\begin{pmatrix} (k_O)_\alpha \\ 2 \end{pmatrix}}, \qquad (34)$$

where the first term corresponds to the complete subgraph of the $(k_H)_\alpha$ neighbours of α in G_H, and the sum contributes the edges from one complete subgraph formed by node α, β, and β $(k_H)_{\alpha,\beta} - 1$ neighbours, excluding α. The global clustering coefficient, $CC(G_O)$, is defined as the average of Eq. (34) over

all nodes in the network. In simulations, $CC(G_O) \approx 0.771$ for large t.

Path lengths. Steinbock et al.[16] calculate the distribution of shortest path lengths for the UCM (referred to in their paper as the corded node duplication model). Specifically, the authors calculate the probability that two randomly chosen nodes, i and j, will be separated by a shortest path of length ℓ, denoted as $\mathcal{P}(L = \ell; t)$, at time t.

The UCM with $p = 0$ corresponds to a random recursive tree and is therefore equivalent to G_H in the CCM. Hence, for the hidden network, we can lift the path length distribution, $\mathcal{P}_H(L_H = \ell; t)$, and the mean shortest path, $\langle L_H(t) \rangle$, from Steinbock et al.[16]. We can then exploit a convenient mapping to calculate the distribution of shortest path lengths in G_O from G_H.

Consider two randomly chosen nodes i and j. In G_H, there is a unique path (due to its tree structure) from i to j of length $(\ell_H)_{ij}$. In G_O, the enforced triadic closure process means that for every two steps on the path from i to j in G_H, an observed edge exists in G_O, which acts as a shortcut, reducing the path length by one. Hence, if the path length $(\ell_H)_{ij}$ is even, the path length in G_O is given by $\ell_O = \ell_H/2$; if the path length is odd $\ell_O = (\ell_H + 1)/2$. Using this mapping, we can write

$$\mathcal{P}_O(L_O = \ell; t) = \mathcal{P}_H(L_H = 2\ell; t) + \mathcal{P}_H(L_H = 2\ell - 1; t), \quad (35)$$
$$\text{for } \ell \geq 1.$$

If we assume that, for large t, there are an approximately equal number of odd and even shortest paths in G_H, the average shortest path length in G_O is

$$\langle L_O(t) \rangle \approx \frac{\langle L_H(t) \rangle}{2} + \frac{1}{4}, \quad (36)$$

where the 1/4 term accounts for the discrepancy in the mapping for odd and even paths.

From Steinbock et al.[16], we note that the mean shortest path length for G_H scales as

$$\langle L_H(t) \rangle \sim 2 \cdot \ln(t), \quad (37)$$

which indicates that the hidden network exhibits the small-world property[41]. We have omitted constants which are negligible at large t. Hence, applying the mapping in Eq. (36) and omitting the 1/4 term for simplicity, the mean shortest path length for G_O is given by

$$\langle L_O(t) \rangle \sim \ln(t), \quad (38)$$

indicating that the observed network also exhibits the small-world phenomenon. This mapping is confirmed by simulations.

For interest, we note that for $0 < p < 1$, the shortest paths for the UCM are in general not unique; there may be multiple paths between nodes i and j, which are equally short. Unusually for a non-tree network, all shortest paths are unique in the CCM.

General correlated copying model. The GCCM, is defined analogously to the CCM, starting with observed and hidden networks initialised at $t = 1$. Like the UCM and CCM, the GCCM is a corded node duplication model. For practical reasons, we initialise the graph with three nodes which form a complete graph in G_O, and a wedge in G_H. This ensures that the initial graph contains some edges found in G_H, and some edges found only in G_O.

