Popularity versus similarity in growing networks

Fragkiskos Papadopoulos,[1] Maksim Kitsak,[2] M. Ángeles Serrano,[3] Marián Boguñá,[3] and Dmitri Krioukov[2]

[1]*Department of Electrical Engineering, Computer Engineering and Informatics,*
Cyprus University of Technology, 33 Saripolou Street, 3036 Limassol, Cyprus
[2]*Cooperative Association for Internet Data Analysis (CAIDA),*
University of California, San Diego (UCSD), La Jolla, CA 92093, USA
[3]*Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*

Supplementary Methods

Here we provide details on the real-world network data used in the main text to validate the popularity×similarity optimization approach. We have considered the AS Internet, the *E.coli* metabolic network, and the web of trust among people extracted from Pretty-Good-Privacy (PGP) data. That is, we have validated our approach against three paradigmatic real networks, from three different domains—technology, biology, and society. We also describe the network mapping method used to infer the popularity and similarity coordinates in the considered real networks, and show that this method yields meaningful results, without overfitting or other artifacts.

I. REAL-WORLD NETWORKS

A. Internet

The Internet data used in Fig. 3(a) of the main text and in Fig. S11 of Section IX is collected and prepared as follows. First, we obtain 11 lists of all the autonomous systems (ASs) observed in a collection of Border Gateway Protocol (BGP) data exactly as described in [31]. These AS lists are linearly spaced in time with the interval of three months: time $t = 0$ corresponds to January 2007, $t = 1$ is April 2007, and so on until $t = 10$, June 2009. We denote the obtained AS lists by L_t. For any pair of t and $t' > t$, we call the ASs present both in L_t and $L_{t'}$ the *old* ASs, and the ASs present only in $L_{t'}$ but not in L_t are called the *new* ASs. The number of ASs in L_0 is 17258, while the numbers of new ASs in $L_{t'}$ with $t' = 1, 2, \ldots, 10$ compared to $t = 0$ are 806, 1614, 2389, 3103, 3973, 4794, 5434, 5843, 6207, and 6426. We then take the Archipelago AS topology [32] of June 2009, available at [33], and for each $t = 0, 1, \ldots, 10$ we remove from it all ASs and their adjacent links that are not in L_t, thus obtaining a time series of historical AS topology snapshots S_t. We then map each S_t to the hyperbolic space as described in Section II, and for each $t = 0, 1, \ldots, 9$ and $t' = t + 1$ we compute the empirical probability $p(x)$ of connections between new and old ASs as a function of hyperbolic distance x between the ASs. To compute $p(x)$, we linearly bin distance x, and show in each bin the ratio of the number of connected ASs to the total number of AS pairs located at hyperbolic distances falling within this bin. To avoid clutter, Fig. 3(a) in the main text shows the results for the first and last pairs of consecutive snapshots, i.e., S_0, S_1 and S_9, S_{10}. The number of new ASs in each pair is respectively 806 and 259. Similar results hold for all intermediate snapshot pairs. The figure also shows the results of PA emulations. To emulate PA in a snapshot pair, the links adjacent to a new AS are first disconnected from the old ASs to which these links are connected in reality, and then reconnected to old ASs chosen randomly with the normalized probability $\sim k + \bar{k}(\gamma - 2)/2$, where k is the number of connections the AS has to other old ASs, and $\bar{k} = 5.3$, $\gamma = 2.1$, taken from the Internet. The average clustering in the Internet is $\bar{c} = 0.61$, and both $\bar{k}$ and $\bar{c}$ are stable across the considered period. Fig. S1(a) validates effective preferential attachment for the pairs of Internet snapshots considered in Fig. 3(a) of the main text.

B. *E.coli* metabolic network

We use the bipartite metabolic network representation of the *E.coli* metabolism from [34], reconstructed from data in the BiGG database [35, 36], *i*AF1260 version of the K12 MG1655 [37] strain. The bipartite representation differentiates two subsets of nodes, metabolites and reactions, mutually interconnected through unweighted and undirected links, without self-loops or dead end reactions. Reactions that do not involve direct chemical transformations, such as diffusion and exchange reactions, are avoided and isomer metabolites are differentiated. To enhance the resolution of the mapping procedure, currency metabolites are eliminated (h, h2o, atp, pi, adp, ppi, nad, nadh, amo, nadp, nadph), altogether with a few isolated reaction-metabolite pairs and reaction-metabolite-reaction triplets. This leads to a globally connected set of 1512 reactions and 1010 metabolites. Starting from this bipartite network, we construct its one mode projection over the space of metabolites, that is, we consider only metabolites and declare two metabolites as connected if they participate in the same reaction in the original bipartite network. The resulting unipartite network of metabolites has a power law degree distribution with exponent $\gamma = 2.5$, average degree $\bar{k} = 6.5$, and the average clustering is $\bar{c} = 0.48$.

Empirical data for ancestral metabolic networks is not available. However, it has been argued that there exists a direct relation between the evolutionary history of metabolism and the connectivity of metabolites. The hypothesis is that metabolic networks grew by adding new metabolites, such that the most highly connected metabolites should also be the phylogenetically oldest [38–41]. Following this idea, we sorted the network of metabolites by degree to construct an ancestor core metabolic network of 460 metabolites with degrees larger than 4, and two shells including metabolites of degrees 4 and 3, respectively. Each shell is meant to represent the addition of new metabolites in subsequent evolutionary steps. The first shell consists of 142 new metabolites and the second shell of 171 new metabolites. Time $t = 0$ corresponds to the core network S_0. Time $t = 1$ corresponds to the snapshot of the topology S_1 consisting of the metabolites in S_0 and the new metabolites in the first shell. And, time $t = 2$ corresponds to the snapshot of the topology S_2 consisting of the metabolites in S_1 and the new metabolites in the second shell. We map S_0, S_1, S_2 to the hyperbolic space and compute the empirical connection probability, following the same procedure as in the previous subsection for the Internet. As before, we also perform PA emulations. The results are shown in Fig. 3(b) in the main text and in Fig. S1(b), and are very similar to those of Figs. 3(a), S1(a). The data from this section are also used in Fig. S12 of Section IX.

C. PGP web of trust

Pretty-Good-Privacy (PGP) is a data encryption and decryption computer program that provides cryptographic privacy and authentication for data communication [42]. PGP web of trust is a directed network where nodes are certificates consisting of public PGP keys and owner information. A directed link in the web of trust pointing from certificate A to certificate B represents a digital signature by owner of A endorsing the owner/public key association of B. We use temporal PGP web of trust data collected and maintained by Jörgen Cederlöf [43].

The PGP web of trust (WoT) data is analyzed as follows. We consider two closely spaced in time pairs of WoT snapshots taken in April 2003, October 2003, and December 2005, December 2006. For each of the directed graphs we form their undirected counterparts by taking into account only bi-directional trust links between the certificates. For each of the undirected counterparts we isolate its largest connected component. Then, for each pair of the snapshots we identify old and new sets of nodes. As before, the set of old nodes contains all nodes present in both snapshots, while the set of new nodes contains nodes that are present in the newer snapshot and not in the older snapshot. We refer to the obtained undirected connected subgraphs of the WoT, as snapshots S_0, S_1, S_2, S_3, for April 2003, October 2003, December 2005, December 2006, respectively. The numbers of nodes in S_0, S_1, S_2, S_3 are $14367, 17155, 23797, 26701$, while the average degree $\bar{k}$ is $5.3, 6.2, 7.9, 8.1$ and the average clustering is $\bar{c} = 0.47\text{-}0.48$. The degree distribution can be roughly approximated by a power-law with exponent $\gamma = 2.1$, yet we observe some deviations from this power at high degrees, see Fig. S13(a). We map S_0, S_1, S_2, S_3 to the hyperbolic space and compute the empirical connection probability, following the same procedure as in the previous two subsections. As before, we again perform PA emulations. The results are shown in Fig. 3(c) in the main text and in Fig. S1(c), and are very similar to Figs. 3(a), 3(b) and Figs. S1(a), S1(b). The data from this section are also used in Fig. S13 of Section IX.

By using the PGP data as described, we strengthen the social component of the WoT, since we only

consider bi-directional signatures, i.e., pairs of users (owners of PGP keys) who have reciprocally signed each other's keys. This filtering process increases the probability that the connected users know each other, and makes the extracted network a reliable proxy to the underlying social network. We consider the PGP WoT since it is a massive evolving unipartite graph, which represents real social relationships of trust among individuals, and for which complete historical data is available.

II. INFERRING THE POPULARITY AND SIMILARITY COORDINATES

Given a snapshot of a real network consisting of t nodes, we use the Markov Chain Monte Carlo (MCMC) method described in [44] to compute the current radial (popularity) $r_s(t)$ and angular (similarity) θ_s coordinates for each node s in the network. In this section we briefly describe this method. See [44] for further details.

To infer the radial coordinates is relatively easy. In Section IV, we derive the exact relation between the expected current degree $\overline{k_s(t)}$ of node s and its current radial coordinate $r_s(t)$ in the model, $\overline{k_s(t)} \sim e^{r_t - r_s(t)}$, where r_t is the current radius of the hyperbolic disc. Therefore to infer the radial coordinates in a real network, we use the same expression substituting in it the real degrees $k_s(t)$ of nodes instead of their expected degrees.

The inference of the angular coordinates is much more involved. In summary, we first measure the average degree, power-law exponent, and average clustering in the network to determine m, β, and T, and then execute the Metropolis-Hastings algorithm trying to find the angular coordinates that would maximize the probability (or likelihood)

$$\mathcal{L} = \prod_{i<j} p(x_{ij})^{a_{ij}} [1 - p(x_{ij})]^{1-a_{ij}}, \tag{1}$$

$$p(x_{ij}) = \frac{1}{1 + e^{\frac{x_{ij} - r_t}{T}}}, \tag{2}$$

that a given real network with adjacency matrix a_{ij}, $i, j = 1, 2, \ldots, t$, and with given node coordinates defining the hyperbolic distances x_{ij} between nodes, is produced by the model with the measured parameters. The algorithm operates by repeating the following steps:

1. Compute the current likelihood $\mathcal{L}_c$;

2. Select a random node;

3. Move it to a new random angular location;

4. Compute the new likelihood $\mathcal{L}_n$;

5. If $\mathcal{L}_n > \mathcal{L}_c$, accept the move;

6. Otherwise, accept the move with probability $\mathcal{L}_n/\mathcal{L}_c$;

and some manual intervention and guidance are needed for this algorithm to actually succeed in a reasonable amount of computing time [44].

Given a historical series of real network topology snapshots $S_0, S_1, S_2, \ldots$, we first map S_0 to the hyperbolic space exactly as just described, i.e., we compute the current radial coordinate, and angular position of each node. In the previous section, we have considered a series of 11 AS Internet snapshots $S_0, S_1, \ldots, S_{10}$, a series of 3 $E.coli$ metabolic network snapshots S_0, S_1, S_2 and two series of two PGP web of trust snapshots S_0, S_1 and S_2, S_3. For new nodes in consecutive snapshots of a series we compute their hyperbolic coordinates keeping the coordinates of old nodes fixed. That is, once a node appears at some time and gets its coordinates computed, its coordinates never change. Although according to the model the radial coordinates of nodes should increase with time (unless $\gamma = 2$), here, for simplicity, we keep them fixed. This simplification is justified because the difference $\Delta r = \left(\frac{\gamma-2}{\gamma-1}\right) \ln \frac{t+\Delta t}{t}$ in the radial coordinate of every node in snapshots with $t + \Delta t$ and t nodes is not significant in the closely spaced snapshot series that we consider. In particular, the maximum value of Δr in the Internet, metabolic and PGP snapshot series is respectively $0.058, 0.35, 0.032$. Another simplification is that since the old node coordinates are fixed, we compute the angular coordinate

for new nodes i using only their local contributions to the total likelihood in Equation (1), i.e., instead of (1) we use $\mathcal{L}_i = \prod_{j \neq i} p(x_{ij})^{a_{ij}} [1 - p(x_{ij})]^{1-a_{ij}}$.

Fig. 3 in the main text shows that the empirical connection probabilities between new and old nodes in the Internet, *E.coli* metabolic network, and PGP web of trust, follow their theoretical predictions. These results signify that new connections in these networks are established as our popularity×similarity optimization framework predicts.

III.   DISCUSSION OF THE MAPPING METHOD

The number of parameters in the model is large. It is proportional to the network size, since we have to infer coordinates for each node. Therefore a natural question that arises is whether the mapping method described above yields meaningful results. In particular, could it be the case that the good match between empirical and theoretical connection probabilities in Fig. 3 in the main text is due to overfitting?

In this section we show that the inference results are indeed meaningful, since we find strong correlations between inferred coordinates and network-specific node attributes in each considered network. We also compute the logarithmic loss, which is the metric of the inference quality for statistical inference methods based on maximum-likelihood estimation. We show that this quality is good for each considered network, confirming that the results in Fig. 3 are not an (overfitting) artifact. Finally, we provide an example of real network (IMDb), where this quality is poor, and so is the logarithmic loss. Collectively, these results show that the inference method does not suffer from overfitting. In particular, if it were the case, then this method would yield statistically good results for *any* network.