At $t = t_\alpha$, node α is added to both networks and a single target node, β, is chosen uniformly at random. We label the k_O^β neighbours of β in G_O with the index γ_j. In the observed network, the copied edge (α, γ_j) is formed with probability p_{hid} if the edge $(\beta, \gamma_j) \in E_H$ (inner circle copying), and probability p_{obs} otherwise
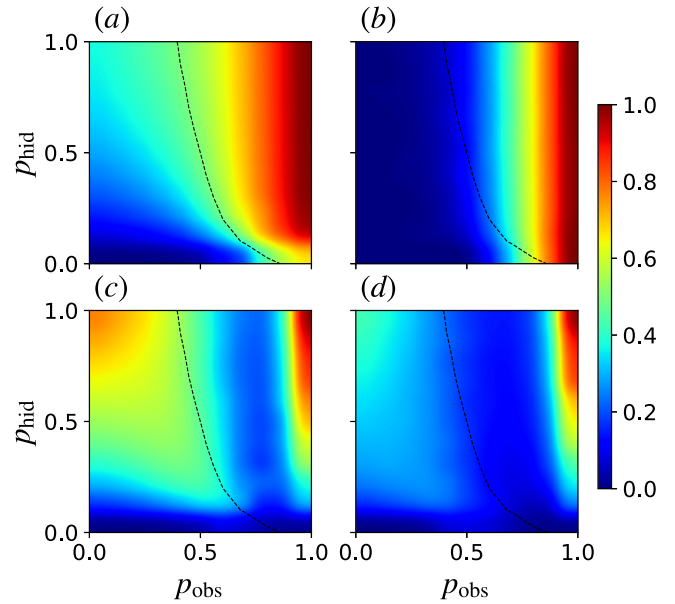

(a) (b) (c) (d)

Fig. 4 Properties of the observed network, G_O, in the general correlated copying model (GCCM). Numerical results for 10^4 nodes as a function of the hidden/inner copying probability, p_{hid}, and the outer copying probability, p_{obs}. **a** The effective copying probability. **b** The densification exponent. **c** The average local clustering coefficient, $CC(G_O)$. **d** The transitivity, τ_{G_O}. Black dashed contour: effective copying probability of 0.5 at $t = 10^4$, calculated numerically. Values have been smoothed for clarity.

(outer circle copying). The direct edge (α, β) is added to both G_O and G_H.

The GCCM encapsulates a wide spectrum of heterogeneous copying. Setting $p_{\text{hid}} = 1$ and $p_{\text{obs}} = 0$ reduces the GCCM to the CCM, whereas setting $p_{\text{hid}} = p_{\text{obs}} = p$ reduces the GCCM to the UCM. We have discussed the social motivation for the case where $p_{\text{hid}} > p_{\text{obs}}$, representing a copying bias towards the inner social circle of a node. However, the GCCM can also be tuned to the reverse case where $p_{\text{hid}} < p_{\text{obs}}$, resulting in a bias against inner circle nodes. We are not aware of a clear physical motivation for this latter case. However, the structural diversity of these anti-correlated networks warrants their discussion here.

Figure 4 shows numerical results for (a) the effective copying probability, (b) the densification exponent, (c) the average local clustering coefficient, and (d) the transitivity, for the GCCM with 10^4 nodes.

The effective copying probability corresponds to the fraction of target node neighbours which appear to be copied in the observed network. Formally, we can write the average effective copying probability at time t as

$$p_{\text{eff}}(t) = \left\langle \frac{p_{\text{hid}} k_H^\beta + p_{\text{obs}}(k_O^\beta - k_H^\beta)}{k_O^\beta} \right\rangle, \quad (39)$$

where β is the index of the target node at time t, the first term represents edges copied from node β's inner circle, and the second term represents edges copied from the outer circle.

The dashed contour in Fig. 4b corresponds to an effective copying probability of 0.5, calculated numerically by averaging over the preceding 10^4 time steps. We note that $p_{\text{eff}} = 0$ if $p_{\text{hid}} = p_{\text{obs}} = 0$ (random tree), $p_{\text{eff}} = 1$ if $p_{\text{hid}} = p_{\text{obs}} = 1$ (complete graph), and $p_{\text{eff}} = p$ if $p_{\text{hid}} = p_{\text{obs}} = p$ (UCM). In general, the rise in p_{eff} is faster with increasing p_{obs} than increasing p_{hid}, although for $p_{\text{hid}} = 0$ we find very small p_{eff}, even for large p_{obs}. However, this observation is somewhat deceptive since, if the GCCM is in the dense regime and $p_{\text{hid}} \neq p_{\text{obs}}$, p_{eff} is not stationary. Calculated

over longer time frames, we note that the effective copying probability appears to slowly converge to the outer copying probability, $p_{eff} \to p_{obs}$, since for $t \to \infty$, the ratio of the number of edges in the hidden network to the number of edges in the observed network tends to zero. This suggests that the dashed $p_{eff} = 0.5$ contour will converge to the $p_{obs} = 0.5$ line as $t \to \infty$.

We test whether the GCCM is in the sparse or dense regime explicitly by tracking the growth in the number of edges in the observed network. Let us define the densification exponent, δ, $0 \le \delta \le 1$, using $E_O(t) \propto t^{1+\delta}$, which relates the number of edges in the observed network to the number of nodes t. If $\delta \approx 0$, the GCCM is sparse. If $\delta = 1$, the GCCM grows as a complete graph. For intermediate values, the GCCM undergoes densification. For the UCM, the transition from the sparse to dense regime is known to take place at $p = 0.5$[3]. We have not analytically calculated the transition for the GCCM, but may intuitively expect the transition at $p_{obs} = 0.5$ since the hidden network is a random tree. This seems to be supported by the numerical values of δ in Fig. 4b, although the transition from zero to non-zero δ is shifted to slightly larger p_{obs} for $p_{hid} = 0$, and to smaller p_{obs} for $p_{hid} = 1$; this shift is likely to disappear as $t \to \infty$.

Figures 4c and d shows the average local clustering coefficient, $CC(G_O)$, and transitivity (global clustering), τ_{G_O}, for the GCCM. Patterns are similar between the two figures, although local clustering generally exceeds global clustering in the sparse regime. For the UCM it is known that, in the dense regime, τ_{G_O} slowly converges to zero as $t \to \infty$, unless $p = 1$[2]. In contrast, the local clustering appears to remain non-zero.

As expected, clustering is minimised at $p_{hid} = p_{obs} = 0$ (random tree) and maximised for a complete graph, $p_{hid} = p_{obs} = 1$. However, in the sparse regime we find that the maximum clustering is found at $p_{hid} = 1$, $p_{obs} = 0$ which corresponds to the CCM. Bhat et al.[2] note that local and global clustering for the UCM is not a monotonically increasing function of the copying probability p, with a local maxima in the sparse regime at non-zero p. This bimodal clustering is also present in the GCCM. In the anti-correlated regime where $p_{hid} \approx 0$, we find near-zero clustering values. In particular if $p_{hid} = 0$ and $p_{obs} = 1$, we observe the unusual property that $\delta \approx 1$, such that the network scales as (but is not) a complete graph, yet both the local and global clustering are approximately zero.

Extracting the degree distributions for the GCCM for various p_{hid} and p_{obs} shows similarly diverse behaviour, see Fig. 5. Each distribution is averaged over 100 instances, but points are left deliberately unbinned to illustrate the significant fluctuations observed in the dense regime. For $p_{hid} = p_{obs} = 0$ (bottom left) the GCCM reduces to a random recursive tree, see Eq. (3). The CCM case with $p_{hid} = 1$, $p_{obs} = 0$ (top left) follows Eq. (14), where the tail can be approximated as a stretched exponential. This distribution is also shown in Fig. 2. Along the diagonal where $p_{hid} = p_{obs}$ (UCM), the degree distribution has a power-law tail in the sparse regime, and exhibits anomalous scaling in the dense regime ($p \ge 0.5$). For $p_{hid} = p_{obs} = 1$, the GCCM reduces to a complete graph and all nodes have degree $t - 1$.

For $p_{obs} = 0$, the power-law scaling observed in the UCM is completely suppressed, with a gradual transition from exponential decay to a stretched exponential tail as p_{hid} is increased from 0 to 1. In the sparse regime with $p_{obs} \ne 0$, all degree distributions appear fat-tailed with only small deviations from the power-laws observed for the UCM. However, unusual scaling is observed for $p_{hid} = 0$, $p_{obs} \ne 0$, where the distributions exhibit initial exponential decay at small k_O, attributable to the hidden network, before a second fat-tailed regime starting at intermediate k_O.