A.   The mapping yields meaningful results

1.   Internet

In [44], where we study Internet routing, we use the method described in Section II to map the Archipelago AS topology of June 2009, used in Section I A. The mapping yields meaningful results, since ASs belonging to the same country are mapped close to each other, see Figs. 3 and 5 in [44]. More precisely, one can see from Figs. 3,5 in [44] that for the majority of countries, their ASs are localized in narrow angular regions. That is, even though the mapping method is completely geography-agnostic, it discovers meaningful groups or communities of ASs belonging to the same country.

The reason for this effect is that ASs belonging to the same country are usually connected more densely to each other than to the rest of the world, and the method correctly places all such ASs in narrow regions close to each other. We can also see from Fig. 3 in [44] that in many cases, geographically or politically close countries are located close to each other on the circle. These results prove that the angular coordinates inferred by the method reflect reality well, as is the case with the other two networks that we consider here.

2.   E.coli metabolic network

Distances in metabolic network maps give a measure of the chemical potential of metabolites to participate jointly in reactions, such that higher reaction likelihoods are naturally associated to metabolites which are closer in the underlying space. It is then expected that metabolites participating in reactions in the same biochemical pathway would cluster in specific regions in the inferred space.

This was indeed observed in [34] for the cartographic network representation of the metabolism of *E.coli*, see Figs. 2, 3 and 4 there. The geometric embedding of the metabolic network in Fig. 2 in [34], obtained by the same mapping method we use in this paper, shows that metabolites that participate jointly in reactions are mapped close to each other, i.e., in the same angular regions. In particular, pathways—classically understood as chains of step-by-step reactions which transform a principal chemical into another—are in general strongly localized, even though some adopt either a discrete bi-modal or a multi-peaked form, and only a very small fraction transversally spread over the circle (Fig. 3 in [34]). Furthermore, pathways in related functional categories tend to concentrate into well defined sectors (Fig. 4 in [34]). Therefore our model discriminates well the concentrated pathways, most frequent and consistent with the classical view of modular subsystems,

from others, formed of subunits, and even from those responsible of producing or consuming metabolites used extensively in many other pathways.

3. PGP web of trust

To see if the mapping method yields meaningful results in the case of the PGP web of trust (WoT), we consider its mapped topology of April 2003 from Section I C. For each node (PGP certificate) in the topology, the data we use [43] contain the email address of the corresponding owner of the PGP certificate. For each email address we can identify the top-level domain that the email address belongs to, which is the last part of the email address. Examples of top-level domains are .com, .net, .org, .de, .fr., .it and other country codes. Therefore, for each node in the PGP network we can identify the top-level domain that the node belongs to. If the inferred angular coordinates reflect reality well, then we expect PGP nodes belonging to the same country code to be mapped to angular locations close to each other, since in general people in the WoT are expected to trust other people from their own country more. By contrast, we do not expect this to be the case for generic top-level domains such as .net. In Fig. S15 we show that the mapping method indeed yields meaningful results, as expected.

B. Overfitting considerations and logarithmic loss

1. The number of parameters versus the number of predictions

In general, a statistical inference method may suffer from overfitting if the number of parameters in the model is comparable or larger than the number of predicted parameters. Here we show that for any reasonably sized network, the former is much smaller than the latter in our model.

Indeed, given a snapshot of a real network consisting of t nodes, the mapping method in Section II finds the angular coordinate of every node in the network such that the likelihood that the network is produced by the model is maximized. That is, if there are t nodes in the network the method infers t parameters (angular coordinates).

However, we stress that if a network consists of t nodes, then there are $O(t^2)$ node pairs in the network, and for each node pair $i, j \leq t$, the model predicts an independent probability of the existence of a link between this node pair $p(x_{ij})$. If the mapping is successful, then *every* node pair $i, j \leq t$ is placed at the right hyperbolic distance. In other words, if the fitting is successful, then with *just* t parameters, the model manages to successfully make $O(t^2)$ predictions.

If we consider new nodes in a subsequent snapshot, the method infers their angular coordinates such that they are all placed at the right hyperbolic distances with respect to the old nodes. If there are Δt new nodes and t existing nodes, and the mapping is successful, then with *just* Δt parameters, the model makes $O(t\Delta t)$ predictions.

If a real network is well described by the model, then the fitting of the large number of unknowns with a significantly smaller number of parameters is expected to be successful, as Fig. 3 in the main text illustrates. However, to compute the empirical connection probability in Fig. 3, we have to bin the hyperbolic distances into a small number of bins to have statistically reliable results for ratios of the number of connected node pairs to the total number of node pairs at distances within each bin. Instead of a large ensemble of graphs generated with the same parameters, we have only one real network, and we do not have any other method to compute the empirical connection probability for it. Therefore it is desirable to assess the mapping quality using an appropriate metric independent of any binning. Such metric for maximum-likelihood inference methods is logarithmic loss.

2. Logarithmic loss

In general, the logarithmic loss [45] is defined as

$$L \equiv -\log \mathcal{L}, \tag{3}$$

Network Name	L	$\tilde{L}, \delta = 0.05$	$\tilde{L}, \delta = 0.1$	L_{rand}	$\mathcal{L}/\mathcal{L}_{rand}$
Internet (April 2007)	1.4×10^5	1.8×10^5	2.3×10^5	2.7×10^5	$\exp(1.3 \times 10^5)$
$E.coli$ metabolic (S_0)	7.3×10^3	9.0×10^3	9.6×10^3	1.4×10^4	$\exp(6.7 \times 10^3)$
PGP web of trust (April 2003)	6.9×10^4	1.7×10^5	2.4×10^5	3.0×10^5	$\exp(2.3 \times 10^5)$

TABLE I: Logarithmic losses calculated with the inferred coordinates, distorted angular coordinates with $\delta = 0.05, 0.1$ radians, and fully randomized angular coordinates, for the Internet (April 2007 snapshot), $E.coli$ metabolic network (S_0 snapshot), and PGP web of trust (April 2003 snapshot). The last column shows the ratios of likelihoods $\mathcal{L}/\mathcal{L}_{rand} = \exp(L_{rand} - L)$, which are the ratios of the probability that the network is produced by the model with the inferred angular coordinates, to the same probability with all these coordinates being random.

where $\mathcal{L}$ is likelihood. Since maximum-likelihood inference methods operate by maximizing likelihood, logarithmic loss is a natural metric of the quality of the results that these methods produce. Specifically, if the results are good, then logarithmic loss is small. To estimate how small is "small" here, one usually compares against the case with random parameter assignments.

In our case, the likelihood $\mathcal{L}$ is defined in Equation (1). That is, for a given real network and a given set of inferred coordinates, the logarithmic loss is

$$L \equiv -\sum_{i \neq j} [a_{ij} \log [p(x_{ij})] + (1 - a_{ij}) \log [1 - p(x_{ij})]], \qquad (4)$$

where the above sum goes over all $O(t^2)$ pairs of nodes i, j, where t is the network size. That is, we stress that logarithmic loss depends on all the $O(t^2)$ predicted probabilities $p(x_{ij})$. The logarithmic loss is nothing but the absolute value of the logarithm of the probability that the network is generated by the model, given the set of inferred node coordinates.

We compute logarithmic losses for the Internet, $E.coli$ metabolic network, and the PGP web of trust, with the node coordinates inferred by our mapping method. We contrast these logarithmic losses against those obtained for the same networks with random angular coordinates. That is, we first assign to each node an angular coordinate drawn uniformly at random from $[0, 2\pi]$. The randomized logarithmic loss is then

$$L_{rand} \equiv -\sum_{i \neq j} [a_{ij} \log [p(\tilde{x}_{ij})] + (1 - a_{ij}) \log [1 - p(\tilde{x}_{ij})]], \qquad (5)$$

where $\tilde{x}_{ij}$ is the hyperbolic distance between nodes i and j with random angular coordinates. The smaller the L compared to L_{rand}, the better the quality of the mapping, i.e., the better our model describes a given real network.

To test the robustness of the inferred coordinates we also calculate logarithmic losses after distorting inferred angular coordinate θ_i to

$$\tilde{\theta}_i = \theta_i + \delta\epsilon, \qquad (6)$$

where $\delta = 0.05, 0.1$ radians and ϵ a random variable drawn uniformly from the interval $[-1, 1]$.

The logarithmic loss values are reported in Table I. From the table we observe that the logarithmic losses calculated using the inferred angular coordinates are significantly smaller than those with random angular coordinates, indicating that the considered real networks are well described by our model, corroborating the results in Fig. 3 of the main text.

We also compute logarithmic losses considering only the links between new and old nodes. That is, given two consecutive snapshots of a network S_t and S_{t+1}, we define the logarithmic loss as

$$L \equiv -\sum_{i,j} [a_{ij} \log [p(x_{ij})] + (1 - a_{ij}) \log [1 - p(x_{ij})]], \qquad (7)$$

where summation is now over only $O(t\Delta t)$ new-old node pairs, and where t is the number of old nodes, and Δt the number of new nodes. Again, the logarithmic losses using the inferred angular coordinates of new nodes are significantly smaller than those obtained using randomized angular coordinates, see Table II, signifying that new connections in these networks are well described by the popularity×similarity optimization.

Network Name	L	L_{rand}	$\mathcal{L}/\mathcal{L}_{rand}$
Internet (Jan-Apr 2007)	1100	1800	exp(700)
$E.coli$ metabolic (S_0-S_1)	400	600	exp(200)
PGP web of trust (Apr-Oct 2003)	5900	9000	exp(3100)

TABLE II: Logarithmic loss calculated only for new-old pairs of nodes.

C. Example of a network that is not well described by the model

We finally present an example of a real network for which our mapping method does not produce good results. Specifically, we consider the actor network from the Internet Movie Database (IMDb) [46]. To build the network we connect two actors if they have co-starred in at least one film, limiting our consideration only to films labeled as comedies. The largest connected subgraph of the resulting network in the year of 2000 consists of 44936 actors and has average degree $\bar{k} = 13.6$, and average clustering $\bar{c} = 0.55$. This actor network is another example of a growing network with strong clustering and heterogeneous node degrees. However, the mapping of this network using our method is poor, as illustrated by the connection probability in Fig. S16.

We also compute the logarithmic loss for this network using the inferred node coordinates. The result is $L = 6.0 \times 10^6$. This value is larger than the one obtained after randomizing the node angular coordinates, $L_{rand} = 3.8 \times 10^6$, so that $\mathcal{L}/\mathcal{L}_{rand} = \exp(L_{rand} - L) = \exp(-2.2 \times 10^6)$.

The reason why our model does not describe the actor network well is the following. By construction, the network is overinflated with fully connected subgraphs, since many modern film crews include hundreds of dissimilar actors. Any pair of such actors participating at least once in such a large-scale film project, are connected, leading to an abundance of large cliques in the network. As a result of this overinflation, even not so famous actors that may join the project coming from many different countries, have high chances to be connected. That is, connections in this network are not well described by popularity×similarity optimization, because even fairly dissimilar and unpopular actors may be connected with high probability. Therefore, the fact that this network cannot be successfully mapped by our method is quite expected.

Supplementary Notes

IV.   THE POPULARITY×SIMILARITY MODEL: ANALYSIS AND SIMULATIONS

In this section we discuss the formulation details of the popularity×similarity model, analyze its properties, and verify them in simulations. We start with the simplest version of the model: (1) initially the network is empty; (2) at time $t \geq 1$, new node t appears having coordinates (r_t, θ_t), where $r_t = \ln t$, while θ_t is uniformly distributed on $[0, 2\pi]$, and every existing node s, $s < t$, moves increasing its radial coordinate according to $r_s(t) = \beta r_s + (1 - \beta)r_t$ with parameter $\beta \in [0, 1]$; and (3) node t connects to the m hyperbolically closest nodes s, $s < t$; at early times $t \leq m$, node t connects to all the existing nodes. The value of m controls the average degree in the network $\bar{k} = 2m$. The hyperbolic distance between two points (r_s, θ_s) and (r_t, θ_t) is given by [47]

$$
\begin{aligned}
x_{st} &= \frac{1}{2}\operatorname{arccosh}\left(\cosh 2r_s \cosh 2r_t - \sinh 2r_s \sinh 2r_t \cos\theta_{st}\right) \\
&\approx r_s + r_t + \ln(\theta_{st}/2), \quad \text{where } \theta_{st} = \pi - |\pi - |\theta_s - \theta_t||.
\end{aligned}
$$

This expression gives the distance between two points on the hyperbolic plane of curvature $K = -4$ [47]. The model can be generalized for any curvature value (see Section VI) without affecting the results since changing the value of curvature corresponds to simple rescaling of all distances, thus preserving the distance-induced ordering of nodes, e.g., the sets of m closest nodes, etc. We call the above model Model$_1$.