In the dense regime, all degree distributions exhibit anomalous scaling, such that individual instances are not self-averaging. For $p_{obs} = 0.75$, the tail of the degree distributions is largely consistent across all p_{hid}. However, the probability of finding nodes with small degree is large for $p_{hid} = 0$, and is gradually suppressed as $p_{hid} \to 1$. These effects are most pronounced for $p_{obs} = 1$ where the modal degree is 1 for $p_{hid} = 0$, and $t - 1$ for $p_{hid} = 1$, with a gradual transition in between. Throughout this transition, the degree distribution appears almost uniform at $p_{hid} = 0.25$, where the probability of finding nodes with any given degree is approximately constant up until the large k_O limit. However, this effect is only observed when averaging over many instances, with a much smaller degree range observed in individual networks.

It is possible to extend the GCCM further by adding copied edges from G_O to the hidden network, G_H, with probability q. Results are shown in Supplementary Note 4 for $q > 0$ where clustering is enhanced if $p_{hid} > p_{obs}$ and suppressed if $p_{hid} < p_{obs}$, relative to the UCM. In the limiting case of $q = 1$, the GCCM is independent of p_{obs} and equivalent to the UCM with $p = p_{hid}$. The $p_{hid} = p_{obs}$ line (UCM) is invariant under changes in q. One potential application of the $q \ne 0$ case is for generating random simplicial complexes[42] by combining the hidden and observed networks into a single structure. Such a construction may be interesting since it explicitly distinguishes between cliques of strong ties, where all nodes are within each other's inner circle, and cliques of weak ties, see Supplementary Note 5.

Comparing copying models. We have introduced a simple model of heterogeneous node copying, motivated by arguments that triadic closure may not be structurally homogeneous in real networks.

Comparing the CCM, for which we have analytical results, to the UCM with the equivalent effective copying probability ($p = p_{eff} = 0.374$) we find significant differences in network structure. Both the average local clustering coefficient, $CC(G_O)$, and the transitivity, τ_{G_O}, are significantly larger in the CCM than the UCM. The CCM suppresses the power-law tail observed in the UCM for the sparse regime, and consequently, the degree variance observed in the CCM is smaller than for the UCM. CCM: $\sigma^2(k_O) \approx 26$; UCM: $\sigma^2(k_O) \approx 192$. The CCM also has the unusual property, not found in the UCM, that the growth in the number of cliques of size n scales independently of n as $t \to \infty$. For both the UCM and CCM, the mean shortest path lengths scale as $\ln(t)$ indicative of the small-world property.

The above comparison uses a single effective copying probability, but key differences are robust for variable p in the sparse regime. Specifically, the UCM degree distribution always exhibits a power-law tail, and the largest measured clustering coefficients fall below the values seen for the CCM, see Table 1. Relaxing the CCM to the GCCM, we note that for large p_{hid} and small p_{obs}, the measured clustering values regularly exceed those observed in the UCM, with the UCM only reaching similar values far into the dense regime. Given the continuing debate about the ubiquity of power-laws in real networks[43], the observation that power-laws are suppressed in the GCCM as soon as the UCM symmetry is broken supports the view that power-law network scaling is an idealised case which in practice is rarely observed for real networks.