We show in Section IV B that clustering is strongest possible in the networks generated by Model$_1$. To weaken clustering we allow connections to nodes farther apart. To do so, we modify step (3) of Model$_1$ as follows: (3) new node t picks a randomly chosen node s, $s < t$, and given that it is not already connected to it, it connects to it with probability $p(x_{st}) = 1/[1 + e^{(x_{st} - R_t)/T}]$, where parameter T is called network temperature, and $R_t \sim r_t$—the exact value of R_t is specified below. Node t repeats this step until it gets connected to m nodes. The connection probability $p(x_{st})$ is nothing but the Fermi-Dirac distribution [48]. We call this model Model$_2$.

We also show in Section IV B that clustering is a decreasing function of temperature, and that at zero temperature we recover the strongest clustering case, where new nodes connect to the hyperbolically closest existing nodes. But first we show that for any $\beta \in (0, 1)$ both models produce scale-free networks with the power-law degree distribution identical to the degree distribution in networks growing according to preferential attachment (PA) [49], and having power-law exponent $\gamma = 1 + \frac{1}{\beta}$.

A.   Degree distribution

We start with Model$_1$. Consider new node t, let R_t be the radius of a hyperbolic disc centered at this node, and let it connect to all nodes s, $s < t$, that lie within this disc. The probability that there is a connection to node s is

$$
P\left[x_{st} \leq R_t\right] = P\left[\theta_{st} \leq 2e^{-(r_s(t) + r_t - R_t)}\right] \approx \frac{2}{\pi}e^{-(r_s(t) + r_t - R_t)}. \tag{8}
$$

The average number of existing nodes lying within R_t is

$$
\overline{N(R_t)} = \int_1^t P(x_{it} \leq R_t)di = \frac{2}{\pi}e^{-(r_t - R_t)}\int_1^t e^{-r_i(t)}di = \frac{2}{\pi}e^{-(r_t - R_t)}\frac{1}{1 - \beta}\left(1 - e^{-(1-\beta)r_t}\right). \tag{9}
$$

Therefore

$$
R_t = r_t - \ln\left[\frac{2}{\pi}\frac{\left(1 - e^{-(1-\beta)r_t}\right)}{\overline{N(R_t)}(1 - \beta)}\right], \tag{10}
$$

is the radius of the hyperbolic disc centered at node t, which contains on average the closest $\overline{N(R_t)}$ existing nodes. Setting $\overline{N(R_t)} = m$ and substituting R_t from Equation (10) into Equation (8), we find the probability that an existing node that appeared at time s attracts a link from a new node t, if node t connects on average to the m closest existing nodes

$$\Pi(r_s(t)) = P(x_{st} \leq R_t) = \frac{m}{\frac{1}{1-\beta}\left(1 - e^{-(1-\beta)r_t}\right)} e^{-r_s(t)}. \tag{11}$$

The above equation also holds if the new node t always connects to exactly m closest nodes. Further, since $\int_1^t e^{-r_i(t)} di = \frac{1}{1-\beta}\left(1 - e^{-(1-\beta)r_t}\right)$, we can rewrite Equation (11) as

$$\Pi(r_s(t)) = m \frac{e^{-r_s(t)}}{\int_1^t e^{-r_i(t)} di} = m \frac{e^{-(\beta r_s + (1-\beta)r_t)}}{\int_1^t e^{-(\beta r_i + (1-\beta)r_t)} di} = m \frac{\left(\frac{s}{t}\right)^{-\beta}}{\int_1^t \left(\frac{i}{t}\right)^{-\beta} di} \equiv \Pi_{\text{Model}_1}(s,t). \tag{12}$$

We now recall how connections are made in PA [49], where at sufficiently large times t, an existing node s with degree $k_s(t)$ attracts a link from a new node t with probability

$$\Pi(k_s(t)) = m \frac{k_s(t) - m + A}{(m + A)t}, \tag{13}$$

where m is the number of existing nodes that each new node connects to, $A = (\gamma - 2)m$ is a parameter called initial attractiveness of each node, and γ is the exponent of the target power law degree distribution. Notice that since each new node brings m connections, at large times t the denominator in Equation (13) can be written as

$$(m + A)t = \int_1^t (\overline{k_i(t)} - m + A) di. \tag{14}$$

Further, it has been shown [49] that

$$\overline{k_s(t)} = m + A \left[\left(\frac{s}{t}\right)^{-\beta} - 1 \right], \tag{15}$$

where $\beta = \frac{1}{\gamma - 1}$, $\beta \in (0,1)$.

The connection probability given by Equation (13) is conditioned on the exact value of the degree of the node s, $k_s(t)$. Therefore, the unconditional probability that an existing node s attracts a link from a new node t, which can be obtained by Equation (13) after replacing $k_s(t)$ with its expected value, is

$$\Pi(\overline{k_s(t)}) = m \frac{\overline{k_s(t)} - m + A}{\int_1^t (\overline{k_i(t)} - m + A) di} = m \frac{\left(\frac{s}{t}\right)^{-\beta}}{\int_1^t \left(\frac{i}{t}\right)^{-\beta} di} \equiv \Pi_{\text{PA}}(s,t). \tag{16}$$

From Equations (12) and (16) we conclude that

$$\Pi_{\text{Model}_1}(s,t) = \Pi_{\text{PA}}(s,t). \tag{17}$$

This means that for fixed m and $\beta = \frac{1}{\gamma - 1}$ the probability that an existing node s, $s < t$, attracts a link from a new node t, is the same in Model_1 and PA. This, in turn, means that the resulting degree distribution in Model_1 is identical to PA, i.e., it is the same power law with exponent $\gamma = 1 + \frac{1}{\beta}$, whose exact expression is given by [49]

$$P(k) = (\gamma - 1) \frac{\Gamma[(m+1)(\gamma - 2) + 1]\Gamma[k + m(\gamma - 3)]}{\Gamma[m(\gamma - 2)]\Gamma[k + m(\gamma - 3) + \gamma]}. \tag{18}$$

Further, knowing the current degree of a node k, the node attracts a link from a new node t with probability as in Equation (13)

$$\Pi(k) = m \frac{k - m + A}{(m + A)t}. \tag{19}$$

Probabilities $P(k)$ and $\Pi(k)$ are both defined for $k \geq m$. Finally, using Equation (15), we can deduce that

$$\overline{k_s(t)} = m + A\left[e^{-(r_s(t)-r_t)} - 1\right] \sim e^{-(r_s(t)-r_t)}. \tag{20}$$

In contrast to PA where the case $\gamma = 2$ is problematic [49], there are no problems with $\gamma = 2$ in Model_1, where $\gamma = 2$ corresponds to $\beta = 1$, i.e., to the case where nodes do not move. It is easy to check that for $\beta \to 1$, $\int_1^t e^{-r_i(t)} di = \frac{1}{1-\beta}\left(1 - e^{-(1-\beta)r_t}\right) \to r_t$, and Equations (9), (10), and (11) are all well defined.

We now move to Model_2, and show that the same results with respect to the degree distribution hold there as well. Recall that in Model_2 a new node t, instead of connecting to the m closest nodes, picks a random existing node s, $s < t$, and given that it is not already connected to it, it connects to it with probability $p(x_{st}) = 1/[1 + e^{(x_{st} - R_t)/T}]$. It then repeats this procedure until it gets connected to m nodes. Notice that at long times $t \gg m$, the probability that node t selects a random node s to which it is already connected, is insignificant and can be ignored to ease analysis. Further, notice that the probability $p(x_{st})$ can be also written as

$$p(x_{st}) = \frac{1}{1 + \left(X(s,t)\frac{\theta_{st}}{2}\right)^{\frac{1}{T}}}, \quad \text{where } X(s,t) = e^{(r_s(t)+r_t-R_t)}. \tag{21}$$

Since node t picks a random existing node and θ_{st} is uniformly distributed in $[0, \pi]$, the probability that node t connects to node s is

$$P(s,t) = \frac{1}{t}\frac{1}{\pi}\int_0^\pi \frac{1}{1 + \left(X(s,t)\frac{\theta_{st}}{2}\right)^{\frac{1}{T}}} d\theta_{st} \approx \frac{2T}{t\sin T\pi}\frac{1}{X(s,t)}. \tag{22}$$

The approximation in Equation (22) holds for $T < 1$. Now, the probability that node t connects to any node is

$$P(t) = \int_1^t P(i,t)di. \tag{23}$$

Since node t brings m new links, then at sufficiently large times t, the probability that node s attracts a link is

$$\Pi_{\text{Model}_2}(s,t) = m\frac{P(s,t)}{P(t)} = m\frac{e^{-r_s(t)}}{\int_1^t e^{-r_i(t)}di} = \Pi_{\text{Model}_1}(s,t) = \Pi_{\text{PA}}(s,t). \tag{24}$$

This means that for fixed m and $\beta = \frac{1}{\gamma-1}$, the degree distribution and link attraction probability in Model_2 are the same as in Model_1, i.e., given by Equations (18) and (19). The limit $\beta \to 1$ is also well defined.

Notice that as $T \to 0$, $p(x_{st}) \to 1$ if $x_{st} \leq R_t$, and $p(x_{st}) \to 0$ if $x_{st} > R_t$. In this case, setting R_t as in Equation (10) with $\overline{N(R_t)} = m$, constrains the connections of a new node t to its m hyperbolically closest nodes, and Model_2 becomes identical to Model_1. In Model_2, we can also compute the average number of existing nodes lying within R_t from a new node t

$$\overline{N(R_t)} = tP(t) = \frac{2T}{\sin T\pi}e^{-(r_t-R_t)}\frac{1}{1-\beta}\left(1 - e^{-(1-\beta)r_t}\right). \tag{25}$$

Therefore, in analogy to Model_1, setting $\overline{N(R_t)} = m$ we can fix R_t

$$R_t = r_t - \ln\left[\frac{2T}{\sin T\pi}\frac{\left(1 - e^{-(1-\beta)r_t}\right)}{m(1-\beta)}\right]. \tag{26}$$

Equation (26) is valid for $0 < T < 1$, and for $T \to 0$ it becomes Equation (10) as expected.

Fig. S2 shows simulation results for Model_2, and Fig. 2(a) in the main text with Fig. S3(a) show simulation results for Model_1, validating our analysis. Fig. S3(b) also shows that clustering is strong in networks growing according to popularity×similarity optimization, as opposed to PA. We study clustering in the next section.

B. Clustering

We have shown that networks grown according to popularity×similarity optimization have an effective hyperbolic geometry underneath, from which power-law degree distributions emerge. We now show that the metric property of this geometry, i.e., the triangle inequality, leads to strong clustering in these networks, i.e., the large number of triangular subgraphs.

Intuitively, if node a is hyperbolically close to a node b, and b is close to a third node c, then a is also close to c because of the triangle inequality. Since all three nodes are close to each other, links between all of them forming triangle abc exist with high probability. This probability depends on the value of the temperature $T \in [0, 1)$.

We show next that average clustering at time t, $\bar{c}(t)$, is a decreasing function of temperature: clustering is maximized at $T = 0$, and it gradually decreases to zero as $T \to 1$.

1. Analysis

Let $\bar{c}(s, t)$ be the average clustering of node s at time t. Then

$$\bar{c}(t) = \frac{1}{t} \int_1^t \bar{c}(s, t) ds. \tag{27}$$

where $\bar{c}(s, t)$ is given by [50]

$$\bar{c}(s, t) = \frac{2\overline{T_s(t)}}{[\overline{k_s(t)}]^2}, \tag{28}$$

and $\overline{T_s(t)}$ is the expected number of triangles that contain node s at time t, while $\overline{k_s(t)}$ is s'es expected degree given by Equation (15). To compute $\overline{T_s(t)}$ we break it into two parts: (i) $\overline{T_s^{\mathrm{old}}}$, which is the expected number of triangles formed when node s appeared, i.e., by connections from node s to existing pairs of connected nodes; and (ii) $\overline{T_s^{\mathrm{new}}(t)}$, which is the expected number of triangles formed by new nodes appearing after node s, i.e., by connections from new nodes to old pairs of connected nodes where one of the nodes is node s. Clearly, $\overline{T_s(t)} = \overline{T_s^{\mathrm{old}}} + \overline{T_s^{\mathrm{new}}(t)}$.

The probability that two nodes $s < t$ are connected in Model$_2$ given the hyperbolic distance x_{st} between them, is $m \frac{\frac{1}{t} p(x_{st})}{P(t)} = m \frac{\frac{1}{t} p(x_{st})}{\frac{m}{t}} = p(x_{st})$, i.e., they connect with probability given by Equation (21). Introducing notation $\chi_{st} = X(s, t) \frac{\theta_{st}}{2}$, we can write

$$p(x_{st}) = \frac{1}{1 + \chi_{st}^{\frac{1}{T}}} = \widetilde{p}(\chi_{st}). \tag{29}$$

Since $T < 1$, the function $\widetilde{p}(\chi)$ is integrable

$$I = \int_0^\infty \frac{1}{1 + \chi^{\frac{1}{T}}} d\chi = \frac{T\pi}{\sin T\pi}. \tag{30}$$

Further, since $X(s, t) = e^{(r_s(t) + r_t - R_t)}$, with R_t given by Equation (26) we can also write

$$X(s, t) = \frac{2T}{\sin T\pi} f(s, t), \quad \text{where } f(s, t) = \frac{1 - e^{-(1-\beta)r_t}}{m(1 - \beta)} e^{r_s(t)} = \frac{t^{1-\beta} - 1}{m(1 - \beta)s^{-\beta}}. \tag{31}$$

Now, the probability that three nodes $s, t', t'' < t$ form a triangle, is the probability that the three nodes are connected. Let $\theta_{t'}, \theta_{t''}$ be the angular coordinates of nodes t' and t'' respectively, and θ_s be the angular coordinate of node s. As the angular coordinate is uniformly distributed, we can set without loss of generality

$\theta_s = 0$. Therefore, with $\theta_{t's} = \theta_{st'} = \theta_{t'}, \theta_{t''s} = \theta_{st''} = \theta_{t''}$, and $\theta_{t't''} = \theta_{t''t'} = \theta_{t'} - \theta_{t''}$, it is easy to see that

$$\overline{T_s^{\mathrm{old}}} = \frac{1}{4\pi^2} \int_1^s dt' \int_1^{t'} dt'' \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} d\theta_{t'} d\theta_{t''} \widetilde{p}(|\chi_{t's}|) \widetilde{p}(|\chi_{t''t'}|) \widetilde{p}(|\chi_{t''s}|).$$

$$\overline{T_s^{\mathrm{new}}(t)} = \frac{1}{4\pi^2} \int_s^t dt' \left\{ \int_1^s dt'' \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} d\theta_{t'} d\theta_{t''} \widetilde{p}(|\chi_{st'}|) \widetilde{p}(|\chi_{t''t'}|) \widetilde{p}(|\chi_{t''s}|) \right.$$
$$\left. + \int_s^{t'} dt'' \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} d\theta_{t'} d\theta_{t''} \widetilde{p}(|\chi_{st'}|) \widetilde{p}(|\chi_{t''t'}|) \widetilde{p}(|\chi_{st''}|) \right\}. \tag{32}$$

Changing the θ integration variables in Equation (32) to the corresponding χ variables gives

$$\overline{T_s^{\mathrm{old}}} = \frac{1}{(2I)^2} \int_1^s \frac{dt'}{f(t',s)} \int_1^{t'} \frac{dt''}{f(t'',s)}$$
$$\times \int_{-If(t',s)}^{If(t',s)} d\chi' \int_{-If(t'',s)}^{If(t'',s)} d\chi'' \widetilde{p}(|\chi'|) \widetilde{p}\left(f(t'',t') \left| \frac{\chi'}{f(t',s)} - \frac{\chi''}{f(t'',s)} \right| \right) \widetilde{p}(|\chi''|).$$

$$\overline{T_s^{\mathrm{new}}(t)} = \frac{1}{(2I)^2} \int_s^t \frac{dt'}{f(s,t')}$$
$$\times \left\{ \int_1^s \frac{dt''}{f(t'',s)} \int_{-If(s,t')}^{If(s,t')} d\chi' \int_{-If(t'',s)}^{If(t'',s)} d\chi'' \widetilde{p}(|\chi'|) \widetilde{p}\left(f(t'',t') \left| \frac{\chi'}{f(s,t')} - \frac{\chi''}{f(t'',s)} \right| \right) \widetilde{p}(|\chi''|) \right.$$
$$\left. + \int_s^{t'} \frac{dt''}{f(s,t'')} \int_{-If(s,t')}^{If(s,t')} d\chi' \int_{-If(s,t'')}^{If(s,t'')} d\chi'' \widetilde{p}(|\chi'|) \widetilde{p}\left(f(t'',t') \left| \frac{\chi'}{f(s,t')} - \frac{\chi''}{f(s,t'')} \right| \right) \widetilde{p}(|\chi''|) \right\}, \tag{33}$$

which cannot be written as a closed-form expression. However, these equations allow us to infer the relationship between the clustering strength of the network and parameter T. As $T \to 1$, $I \to \infty$, and therefore, $\overline{T_s^{\mathrm{old}}} \to 0$, $\overline{T_s^{\mathrm{new}}(t)} \to 0$, $\forall s, t$, meaning that clustering goes to zero. As $T \to 0$, $I \to 1$, and clustering is maximized. To see this, consider the node with the smallest degree, i.e., the node that appeared at time $s = t$, whose degree is $k_t(t) = m$. Clearly, $\overline{T_t^{\mathrm{new}}(t)} = 0$. To compute $\overline{T_t^{\mathrm{old}}}$, observe that when $T \to 0$, $\widetilde{p}(\chi) \to \Theta(1 - \chi)$, and therefore, the inner integrals taken over the variables χ', χ'', reduce to the area of intersection of the square defined by $\{|\chi'| < 1; |\chi''| < 1\}$, and the stripe $f(t'', t') \left| \frac{\chi'}{f(t',t)} - \frac{\chi''}{f(t'',t)} \right| < 1$. For most of the combinations of t', t'' the stripe is so wide that it fully contains the square whose area is 4, yielding at large t, $\overline{T_t^{\mathrm{old}}} \approx \frac{m^2}{2}$. Given Equation (28), this means that $\bar{c}(t,t) \approx 1$, proving that clustering is maximized at the zero temperature. Recall that clustering cannot be equal to its maximum possible value of 1 for all node degrees because of structural constraints imposed by power-law degree distributions [51]. For arbitrary values of $s < t$ we need to compute $\overline{T_s^{\mathrm{new}}(t)}$, but the inner integration region defined by the χ', χ'' variables in the expression for $\overline{T_s^{\mathrm{new}}(t)}$ (33) depends on the exact mutual relationship between s, t', and t'', making the analytic computation unfeasible. However, one can check that $\bar{c}(s,t)$ increases as s increases, and that average clustering decreases almost linearly with $T \in (0,1)$.

2. Simulations

Fig. S4 shows average clustering in simulated networks. As predicted by our analysis, clustering decreases as T increases, and vanishes as T approaches 1. Clustering is also the stronger, the smaller the γ.

To confirm that zero temperature yields the strongest *possible* clustering (modulo fluctuations), we perform the following experiment. We grow three networks up to $t = 1000$ nodes according to Model$_1$ with $\gamma = 2.1, 2.5, 3.0$ and $m = 3$. The average clustering in these networks is $\bar{c} = 0.83, 0.76, 0.72$, respectively. For each network we then perform a number of random link rewirings preserving the degree distribution in the network and trying to increase its clustering if possible [52]. Specifically, we select a random pair of links A–B and C–D in the network, and rewire them to A–D and B–C, provided that none of these links already

exist in the network and that the rewiring will not decrease clustering. If these two conditions are met, then the rewiring is accepted, otherwise it is aborted, and a new pair of links is selected. This way each accepted rewiring step preserves the degree distribution in the network, and can only increase its average clustering. For each network we run the experiment until 2000 rewiring steps were accepted, measuring the new average clustering $\bar{c}_{new}$ every 100 accepted rewirings. Fig. S5 shows the results. From the figure, we observe only a minor increase of clustering from its original value, quickly reaching saturation as the number of accepted rewirings increases, as expected. After 2000 accepted rewirings the average clustering is $\bar{c}_{new} = 0.86, 0.81, 0.78$ for $\gamma = 2.1, 2.5, 3.0$.

C.   Connecting to nodes within distance R_t, and densification

We now consider a variant of the popularity×similarity model, where a new node t, instead of connecting to *exactly* m existing nodes as in Model$_2$, looks at every existing node s, $s < t$, only once and connects to it with probability $p(x_{st})$ given by Equation (21). We call this variant Model$_{2'}$. In this case, the probability that node s attracts a link from node t is $\Pi_{\text{Model}_{2'}}(s,t) = tP(s,t)$, where $P(s,t)$ as given by Equation (22). The average number of nodes that node t connects to is $\overline{N(R_t)} = \int_1^t \Pi_{\text{Model}_{2'}}(i,t)di = t\int_1^t P(i,t)di = tP(t)$, with $P(t)$ given by Equation (23). That is, $\overline{N(R_t)}$ is given again by Equation (25) and can be fixed to m by setting R_t as in Equation (26). Further, since $t = \frac{\overline{N(R_t)}}{P(t)} = \frac{m}{P(t)}$, we have

$$\Pi_{\text{Model}_{2'}}(s,t) = m\frac{P(s,t)}{P(t)} = m\frac{e^{-r_s(t)}}{\int_1^t e^{-r_i(t)}di}. \tag{34}$$

That is, Model$_{2'}$ is equivalent to Model$_2$ (cf. Eq. (24)) with the difference that in Model$_{2'}$ a new node t connects on *average* to m existing nodes.

Parameter $T \in [0,1)$ can be used again to tune clustering. As $T \to 0$ a new node t connects only to all nodes within distance R_t from it, and we have a variant of Model$_1$ where clustering is maximized. Indeed, in this case, the probability that node s attracts a link from a new node t is given again by Equation (8), which means that Equations (9), (10), (11) and (12) hold here as well. The difference here is that the new node t connects to closest nodes whose *average* number is m.

To quantify the difference between Model$_2$ and Model$_{2'}$, we need to consider the distribution of the number of existing nodes that a new node t connects to in Model$_{2'}$, and to check how narrowly distributed this number is around its average value m. The connection events are statistically independent, so that the number of connections to existing nodes is a sum of independent Bernoulli trials with different success probabilities $\Pi_{\text{Model}_{2'}}(s,t)$. Hence, the distribution of $N(R_t)$ follows the Poisson-Binomial distribution with average m and variance $\sigma^2(t) \approx \int_1^t \left(1 - \Pi_{\text{Model}_{2'}}(i,t)\right) \Pi_{\text{Model}_{2'}}(i,t)di$. We do not use strict equality in the formula for $\sigma^2(t)$ as we replace the summation with the integration to ease the calculations. Performing the integration we can see that

$$\sigma^2(t) \approx m - g(m,\beta,t), \tag{35}$$

where $g(m,\beta,t)$ a function of m, β, and t that goes to zero as $t \to \infty$. Therefore at $t \to \infty$ the variance $\sigma^2(t)$ approaches m, which is the variance of a Poisson distribution with the average equal to m. Indeed, by Le Cam's Theorem [53] $\sum_{i=0}^{\infty} |P(N(R_t) = i) - \frac{\lambda^i e^{-\lambda}}{i!}| < 2\int_1^t \left(\Pi_{\text{Model}_{2'}}(i,t)\right)^2 di \to 0$ at $t \to \infty$, and therefore the distribution of $N(R_t)$ converges to the Poisson distribution with the average at m.

The simulation results in Fig. S6 confirm the analysis above. Fig. S7 shows the simulation results for the average clustering as a function of temperature, where the behavior is similar to Fig. S4 as expected. Finally, in Fig. S8 we repeat the same experiment with the same parameter values as in Fig. S5, verifying that networks grown according to Model$_{2'}$ with $T = 0$ have maximum possible clustering.

Finally, if the connection disc radius is $R_t = r_t$ instead of Equation (26), then the average degree is not constant $\bar{k} = 2m$, but grows with the network size t, an effect known as network *densification* [54]. Specifically, the average degree in this case is given by

$$\overline{k(t)} \approx \frac{4T}{\sin T\pi}\frac{1}{1-\beta}\left[1 - \frac{1}{t} + \frac{1}{\beta t} - \frac{1}{\beta t^{1-\beta}}\right] \xrightarrow[\beta \to 1]{} \frac{4T}{\sin T\pi}\left[\ln t - 1 + \frac{1}{t}\right], \tag{36}$$

where $\gamma = 1 + \frac{1}{\beta}$ is the exponent of the degree distribution as before. We see that the average degree grows logarithmically with the network size if $\gamma \to 2$. More generally, if $R_t = \delta r_t$ with $\delta \geq 1$, then we have

$$\overline{k(t)} \approx \frac{4T}{\sin T\pi} \frac{1}{1-\beta} \left[\frac{1}{\delta t^{1-\delta}} - \frac{1}{t\delta} + \frac{1}{(\beta+\delta-1)t} - \frac{1}{(\beta+\delta-1)t^{2-\delta-\beta}} \right], \tag{37}$$

so that for large t and $\gamma \to 2$, the average degree grows polynomially with the network size, $\overline{k(t)} \sim t^{\delta-1} \ln t$, if $\delta > 1$. In this case the average shortest path distance and effective diameter do not increase but decrease with the network size, thus reproducing the shrinking diameter effect [54], see Fig. S9.

V. CONNECTION TO THE FITNESS MODEL

In this section we consider the popular fitness model [55] and show that it can be also mapped to our geometric optimization framework.

The main motivation behind the fitness model is that in some real networks the popularity of a node does not depend only on its birth time, but also on its ability (fitness) to compete for links. Examples include the Web, where new sites may attract considerably more links than old ones, social networks where new individuals may have more friends, and citation networks where new research papers may acquire a large number of citations quickly.