Whether such extreme bias is plausible in real networks is uncertain. However, observations in academic collaboration networks suggest that extreme bias may be possible[33]. For instance, Kim and Diesner[33] show that the ratio of triadic closure between two nodes is approximately zero if the number of shared collaborators is zero, rises rapidly as the number of shared collaborators increases, and plateaus at a ratio of one.
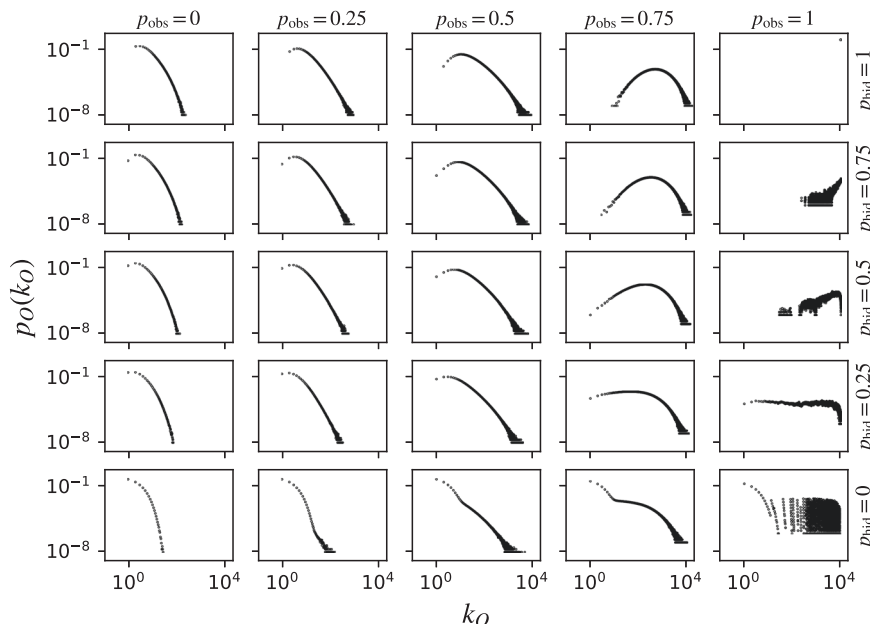
Fig. 5 The observed degree distributions for the general correlated copying model (GCCM). The degree probability, $p_O(k_O)$, is plotted as a function of the observed degree, k_O, for various values of the outer copying probability, p_{obs} (left to right), and the hidden copying probability, p_{hid} (bottom to top). For $p_{obs} \in \{0, 0.25, \text{ and } 0.5\}$, each network contains 10^6 nodes. For $p_{obs} = 0.75$, each network contains 10^5 nodes. For $p_{obs} = 1$, each network contains 10^4 nodes. Distributions are averaged over 100 instances. In the dense regime ($p_{obs} > 0.5$), network growth is non-ergodic leading to anomalous scaling and noisy degree distributions. The distribution at $p_{hid} = p_{obs} = 0$ corresponds to a random recursive tree, see Eq. (3) (exponential decay). The distribution at $p_{hid} = p_{obs} = 1$ corresponds to a complete graph. The distribution at $p_{hid} = 1$, $p_{obs} = 0$ corresponds to the CCM, see Fig. 2. If $p_{hid} = p_{obs}$, the GCCM is equivalent to the uniform copying model.

Table 1 A comparison between the correlated copying model (CCM) and the uniform copying model (UCM).

	p_{eff}	$CC(G_O)$	τ_{G_O}	Degree dist.
CCM	0.37	0.77	0.43	Str. Exp.
UCM (Max $CC(G_O)$)	0.38	0.52	0.20	Power-law
UCM (Max τ_{G_O})	0.22	0.40	0.28	Power-law

UCM is simulated twice, once with an effective copying probability, p_{eff}, that results in the network with the highest average local clustering coefficient (in the sparse regime), $CC(G_O)$, and once with the effective copying probability that gives the largest transitivity, τ_{G_O}. Values averaged over 50 simulations where each network contains 10^5 nodes. Standard deviations are negligible. All values for the observed network, G_O.

Table 2 A selection of sparse undirected networks which may plausibly grow via a copying mechanism.

	Nodes	Edges	$CC(G_O)$	τ_{G_O}
arXiv Astro coauthors	18.8 K	198.1 K	0.63	0.32
arXiv GR coauthors	5.2 K	14.5 K	0.53	0.63
arXiv CM coauthors	23.1 K	93.4 K	0.63	0.26
arXiv HEP coauthors	22.9 K	2.7 M	0.81	0.31
DBLP coauthors	540.5 K	15.2 M	0.80	0.65
NetSci coauthors	379	914	0.74	0.43
Hollywood collaborations	1.1 M	56.3 M	0.77	0.31
DNC email corecipients	906	12.1 K	0.61	0.56

These networks exhibit larger average local clustering coefficients, $CC(G_O)$, and transitivity, τ_{G_O}, than one may expect if these networks were to grow via a uniform copying mechanism. Network source data from Rossi and Ahmed[51].