To account for the different ability of nodes to compete for links in the fitness model [55], the following attraction probability is introduced

$$\Pi(k_{\eta_s}(t)) = m \frac{\eta_s \left(k_{\eta_s}(t) - m + A\right)}{\int_1^t \eta_i \left(\overline{k_{\eta_i}(t)} - m + A\right) di}, \tag{38}$$

which is a variant of Equation (13). Equation (38) says that the probability that an existing node s, $s < t$, attracts a link from a new node t depends both on the node current degree $k_{\eta_s}(t)$ and on its fitness η_s. Fitness $\eta_s \in (0, \eta_{max}]$ is a parameter assigned to each incoming node s, which remains unchanged in time and follows some distribution $\rho(\eta)$ [55]. Given the fitness of each node, the attraction probability in Equation (38) is conditioned on the exact value of the degree of the node $k_{\eta_s}(t)$, and the unconditional probability can be obtained after replacing $k_{\eta_s}(t)$ with its expected value $\overline{k_{\eta_s}(t)}$. We thus have

$$\Pi_{\text{fitness}}(s, t) = m \frac{\eta_s \left(\overline{k_{\eta_s}(t)} - m + A\right)}{\int_1^t \eta_i \left(\overline{k_{\eta_i}(t)} - m + A\right) di}. \tag{39}$$

Switching to our geometric optimization framework, to account for the fact that the popularity of different nodes can be changing differently with time, we let nodes move with different speeds. That is, our model and its variants remain exactly the same, with the only difference that every existing node s, $s < t$, now drifts away by increasing its radial coordinate using the formula $r_s(t) = \beta(\eta_s)r_s + (1 - \beta(\eta_s))r_t - \ln \frac{\eta_s}{\eta_{max}}$. Parameter $\beta(\eta_s)$ is some function of the fitness of node s, η_s, and therefore its value can be different for different nodes. We call this variant Model$_3$.

Following exactly the same steps as in our earlier analysis, e.g., for Model$_2$, we can see that

$$\Pi_{\text{Model}_3}(s, t) = m \frac{e^{-r_s(t)}}{\int_1^t e^{-r_i(t)} di} = m \frac{\eta_s \left(\frac{s}{t}\right)^{-\beta(\eta_s)}}{\int_1^t \eta_i \left(\frac{i}{t}\right)^{-\beta(\eta_i)} di}, \tag{40}$$

$$\overline{N(R_t)} = \frac{2T}{\sin T\pi} e^{-(r_t - R_t)} \frac{1}{\eta_{max} t} \int_1^t \eta_i \left(\frac{i}{t}\right)^{-\beta(\eta_i)} di,$$

$$R_t = r_t - \ln \left[\frac{2T}{\sin T\pi} \frac{\frac{1}{\eta_{max} t} \int_1^t \eta_i \left(\frac{i}{t}\right)^{-\beta(\eta_i)} di}{m} \right] \quad \text{for } \overline{N(R_t)} = m. \tag{41}$$

Parameter $T \in [0, 1)$ can be used again to tune clustering, and the limit $T \to 0$ is again well defined.

The integral $I(t) = \int_1^t \eta_i \left(\frac{i}{t}\right)^{-\beta(\eta_i)} di$ is in general a random variable that depends on the sequence of η_i's, $i \in (1, t)$, and on the function $\beta(\eta)$. As in [55], we compute the expected value of $I(t)$

$$\overline{I(t)} = \int_1^t \int_0^{\eta_{max}} \eta \left(\frac{i}{t}\right)^{-\beta(\eta)} \rho(\eta) d\eta di \approx tC \quad \text{for large } t, \tag{42}$$

where $C = \int_0^{\eta_{max}} \frac{\eta \rho(\eta)}{1 - \beta(\eta)} d\eta$, and assume that $I(t) \approx \overline{I(t)}$. We then get from Equation (40) that

$$\Pi_{\text{Model}_3}(s, t) \approx \frac{m\eta_s}{tC} \left(\frac{s}{t}\right)^{-\beta(\eta_s)}. \tag{43}$$

Using Equation (43) we compute the average degree of an existing node s at time t, given its fitness η_s

$$
\begin{aligned}
\overline{k_{\eta_s}(t)} &= m + \int_s^t \Pi_{\text{Model}_3}(s, i) di \approx m + \frac{m\eta_s}{\beta(\eta_s)C} \left[\left(\frac{s}{t}\right)^{-\beta(\eta_s)} - 1 \right] \\
&= m + A \left[\left(\frac{s}{t}\right)^{-\beta(\eta_s)} - 1 \right], \quad \text{for } \beta(\eta_s) = \frac{m\eta_s}{AC}.
\end{aligned} \tag{44}
$$

Observe that Equation (44) is similar to Equation (15) with the difference that the exponent is $\beta(\eta_s)$ instead of β, however, we again have $\overline{k_{\eta_s}(t)} \sim e^{-(r_s(t) - r_t)}$. Using Equation (44) in (40) we can see that

$$\Pi_{\text{Model}_3}(s, t) = \Pi_{\text{fitness}}(s, t). \tag{45}$$

This means that for $m, A, \rho(n)$ fixed, and $\beta(\eta) = \frac{m\eta}{AC}$, the probability that node s attracts a link from a new node t is the same between Model$_3$ and the fitness model, which in turn means that the resulting degree distribution is the same. The degree distribution $P(k)$ is a weighted sum of different power laws, which can be computed following the approach in [55]

$$P(k) = \int_0^{\eta_{max}} d\eta \rho(\eta) \frac{C}{m\eta} \left(\frac{A}{k - m + A} \right)^{\frac{AC}{m\eta} + 1}. \tag{46}$$

Note that the attraction probability we consider in Equation (38) is more general than the one used in [55] and degenerates to it when $A = m$. In this case, we see that $\overline{k_{\eta_s}(t)} = m \left(\frac{s}{t}\right)^{-\beta(\eta_s)}$, $\beta(\eta_s) = \frac{\eta_s}{C}$, and $P(k) = \int_0^{\eta_{max}} d\eta \rho(\eta) \frac{C}{m\eta} \left(\frac{m}{k}\right)^{\frac{C}{\eta} + 1}$, as in [55].

We conclude this section with some additional observations. As in [55], we conclude from Equation (44) that the exponent $\beta(\eta_s)$ is bounded, i.e., $0 < \beta(\eta_s) < 1 \; \forall s$, since a node always increases the number of links attached to it with time, $\beta(\eta_s) > 0$, and $\overline{k_{\eta_s}(t)}$ cannot increase faster than t, $\beta(\eta_s) < 1$. This means that $r_s(t) = \beta(\eta_s) r_s + (1 - \beta(\eta_s)) r_t - \ln \frac{\eta_s}{\eta_{max}} > 0$, $\forall s$, as needed. Further, with $\beta(\eta) = \frac{m\eta}{AC}$ and $A = (\gamma - 2)m$, the value of C is computed by the following Equation

$$1 = (\gamma - 2) \int_0^{\eta_{max}} \frac{\rho(\eta)}{\frac{(\gamma - 2)C}{\eta} - 1} d\eta. \tag{47}$$

Since $\beta(\eta) = \frac{\eta}{(\gamma - 2)C} < 1$ the singularity in the above integral is never reached and we also see that $\eta_{max} < (\gamma - 2)C$. Finally, when $\rho(\eta) = \delta(\eta - \tilde{\eta})$, i.e., all fitness equal to some $\tilde{\eta}$, $C = \frac{\gamma - 1}{\gamma - 2}\tilde{\eta}$ and $\beta(\tilde{\eta}) = \beta = \frac{1}{\gamma - 1}$ as expected, since in this case $\Pi_{\text{Model}_3}(s, t) = \Pi_{\text{fitness}}(s, t) = \Pi_{\text{PA}}(s, t)$, i.e., the degree distribution is the same, as if the network was growing according to standard preferential attachment with power-law degree distribution exponent γ.

VI. EXTENSIONS FOR ANY CURVATURE AND TEMPERATURE

The general formula that gives the hyperbolic distance between two points (r_s, θ_s) and (r_t, θ_t) for any value of hyperbolic space curvature $K = -\zeta^2$, $\zeta > 0$ is [47]

$$x_{st} = \frac{1}{\zeta} \text{arccosh}(\cosh \zeta r_s \cosh \zeta r_t - \sinh \zeta r_s \sinh \zeta r_t \cos \theta_{st}) \approx r_s + r_t + \frac{2}{\zeta} \ln(\theta_{st}/2). \tag{48}$$

The popularity×similarity model with $T \in [0, 1)$ can be extended to any $\zeta < \infty$ with the following two simple modifications: (i) the initial radial coordinate of each new node $t \geq 1$ is $r_t = \frac{2}{\zeta} \ln t$ (instead of $\ln t$); and (ii) given the hyperbolic distance x_{st} between new node t and existing node s, node t connects to s with probability $p(x_{st}) = 1/[1 + e^{\zeta(x_{st} - R_t)/(2T)}]$ (instead of $p(x_{st}) = 1/[1 + e^{(x_{st} - R_t)/T}]$). Redoing the analysis for Model$_2$ (or Model$_{2'}$) it is easy to check that exactly the same results hold, and that R_t is now given by the more general formula

$$R_t = r_t - \frac{2}{\zeta} \ln \left[\frac{2T}{\sin T\pi} \frac{\left(1 - e^{-\frac{\zeta}{2}(1-\beta)r_t} \right)}{m(1-\beta)} \right], \tag{49}$$

with the limit $T \to 0$ again well defined. The expected degree of node s at time t in this case is $\overline{k_s(t)} \sim e^{-\frac{\zeta}{2}(r_s(t) - r_t)}$.

The extension for any $T > 1$ is a bit more involved, but we need it for the next section where we consider interesting high-temperature limits. The point $T = 1$ is a phase transition and for $T \geq 1$ the approximation in Equation (22) giving $P(s, t)$ no longer holds. In particular, after performing the change of variables $\chi_{st} = X(s, t)\frac{\theta_{st}}{2}$ as in Equation (29), we see that the corresponding integral (Equation (30)) diverges, and we explicitly have to cut off the integration at the maximum value $X(s, t)\frac{\pi}{2}$. This yields for $T > 1$

$$P(s, t) \approx \left(\frac{2}{\pi} \right)^{\frac{1}{T}} \frac{T}{t(T-1)} \frac{1}{[X(s, t)]^{\frac{1}{T}}}, \tag{50}$$

with $X(s, t) = e^{\frac{\zeta}{2}(r_s(t) + r_t - R_t)}$, $\zeta > 0$. In this high-temperature regime, the model has the same attraction probability and degree distribution as in the low-temperature regime $T < 1$ if the initial radial coordinate of each new node $t \geq 1$ is $r_t = \frac{2T}{\zeta} \ln t$ instead of $r_t = \frac{2}{\zeta} \ln t$, yielding

$$R_t = r_t - \frac{2T}{\zeta} \ln \left[\left(\frac{2}{\pi} \right)^{\frac{1}{T}} \frac{T}{T-1} \frac{1 - e^{-\frac{\zeta}{2T}(1-\beta)r_t}}{m(1-\beta)} \right]. \tag{51}$$

We can now allow $\zeta \to \infty$ if at the same time $T \sim \zeta \to \infty$. The main difference compared to $T < 1$ is that clustering is asymptotically zero for any $T > 1$. We have confirmed this effect and all the expressions in this section in simulations.

VII. CONNECTIONS TO PREFERENTIAL ATTACHMENT, GROWING RANDOM GRAPHS, AND GROWING RANDOM GEOMETRIC GRAPHS

In this section we show that standard PA with asymptotically zero clustering [56], growing random graphs [57] and growing random geometric graphs [58], can all be seen as limiting degenerate cases of popularity×similarity optimization.

To see the connection to standard PA, we need to consider the general formula that gives the hyperbolic distance x_{st} between two points (r_s, θ_s) and (r_t, θ_t) for any value of hyperbolic space curvature $K = -\zeta^2$, $\zeta > 0$, given by Equation (48). By letting curvature go to minus infinity, $\zeta \to \infty$, we transform the hyperbolic space to a tree [47], and kill the θ-dependent term in the expression for x_{st} (48), i.e., the term abstracting the similarity distance. That is, the hyperbolic distance between nodes depends only on their popularity, $x_{st} = r_s + r_t$, as in PA. We can now set $T \sim \zeta$, e.g., $T = \frac{\zeta}{2}$ without loss of generality. This setting yields $r_t = \frac{2T}{\zeta} \ln t = \ln t$, and Equation (51) becomes

$$R_t = r_t - \ln \left[\frac{1 - e^{-(1-\beta)r_t}}{m(1-\beta)} \right]. \tag{52}$$

Further, from Equation (50), the connection probability is now

$$P(s, t) \approx \frac{1}{t} e^{-(r_s(t) + r_t - R_t)}, \tag{53}$$

so that the probability that node s attracts a link from node t is again

$$m\frac{P(s,t)}{P(t)} = m\frac{e^{-r_s(t)}}{\int_1^t e^{-r_i(t)}di} = m\frac{\left(\frac{s}{t}\right)^{-\beta}}{\int_1^t \left(\frac{i}{t}\right)^{-\beta}di} = \Pi_{PA}(s,t). \qquad (54)$$

As before, this means that the degree distribution is a power law with exponent $\gamma = 1 + \frac{1}{\beta}$, but clustering is zero since $T \to \infty$. Fig. S10 shows simulation results validating our analysis.

If we now let $\beta \to 0$, then $\gamma \to \infty$, and the generated networks degenerate to growing random graphs. Indeed, if $\beta = 0$, $r_s(t) = r_t$, $\forall s, t$, i.e., all node pairs have the same popularity as they all lie on the circle of the maximum radius r_t, expanding with time. It is easy to check that the attraction probability is now $m\frac{P(s,t)}{P(t)} = \frac{m}{t-1} \approx \frac{m}{t}$, $\forall s, t$. This probability is similar to the connection probability in classical random graphs $\mathcal{G}_{N,p}$ [57], where each $N(N-1)/2$ pair of N nodes is connected with the same probability $p \approx \bar{k}/N$. The difference is that our graphs are growing, which affects their properties including the degree distribution. The degree distribution in these growing graphs is exponential [59], versus the Poissonian distribution in classical random graphs.

The limit $\beta \to 0$ ($\gamma \to \infty$) also exists at low temperatures $T \in [0, 1)$ with finite clustering controlled by T. In this case, we can check that the attraction probability is still $\frac{m}{t}$, $\forall s, t$, as all nodes are equally popular, but clustering is not zero, as similarity (the angular distance between nodes) matters. When $T = 0$ we have the strongest clustering, and the generated networks degenerate to growing random geometric graphs on the circle. Indeed, we see from Equation (8) that since any two nodes s, t have the same radial coordinate r_t, they are connected only if the distance between them on the circle is less than a constant that depends on t, i.e., t connects to s only if $\theta_{st} \leq 2e^{-(2r_t - R_t)} = \frac{m\pi}{t-1} \approx \frac{m\pi}{t}$.

In equilibrium geometric networks [48], the connections to PA, growing random graphs, and growing random geometric graphs, are, respectively, the connections to the soft configuration model (random graphs with a given expected degree distribution), classical random graphs, and random geometric graphs.

VIII. EXTENSION WITH INTERNAL LINKS

While in some real networks, e.g., citation networks, new connections appear only from new to old nodes, in some other networks, new links may connect pairs of old, previously disconnected nodes. These links are called *internal,* versus *external* links of the previous type. Examples of networks with internal links include the Internet, were existing disconnected ASs may decide to connect at some point, and social networks were existing disconnected individuals may become friends or collaborators. Our geometric optimization framework can be easily extended to account for internal links as we show below.

At each time t, in addition to the m external links introduced by new node t (e.g., using Model$_2$), L internal connections are also created between existing disconnected pairs of nodes. Specifically, a random pair of existing nodes $i, j < t$ is selected, and then connected (given that it is disconnected) with probability $p(x_{ij}) = 1/[1 + e^{(x_{ij} - R_t)/T}]$. The step is repeated until L internal links are created. This procedure is exactly the same as the procedure by which a new node t connects to existing nodes in Model$_2$. The average degree is now $\bar{k} = 2(m + L)$.

Following exactly the same procedure as in Section IV A, and considering any value of hyperbolic space curvature $K = -\zeta^2$, $\zeta > 0$ (Section VI), the probability that existing nodes i, j are selected and connected at time t is

$$P(i,j,t) \approx \frac{2T}{t^2 \sin T\pi} \frac{1}{X(i,j,t)}, \quad \text{where } X(i,j,t) = e^{\frac{\zeta}{2}(r_i(t) + r_j(t) - R_t)}. \qquad (55)$$

The probability that a pair of existing nodes gets connected at time t is $\frac{1}{2}\int_1^t \int_1^t P(i,j,t)didj$. Since L internal links are introduced, the probability that pair i, j attracts a link is

$$\Pi(i,j,t) = 2L\frac{P(i,j,t)}{\int_1^t \int_1^t P(i,j,t)didj}. \qquad (56)$$

Therefore, the probability that node $s < t$ attracts an internal link at time t is

$$\Pi^{\text{internal}}(s,t) = \int_1^t \Pi(s,i,t)di = 2L\frac{e^{-\frac{\zeta}{2}r_s(t)}}{\int_1^t e^{-\frac{\zeta}{2}r_i(t)di}} = 2L\frac{\left(\frac{s}{t}\right)^{-\beta}}{\int_1^t \left(\frac{i}{t}\right)^{-\beta}di}, \qquad (57)$$

which is similar to the probability that node s attracts an external link, with the only difference that here we have the prefactor $2L$ instead of m, see Equation (24). Thus, the total probability that node s attracts a link at time t is the probability that the node attracts an external or an internal link

$$\Pi^{\text{total}}(s,t) = (m + 2L)\frac{\left(\frac{s}{t}\right)^{-\beta}}{\int_1^t \left(\frac{i}{t}\right)^{-\beta} di} = (\bar{k} - m)\frac{\left(\frac{s}{t}\right)^{-\beta}}{\int_1^t \left(\frac{i}{t}\right)^{-\beta} di}. \tag{58}$$

The average degree of node s by time t is now given by

$$\overline{k_s(t)} = m + A'\left[\left(\frac{s}{t}\right)^{-\beta} - 1\right], \tag{59}$$

where $A' = (\bar{k} - m)(\gamma - 2)$. Equation (59) is similar to Equation (15), and is identical to it if $L = 0$ (i.e., if $\bar{k} = 2m$). From Equations (58) and (59) we see that a node of degree k attracts a new link at time t with probability

$$\Pi(k) = (\bar{k} - m)\frac{k - m + A'}{(\bar{k} - m + A')t}. \tag{60}$$

The link attraction probabilities in Equations (60) and (19) are identical if $L = 0$. If $L = \frac{\bar{k} - 2m}{2} > 0$, Equation (60) gives approximately the same probability as Equation (19) for sufficiently large k, i.e., for $k \geq \frac{\bar{k}}{2}$, and the absolute difference between the two probabilities is $\frac{L}{t}$. This observation implies that for a target $\bar{k}$ and $\beta = \frac{1}{\gamma - 1}$ the degree distributions in both cases are nearly identical, and indistinguishable from the degree distribution in networks growing according to standard PA. However, while internal links do not affect the degree distribution, they can affect other topological characteristics, e.g., they can decrease the average distance in the network. We study topological characteristics of networks growing according to popularity×similarity optimization with internal links in the next section.

We conclude this section with some additional notes. First, from the analysis above we see that, similar to external links, PA appears as an emergent effect in the internal link attraction probability as well since a node attracts an internal link with probability which is also proportional to its current degree. Second, temperature T has the same effect on internal connections as on external connections, i.e., smaller values of T increase the probability that hyperbolically close disconnected node pairs get connected, which increases clustering. Finally, the model extension with internal links can be combined with the fitness model extension, described in Section V, as the former does not depend on whether nodes are moving with the same speeds or not. In this combination Equation (57) becomes $\Pi^{\text{internal}}(s,t) = 2L\,e^{-\frac{\zeta}{2}r_s(t)}/\int_1^t e^{-\frac{\zeta}{2}r_i(t)di} = 2L\eta_s\left(\frac{s}{t}\right)^{-\beta(\eta_s)}/\int_1^t \eta_i\left(\frac{i}{t}\right)^{-\beta(\eta_i)} di$, which is similar to Equation (40), and a straightforward analysis as above can be applied.

IX.　PROPERTIES OF REAL-WORLD VERSUS MODELED NETWORKS

In this section we compare several important properties of the real-world networks considered in Section I to the properties of modeled networks growing according to popularity×similarity optimization. Specifically, we consider the following properties:

(a) degree distribution $P(k)$;

(b) average clustering $\bar{c}(k)$ of k-degree nodes;

(c) average degree of neighbors $\bar{k}_{nn}(k)$ of k-degree nodes;

(d) distance distribution $d(l)$, i.e., the distribution of hop lengths l of shortest paths between nodes in the network, or the probability that a random pair of nodes are at the distance of l hops from each other;

(e) average node betweenness $\bar{B}(k)$ of k-degree nodes, which is the average number of shortest paths passing through a k-degree node, normalized by the maximum possible number of such paths.

Property (c) captures *degree correlations* in the network. If $\bar{k}_{nn}(k)$ is an increasing function, then high (low) degree nodes connect, on average, to nodes of high (low) degree, and the network is called *assortative*. Otherwise, nodes of high degree tend to connect to nodes of low degree, and the network is called *disassortative*. Technological and biological networks are usually disassortative, while social networks are usually assortative [60, 61]. Properties (a-c) are *local* statistics reflecting properties of individual nodes and their one-hop neighborhoods, as opposed to *global* properties (d-e) which depend on large-scale organization of the network.

A.　Internet

We take the Archipelago AS Internet topology of June 2009 from Section I A, and compute properties (a)...(e) from above. The network consists of $t = 23748$ nodes, and has $\gamma = 2.1$, $\bar{k} \approx 5$, $\bar{c} = 0.61$. Then we grow a network according to the popularity×similarity model (Model$_{2'}$) up to the same number of nodes as in the real AS Internet, and with the same γ, $\bar{k}$ and $\bar{c}$. We compute the same properties in the resulting network, and compare them to those of the real Internet. The results are shown in Fig. S11, where we observe a good match between the properties of the modeled network and real Internet. This match is even better if we allow for internal connections as described in Section VIII. In this case, each new node connects on average to $m = 1.5$ existing nodes, and at each time $L = 1$ existing disconnected pairs of nodes are connected so that $\bar{k} = 2(m + L) = 5$. With no internal links, $L = 0$ and $m = \bar{k}/2 = 2.5$.

B.　*E.coli* metabolic network

Here we consider the entire network of metabolites from Section I B, and compute properties (a)...(e) for it. Recall that the network consists of $t = 1010$ nodes, and has $\gamma = 2.5$, $\bar{k} = 6.5$, $\bar{c} = 0.48$. We grow a network according to the popularity×similarity model (Model$_{2'}$) up to the same number of nodes as in the metabolic network, and with the same γ, $\bar{k}$, and $\bar{c}$. We use $m = \bar{k}/2 = 3.25$. We compute the same network properties in the resulting network, and compare them to those of the real metabolic network. The results are shown in Fig. S12, where we observe a remarkable match across all five properties.

C.　PGP web of trust

We now take the PGP web of trust snapshot of April 2003 from Section I C, and compute its properties (a)...(e). The network consists of $t = 14367$ nodes, and has $\bar{k} = 5.3$, $\bar{c} = 0.47$. Its degree distribution is shown in Fig. S13(a), where we observe deviations from a clean power law.

This observation motivates us to grow a modeled network using the fitness model extension in Section V, i.e., Model$_3$, which can model non-power-law degree distributions. Recall that in Model$_3$, nodes s, $s < t$, move with different speeds, increasing their radial coordinate according to $r_s(t) = \beta(\eta_s)r_s + (1 - \beta(\eta_s))r_t - \ln \frac{\eta_s}{\eta_{max}}$, where $\beta(\eta_s) \sim \eta_s$, and η_s is the fitness of s. To grow a network according to this model, we need to know $\beta(\eta_s)$, $\forall s \leq t$. Given that $\frac{\beta(\eta_s)}{\beta(\eta_{max})} = \frac{\eta_s}{\eta_{max}}$, we can find $\beta(\eta_s)$ by solving

$$r_s(t) = \beta(\eta_s)r_s + (1 - \beta(\eta_s))r_t - \ln \frac{\beta(\eta_s)}{\beta(\eta_{max})}, \tag{61}$$

since we know $r_t = \ln t$, have $r_s(t)$ inferred in Section II, and can infer r_s as follows. We assume that nodes with smaller current radial coordinates were born earlier, and sort them in the increasing order, thus creating a sequence of current inferred radial coordinates $r_1(t), r_2(t), ..., r_t(t)$ for nodes born at times $s = 1, 2, ..., t$. Nodes for which the current radial coordinate is the same, are assumed to have appeared at the same time. Using $r_s = \ln s$, and setting $\beta(\eta_{max}) = 1$, we have all the ingredients to solve Equation (61) for $\beta(\eta_s)$ for every node $s = 1, 2, ..., t$.

Another peculiarity of the PGP network, compared to the networks considered earlier, is a deviation of the distribution of the inferred angular distances between nodes from the uniform distribution: see Fig. S14 showing these distributions for all the considered real networks. In the PGP network, nodes with small radial coordinates are, on average, at smaller angular distances than what the uniform distribution suggests.

Therefore in growing the modeled PGP network, we use the inferred angular coordinate θ_s for every node $s = 1, 2, ..., t$, even though our analysis in Section V assumes a uniform angular distance distribution.

Figure S13 juxtaposes properties (a)...(e) of a network grown according to Model$_3$ up to $t = 14367$ nodes using the inferred $\bar{\beta}(\eta_s)$'s and θ_s's, temperature $T = 0.2$, $m = 1$, and $L = 1.65$ ($\bar{k} = 2(m + L) = 5.3$), against the corresponding properties of the real PGP snapshot. As with the AS Internet and *E.coli* metabolic network, we also observe a good match between the modeled and real PGP web of trust across all these properties.