A second clue towards heterogeneous copying is the observation of very large clustering values in real networks. A selection of these networks and their clustering coefficients are shown in Table 2. Stressing that both the UCM and GCCM are toy models of node copying, the networks in Table 2 exhibit average local and/ or global clustering far exceeding even the most optimistic values for the UCM. In contrast, the listed clustering values are relatively similar to what may plausibly emerge from heterogeneous copying, although even the clustering observed for the extreme CCM case falls below some of the values shown in Table 2. Future work should go beyond this qualitative analysis and should attempt to measure the degree to which copying symmetry is broken for real networks where these mechanisms are relevant.

Discussion

The UCM, CCM, and GCCM are all examples of corded copying models where an edge forms between a newly added node and the target node which is duplicated. This is in contrast to uncorded duplication models where a new node is formed by copying an existing target node and its neighbours, but an edge is not formed between the new node and the duplicated target node. Corded models are more common in the context of social phenomena and triadic closure, whereas uncorded models are typically more relevant to duplication–divergence processes in protein interaction networks. In the current work, we have focused exclusively on corded models; considering heterogeneous copying in uncorded models[13,14,23,24] would be an appropriate future extension. Heterogeneous copying could also be studied by extending directed models[25,26].

The GCCM and CCM are examples of hidden network models. From a mathematical standpoint, hidden network models can be thought of as a variant of interdependent networks where nodes

in one layer have dependencies of nodes in another layer[38,39]. However, at a conceptual level, hidden network models put an emphasis on how the evolution of network structure can depend on asymmetries not observed in our data.

In this paper, we have focused on copying in social networks, but the ideas naturally extend to other contexts. In economics, our framework may be applied to shareholder networks[44], where nodes are connected if they both own a common asset. Here, the hidden network represents the full set of co-owned assets, whereas the observed network includes publicly disclosed assets. Similarly, the idea can be applied to co-bidding networks in public procurement, where an edge indicates that two companies both placed bids on the same contract. In many jurisdictions, only winning bids (of which there may be multiple) are publicly revealed. Therefore, the observed network may represent the network of winning bids, whereas the hidden network includes all bids. Hidden network models may be a valuable representation in these cases if there are structural reasons for why some data is observed and some data is hidden. For instance, fraudulent behaviour in public procurement has been associated with anomalous structural features in the co-bidding network[45].

Other examples may be found in ecology, where multilayer networks have been used to represent different interactions between a common set of species[46,47]. Kéfi et al.[46] find that the structure of interactions in one layer has significant cross-dependencies to the structure of other layers. This mirrors how interlayer dependencies in the CCM are used to break symmetries in the evolution of the observed network. Finally, hidden networks may find general relevance to other fields where interdependent networks have been influential. This may include studies on energy demand management for power grids[48], and the emergence of synchronisation in multilayer neuronal models[49].

A more unusual application of the hidden network concept is for decomposing complex single-layer networks into simpler two-layer structures. One such example is second-neighbour preferential attachment; an implementation of the Barabási–Albert model where nodes attach proportionally to the number of nodes within two steps of a target node[50]. Using our framework, the model is decomposed into an observed network, and a hidden network (in this case referred to as the influence network) where nodes are connected to all nodes which are two or fewer steps away, representing the node's sphere of influence. Here, second-neighbour preferential attachment is equivalent to conventional first-neighbour preferential attachment followed by a local copying step. Structural heterogeneity that is intrinsic in such a model has profound consequences for the time dependence of network growth[50].

Conclusion
We have introduced a general model of heterogeneous copying, implemented using a hidden network model. In the case of extreme copying bias, we have derived analytical results and have demonstrated significant differences to similar models with uniform copying rules. In particular, power-law degree distributions observed in uniform copying can be suppressed under heterogeneous copying, and networks are significantly more clustered if copying is biased towards a node's inner circle. Although a systematic study of copying in real networks is necessary, evidence suggests that heterogeneous copying may be relevant in a social context.

The heterogeneous copying model is just one simple application of a hidden network model. In general, the framework allows us to deconstruct network growth heterogeneities in a non-arbitrary way, focusing on structural rather than node heterogeneity, and poses questions concerning the role of hidden information in network growth. Exploring these questions is a key aim in upcoming work.

References
1. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**, 167 (2003).
2. Bhat, U., Krapivsky, P., Lambiotte, R. & Redner, S. Densification and structural transitions in networks that grow by node copying. *Phys. Rev. E* **94**, 062302 (2016).
3. Lambiotte, R., Krapivsky, P., Bhat, U. & Redner, S. Structural transitions in densifying networks. *Phys. Rev. Lett.* **117**, 218301 (2016).
4. Bhat, U., Krapivsky, P. L. & Redner, S. Emergence of clustering in an acquaintance model without homophily. *J. Stat. Mech.: Theory Exp.* **2014**, P11035 (2014).
5. Bianconi, G., Darst, R. K., Iacovacci, J. & Fortunato, S. Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**, 042806 (2014).
6. Davidsen, J., Ebel, H. & Bornholdt, S. Emergence of a small world from local interactions: modeling acquaintance networks. *Phys. Rev. Lett.* **88**, 128701 (2002).
7. Hassan, M. K., Islam, L. & Haque, S. A. Degree distribution, rank-size distribution, and leadership persistence in mediation-driven attachment networks. *Physica A: Stat. Mech. Appl.* **469**, 23 (2017).
8. Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J. & Kaski, K. A model for social networks. *Physica A: Stat. Mech. Appl.* **371**, 851 (2006).
9. Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K. & Kivelä, M. Cumulative effects of triadic closure and homophily in social networks. *Sci. Adv.* **6**, eaax7310 (2020).
10. Farid, N. & Christensen, K. Evolving networks through deletion and duplication. *New J. Phys.* **8**, 212 (2006).
11. Pastor-Satorras, R., Smith, E. & Solé, R. V. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199 (2003).
12. Chung, F., Lu, L., Dewey, T. G. & Galas, D. J. Duplication models for biological networks. *J. Comput. Biol.* **10**, 677 (2003).
13. Ispolatov, I., Krapivsky, P. L. & Yuryev, A. Duplication-divergence model of protein interaction network. *Phys. Rev. E* **71**, 061911 (2005).
14. Bhan, A., Galas, D. J. & Dewey, T. G. A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486 (2002).
15. Krapivsky, P. L. & Redner, S. Network growth by copying. *Phys. Rev. E* **71**, 036118 (2005).
16. Steinbock, C., Biham, O. & Katzav, E. Distribution of shortest path lengths in a class of node duplication network models. *Phys. Rev. E* **96**, 032301 (2017).
17. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A. Modeling of protein interaction networks. *Complexus* **1**, 38 (2003).
18. Holme, P. & Kim, B. J. Growing scale-free networks with tunable clustering. *Phys. Rev. E* **65**, 026107 (2002).
19. Vázquez, A. Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E* **67**, 056104 (2003).
20. Peixoto, T. P. Disentangling homophily, community structure and triadic closure in networks. Preprint at https://arxiv.org/abs/2101.02510 (2021).
21. Battiston, F., Iacovacci, J., Nicosia, V., Bianconi, G. & Latora, V. Emergence of multiplex communities in collaboration networks. *PLoS One* **11**, e0147451 (2016).
22. Goldberg, S. R., Anthony, H. & Evans, T. S. Modelling citation networks. *Scientometrics* **105**, 1577 (2015).
23. Li, S., Choi, K. P. & Wu, T. Degree distribution of large networks generated by the partial duplication model. *Theor. Comput. Sci.* **476**, 94 (2013).
24. Bebek, G. et al. The degree distribution of the generalized duplication model. *Theor. Comput. Sci.* **369**, 239 (2006).
25. Steinbock, C., Biham, O. & Katzav, E. Analytical results for the distribution of shortest path lengths in directed random networks that grow by node duplication. *Eur. Phys. J. B* **92**, 1 (2019).
26. Steinbock, C., Biham, O. & Katzav, E. Analytical results for the in-degree and out-degree distributions of directed random networks that grow by node duplication. *J. Stat. Mech.: Theory Exp.* **2019**, 083403 (2019).