To summarize this section, synthetic networks growing according to popularity×similarity optimization reproduce several important structural characteristics of real technological, biological, and social networks. Remarkably, this optimization approach can capture the properties of both disassortative (Figs. S11(c), S12(c)) and assortative (Fig. S13(c)) networks, as well as networks with degree distributions deviating from clean power laws (Fig. S13(a) vs. Figs. S11(a), S12(a)).

X. RELATED WORK

A. Optimization

The work that comes perhaps closest to our approach is by D'Souza et al. [62, 63]. In this work the authors show that PA can emerge in a tradeoff optimization framework requiring only local information. The framework is motivated by how connections in the Internet may take place. Specifically, the motivation is that a new AS may want to establish connections that would minimize the startup costs, while still providing good performance to its users. In the model, a new node is placed on the unit interval where distances abstract the connection fibre costs, and then connects to an existing node minimizing a balance between these costs and the shortest path hop-lengths to the core in the network, the latter abstracting performance in terms of the average delay from the new node to the rest of the network. The authors then focus only on the degree distribution in the graphs produced by this model, showing that with a specific fit of parameters, it matches well the degree distribution of the Internet extracted from the WHOIS data. The basic model studied in this work generates trees, since each incoming node connects to $m = 1$ existing nodes, but the authors suggest at the end that for $m > 1$ the model may lead to some non-zero clustering.

B. PA+similarity information

The fact that similarity between nodes affects the linking probability in networks has been observed, studied, and modeled extensively in the literature [64–74]. Of particular interest are the works by Menczer [67, 68] where he introduces a model for text corpora with linking probability that augments standard PA with document similarity measures. The latter can be the standard cosine similarity for a pair of documents, defined by the normalized count of words common to both documents. The author then shows that this model describes well the degree and similarity distributions in the DMOZ Web data and in a collection of articles published in PNAS. In [67] he also shows that similarity information can help to improve Web navigation, an observation confirmed later in a more abstract context [72], where similarity is modeled by distances on the unit interval. In [74] a modification of the model of [68] is proposed where the linking probability is proportional to the product of the degrees of the documents and their cosine similarity. The authors then show that this model can describe the clustering coefficient in document networks better compared to [68]. In [73] similarity attributes are modeled by vectors in an n-dimensional space. A new node first selects a certain group of existing nodes (community) based on similarity distances between the new and existing nodes. Within the community the attachment then follows standard PA. That is, this model also augments PA with similarity. The authors conclude by showing that the model generates graphs with power-law degree distributions and exponent $\gamma = 3$, and some community structure. No real networks are considered.

C. PA+spatial information

A wider class of models augment PA not with similarity information *per se*, but with some spatial information [75–77], see also Section 4.4 in [78]. In these models, nodes are located in some space, and the linking probability depends not only on node degrees as in standard PA, but also on distances between nodes in the space. If this linking probability decreases with the spatial distance fast enough, then such models generate graphs with strong clustering for a very simple reason: since close nodes have high probability of being connected, then the triangle inequality in the space leads to a large number of triangles in the network. Yet the mechanism responsible for power-law degree distributions in these models is the same PA.

D. Hidden variables

Yet wider and more general class of models, to which our approach actually belongs, are the network models with hidden variables [50, 79]. In these models, some hidden variables are first assigned to nodes, and the linking probability between a pair of nodes is then a function of the values of their hidden variables. For example, in [79] the authors show that a combination of exponentially distributed hidden variables and step-function connection probability leads to power-law degree distributions and strong clustering in modeled networks, while in [50] it is shown that PA itself can be casted as a hidden variable model, where one of the hidden variables is the node birth time.

E. Clustering

A variety of other mechanisms have been proposed to fix the zero-clustering problem with PA. One such mechanism is node activation/deactivation [80] motivated by citation networks. A set of m active nodes is maintained in the model, and the new, initially active node connects to this set by m links. One active node is then deactivated with probability inversely proportional to the node degree. Because of this inverse proportionality imposed by the model, the model effectively implements linear PA. Because the new connections are made to local groups of active nodes, clustering is strong. However, as shown in [81] the model is effectively one dimensional, lacking the small-world property observed in many real networks.

Another popular mechanism enforcing strong clustering is random walks [82, 83]. A new node connects first to a random existing node, and then with some probability to one of its neighbors, and possibly to a neighbor of its neighbor, etc. Clustering is strong because the connections are concentrated in a local neighborhood of the attachment node.

F. Emergent PA

The lack of clustering is not the only problem with standard PA. Another problem is that PA *per se* is simply impossible in a vast majority of real networks because to "implement" PA, the network evolution process must "know" the global current structure of the whole network in order to compute the degree for each node. Since such knowledge is often unavailable in reality, PA must be an emergent phenomenon, i.e., an effective result of some other underlying evolution processes that use only local information. Yet another related problem is that such processes must lead to exactly linear PA, since if the attachment probability is not a linear function of node degree, then the degree distribution in the network is not a power law [84]. Several mechanisms have been proposed to address these two problems as well. The aforementioned random walks, for example, do solve them both because the probabilities of the stationary distribution of a random walk on a graph are linearly proportional to node degrees. Another interesting observation was made in [85] where the authors show that connections based solely on node ranking may lead to power laws, the motivation being that node ranking is a coarser proxy to popularity than the node degree. Yet the simplest and perhaps the first model that addresses the three mentioned concerns with PA—zero-clustering, global knowledge, and linearity—is by Dorogovtsev *et al.* [86]: the new node simply selects a random existing link, and connects to its both ends. Clustering is obviously strong, and linear PA is resurrected because the probability that a random link is attached to a node of degree k is proportional to k. However, this model is clearly a toy model, and there have been no attempts to validate it against any real networks.

G. Discussion

As far as validation is concerned, the model validation methodology is usually limited to generating synthetic graphs according to the model prescription, and comparing one or more of their structural properties, such as the degree distribution, against those in real networks. Remarkably, the core of the network evolution mechanism proposed by a model is quite rarely validated directly, because such validation is either difficult or impossible. In similarity-based models, for example, such validation is difficult because there are too many different similarity measures, and it is usually unclear which one should be used in which case [71, 87], so that cases where model predictions are validated directly against real-world similarity data [67, 68, 74] are rare, and usually limited to specific (types of) networks.

Within our approach, the direct validation of the network evolution mechanism is also difficult but possible. It is possible because we can infer the node coordinates in the generic similarity space as discussed in Section II, and then check if the linking probability in real networks as a function of distances between nodes in this space is close to our model predictions, see Fig. 3 in the main text.

In summary, the salient feature of our approach is that it *simultaneously*:

1. shows that similarity plays an important and *fundamental* role in evolution of complex networks;

2. does so by means of a very simple and *general* geometric model;

3. admits a complete *analytic* treatment;

4. *directly* validates the modeled similarity mechanism and its analytic predictions against drastically different real networks from different domains;

5. reproduces many important *structural* properties of these networks; and

6. resolves all the mentioned concerns with preferential attachment, which appears in the approach as an *emergent phenomenon*.

[31] A. Dhamdhere and K. Dovrolis, in *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement (IMC 2008), Vouliagmeni, Greece, October 20-22, 2008*, edited by K. Papagiannaki and Z.-L. Zhang (ACM, 2008), pp. 183–196, ISBN 978-1-60558-334-1.

[32] K. Claffy, Y. Hyun, K. Keys, M. Fomenkov, and D. Krioukov, in *Proceedings of the 2009 Cybersecurity Applications & Technology Conference for Homeland Security (CATCH 2009), Washington, DC, March 3-4, 2009* (IEEE Computer Society, 2009), pp. 205–211, ISBN 978-0-7695-3568-5, URL `http://www.caida.org/projects/ark/`.

[33] *IPv4 Routed /24 AS Links Dataset*, `http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml`.

[34] M. A. Serrano, M. Boguñá, and F. Sagués, Molecular BioSystems **8**, 843 (2012).

[35] J. Schellenberger, J. O. Park, T. C. Conrad, and B. O. Palsson, BMC Bioinformatics **11**, 213 (2010).

[36] *Bigg database*, `http://bigg.ucsd.edu/`.

[37] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. O. Palsson, Molecular Systems Biology **3**, 121 (2007).

[38] H. J. Morowitz, Complexity **4**, 39 (1999).

[39] D. A. Fell and A. Wagner, Nature Biotechnology **18**, 1121 (2000).

[40] D. A. Fell and A. Wagner, in *In Animating the cellular map, J.-H.S. Hofmeyr, J.M. Rohwer, & J. L. Snoep, eds., Stellenbosch University Press, Stellenbosch, South Africa, pp. 79-85.* (2000).

[41] A. Wagner, Complexity **8**, 15 (2002).

[42] *The Open PGP Alliance*, `http://www.openpgp.org/`.

[43] *OpenPGP web of trust database*, `http://www.lysator.liu.se/~jc/wotsap/wots2/`.

[44] M. Boguñá, F. Papadopoulos, and D. Krioukov, Nature Comms **1**, 62 (2010).

[45] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, 2006).

[46] *The Internet Movie Database (IMDb)*, `http://www.imdb.com/interfaces`.

[47] F. Bonahon, *Low-Dimensional Geometry* (AMS, Providence, 2009).

[48] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá, Phys Rev E **82**, 36106 (2010).

[49] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys Rev Lett **85**, 4633 (2000).

[50] M. Boguñá and R. Pastor-Satorras, Phys Rev E **68**, 36112 (2003).

[51] M. A. Serrano and M. Boguñá, Phys Rev E **72**, 36133 (2005).

[52] S. Maslov, K. Sneppen, and A. Zaliznyak, Physica A **333**, 529 (2004).

[53] L. Le Cam, Pac J Math **10**, 1181 (1960).

[54] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM T Knowl Discov Data **1**, 1 (2007).

[55] G. Bianconi and A.-L. Barabási, Phys Rev Lett **86**, 5632 (2001).

[56] B. Bollobás and O. Riordan, in *Handbook of Graphs and Networks* (Wiley-VCH, Berlin, 2003).

[57] R. Solomonoff and A. Rapoport, B Math Biophys **13**, 107 (1951).

[58] M. Penrose, *Random Geometric Graphs* (Oxford University Press, Oxford, 2003).

[59] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).

[60] S. N. Dorogovtsev, *Lectures on Complex Networks* (Oxford University Press, Oxford, 2010).

[61] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, Oxford, 2010).

[62] R. M. D'Souza, C. Borgs, J. T. Chayes, N. Berger, and R. D. Kleinberg, Proc Natl Acad Sci USA **104**, 6112 (2007).

[63] N. Berger, C. Borgs, J. T. Chayes, R. M. D'Souza, and R. D. Kleinberg, Comb Probab Comput **14**, 697 (2005).

[64] S. Redner, Eur Phys J B **4**, 131 (1998).

[65] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annu Rev Sociol **27**, 415 (2001).

[66] D. J. Watts, P. S. Dodds, and M. E. J. Newman, Science **296**, 1302 (2002).

[67] F. Menczer, Proc Natl Acad Sci USA **99**, 14014 (2002).

[68] F. Menczer, Proc Natl Acad Sci USA **101**, 5261 (2004).

[69] K. Börner, J. T. Maru, and R. L. Goldstone, Proc Natl Acad Sci USA **101**, 5266 (2004).

[70] L. Muchnik, R. Itzhack, S. Solomon, and Y. Louzoun, Phys Rev E **76**, 16106 (2007).

[71] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008), Las Vegas, Nevada, USA, August 24-27, 2008* (ACM, 2008), pp. 160–168, ISBN 978-1-60558-193-4.

[72] O. Simşek and D. Jensen, Proc Natl Acad Sci USA **105**, 12758 (2008).

[73] Y. Li, X. Jin, F. Kong, and J. Li, in *1st IEEE Symposium on Web Society* (IEEE, 2009), pp. 124–128, ISBN 978-1-4244-4157-0.

[74] X.-Q. Cheng, F.-X. Ren, S. Zhou, and M.-B. Hu, New J Phys **11**, 033019 (2009).

[75] S.-H. Yook, H. Jeong, and A.-L. Barabási, Proc Natl Acad Sci USA **99**, 13382 (2002).

[76] M. Dell'Amico, in *Proceedings of the European Conference on Complex Systems (ECCS)* (2006), pp. 1–6.

[77] L. Ferretti and M. Cortelezzi, Phys Rev E **84**, 016103 (2011).

[78] M. Barthélemy, Phys Rep **499**, 1 (2011).

[79] G. Caldarelli, A. Capocci, P. D. L. Rios, and and M. A. Muñoz, Phys Rev Lett **89**, 258702 (2002).

[80] K. Klemm and V. Eguíluz, Phys Rev E **65**, 036123 (2002).

[81] A. Vázquez, M. Boguñá, Y. Moreno, R. Pastor-Satorras, and A. Vespignani, Phys Rev E **67**, 046111 (2003).

[82] A. Vázquez, Phys Rev E **67**, 056104 (2003).