27. Dunbar, R. I. M. The social brain hypothesis. *Evol. Anthropol.: Issues, News, Rev.* **6**, 178 (1998).
28. MacCarron, P., Kaski, K. & Dunbar, R. Calling Dunbar's numbers. *Soc. Netw.* **47**, 151 (2016).
29. McClain, C. R. Practices and promises of facebook for science outreach: becoming a "nerd of trust". *PLoS Biol.* **15**, e2002020 (2017).
30. Easley, D et al. *Networks, Crowds, and Markets* (Cambridge University Press, 2010).
31. Onnela, J.-P. et al. Structure and tie strengths in mobile communication networks. *Proc. Natl Acad. Sci. USA* **104**, 7332 (2007).
32. Scholz, C., Atzmueller, M., Kibanov, M. & Stumme, G. Predictability of evolving contacts and triadic closure in human face-to-face proximity networks. *Soc. Netw. Anal. Min.* **4**, 217 (2014).
33. Kim, J. & Diesner, J. Over-time measurement of triadic closure in coauthorship networks. *Soc. Netw. Anal. Min.* **7**, 9 (2017).
34. Zhou, L., Yang, Y., Ren, X., Wu, F. & Zhuang, Y. Dynamic network embedding by modeling triadic closure process. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 32 (AAAI Press, 2018).
35. Raducha, T., Min, B. & San Miguel, M. Coevolving nonlinear voter model with triadic closure. *EPL (Europhys. Lett.)* **124**, 30001 (2018).
36. Bianconi, G. *Multilayer Networks: Structure and Function* (Oxford University Press, 2018).
37. Cellai, D. & Bianconi, G. Multiplex networks with heterogeneous activities of the nodes. *Phys. Rev. E* **93**, 032302 (2016).
38. Gao, J., Buldyrev, S. V., Stanley, H. E. & Havlin, S. Networks formed from interdependent networks. *Nat. Phys.* **8**, 40 (2012).
39. Danziger, M. M., Bashan, A., Berezin, Y., Shekhtman, L. M. & Havlin, S. An introduction to interdependent networks. In *International Conference on Nonlinear Dynamics of Electronic Systems* 189–202 (Springer, 2014).
40. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629 (2000).
41. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440 (1998).
42. Battiston, F. et al. Networks beyond pairwise interactions: structure and dynamics. *Phys. Rep.* https://doi.org/10.1016/j.physrep.2020.05.004 (2020).
43. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1 (2019).
44. Bardoscia, M. et al. The physics of financial networks. *Nat. Rev. Phys.* **3**, 490−507 (2021).
45. Wachs, J. & Kertész, J. A network approach to cartel detection in public auction markets. *Sci. Rep.* **9**, 1 (2019).
46. Kéfi, S. et al. Network structure beyond food webs: mapping non-trophic and trophic interactions on Chilean rocky shores. *Ecology* **96**, 291 (2015).
47. Pilosof, S., Porter, M. A., Pascual, M. & Kéfi, S. The multilayer nature of ecological networks. *Nat. Ecol., Evol.* **1**, 1 (2017).
48. Iacopini, I., Schäfer, B., Arcaute, E., Beck, C. & Latora, V. Multilayer modeling of adoption dynamics in energy demand management. *Chaos: Interdisciplinary J. Nonlinear Sci.* **30**, 013153 (2020).
49. Majhi, S., Perc, M. & Ghosh, D. Chimera states in a multilayer network of coupled and uncoupled neurons. *Chaos: Interdisciplinary J. Nonlinear Sci.* **27**, 073109 (2017).
50. Falkenberg, M. et al. Identifying time dependence in network growth. *Phys. Rev. Res.* **2**, 023352 (2020).
51. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. In *AAAI* http://networkrepository.com (AAAI Press, 2015).

Acknowledgements

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-021-00694-1.

Correspondence and requests for materials should be addressed to M.F.

Peer review information *Communications Physics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.