[83] M. O. Jackson and B. W. Rogers, Am Econ Rev **97**, 890 (2007).

[84] P. L. Krapivsky, S. Redner, and F. Leyvraz, Phys Rev Lett **85**, 4629 (2000).

[85] S. Fortunato, A. Flammini, and F. Menczer, Phys Rev Lett **96**, 218701 (2006).

[86] S. N. Dorogovtsev, J. Mendes, and A. Samukhin, Phys Rev E **63**, 062101 (2001).

[87] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, in *Proceedings of the 18th International Conference on World Wide Web (WWW)* (ACM, New York, 2009), p. 641, ISBN 9781605584874.

[88] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Eur Phys J B **38**, 205 (2004).

[89] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).
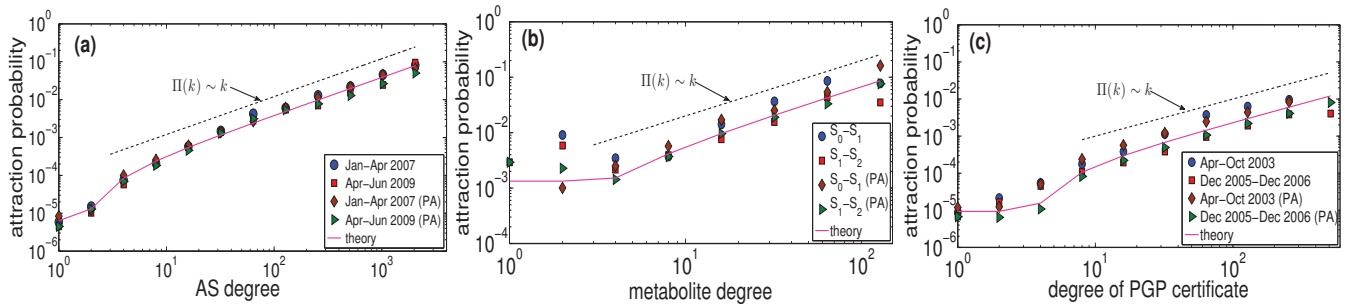
Supplementary Figures

Fig. S1. Plots (a), (b), (c) show the probability $\Pi(k)$ that an old node attracts a new link in the AS Internet, the *E.coli* metabolic network, and the PGP web of trust, respectively. The network snapshots in each case are the ones used in Fig. 3 of the main text. The plots also show the results for the corresponding PA emulations and the theoretical prediction, which is the same linear function in popularity×similarity optimization and PA.
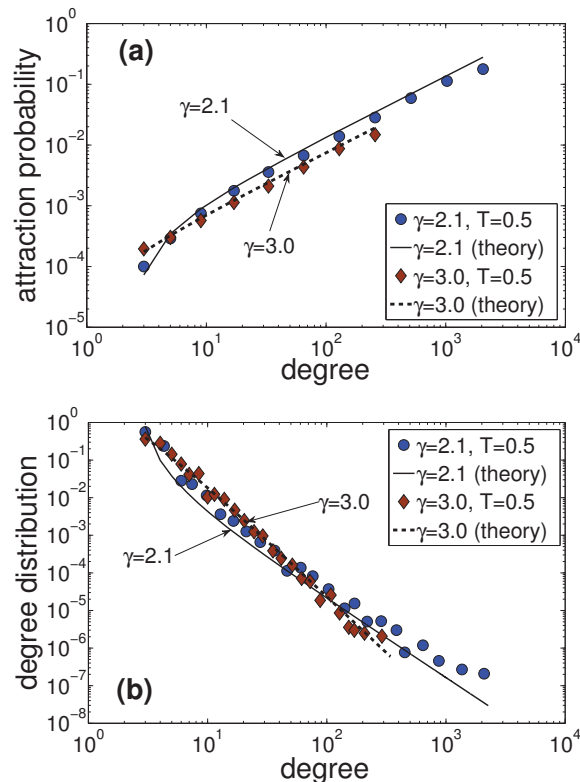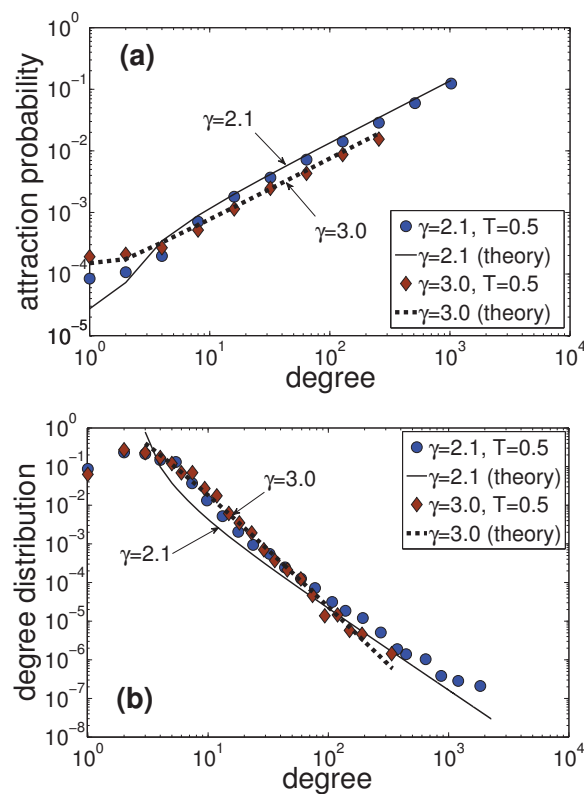

Fig. S2. Plot (a) shows the probability $\Pi(k)$ that an existing node of degree k attracts a link in networks grown up to $t = 10^4$ nodes according to Model$_2$, with $T = 0.5$, $m = 3$, and $\gamma = 2.1, 3.0$. The plot also shows the corresponding theoretical predictions given by Equation (19). Plot (b) shows the distribution $P(k)$ of node degrees in the same networks. The theoretical predictions are given by Equation (18). Small deviations of the theoretical prediction for $\gamma = 2.1$ are due to the increasingly pronounced finite-size effects at $\gamma \to 2$ [88]. Similar results hold for other values of $\gamma \geq 2$, $0 \leq T < 1$, and m, not shown to avoid clutter.
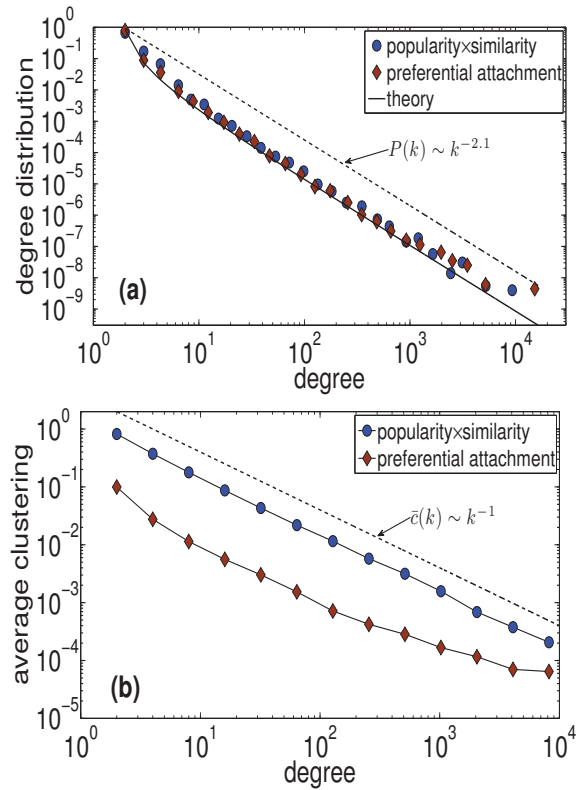
Fig. S3. Plot (a) shows the distribution $P(k)$ of node degrees for the two networks considered in Fig. 2 of the main text. Plot (b) shows for the same two networks the average clustering $\bar{c}(k)$ of k-degree nodes, defined as the ratio of the number of triangles involving a k-degree node to the maximum such number $k(k-1)/2$, averaged over all the k-degree nodes. The $1/k$ scaling of $\bar{c}(k)$ is often considered as a signature of the network's hierarchical organization [89]. The average clustering $\bar{c} = \sum_k \bar{c}(k)P(k)$ in the optimization and PA networks is $\bar{c} = 0.83$ and $\bar{c} = 0.12$, respectively, as mentioned in Fig. 2 of the main text.
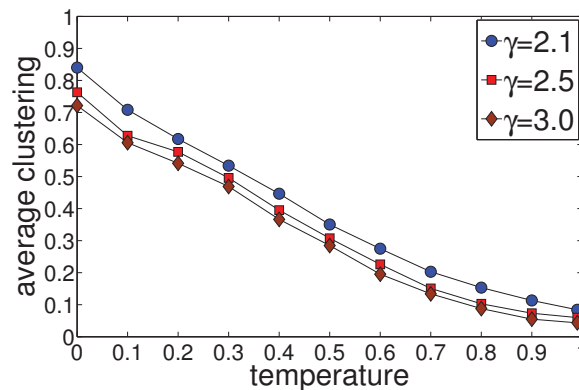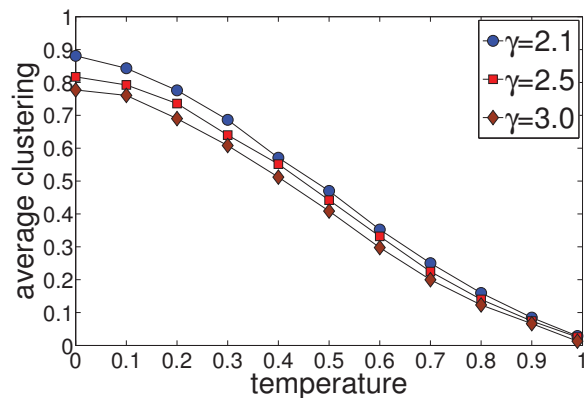

Fig. S4. Average clustering $\bar{c}(t)$ at $t = 10^4$ as a function of temperature $T \in [0, 1)$ in networks grown according to popularity×similarity optimization with $m = 3$.
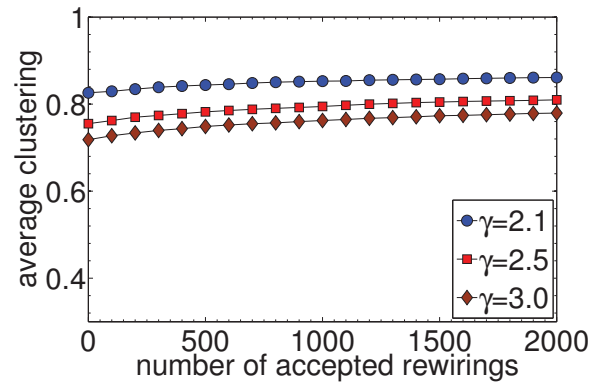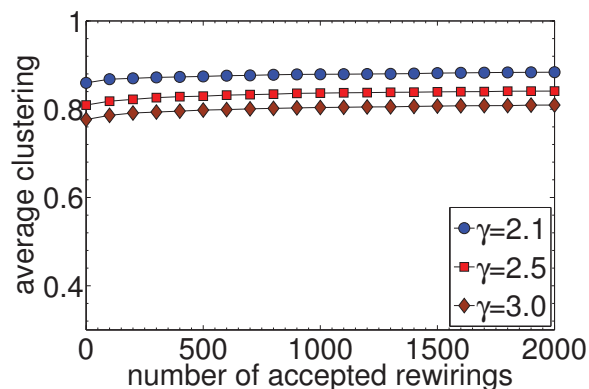
Fig. S5. Average clustering as a function of the number of accepted clustering-increasing rewirings in networks grown according to Model_1 with $m = 3$.

**Fig. S6.** Plot (a) shows the probability $\Pi(k)$ that an existing node of degree k attracts a link in networks grown up to $t = 10^4$ nodes according to $\text{Model}_{2'}$, with $T = 0.5$ and $\gamma = 2.1, 3.0$. Each new node connects on average to $m = 3$ existing nodes. The theoretical predictions are given by Equation (19) when $k \geq m$, and when $k < m$ are given by the formula $\Pi(k) = m \frac{A}{(m+A)t}$. Plot (b) shows the distribution $P(k)$ of node degrees in the same networks. The theoretical predictions are given by Equation (18), which is defined only for $k \geq m$. Compared to Fig. S2, here we observe stronger deviations of the distributions from the power laws at small degrees k, due to fluctuations of the initial degree of a node around its average value $m = 3$. Similar results hold for other values of $\gamma \geq 2$, $0 \leq T < 1$, and m, not shown to avoid clutter.

Fig. S7. Average clustering $\bar{c}(t)$ at $t = 10^4$ as a function of temperature $T \in [0, 1)$ in networks growing according to Model$_{2'}$, where each new node connects on average to $m = 3$ existing nodes. Clustering is calculated excluding nodes of degree 1, whose clustering is always zero.

Fig. S8. Average clustering as a function of the number of accepted clustering-increasing rewirings for networks grown according to Model$_{2'}$ with $T = 0$ and $m = 3$.
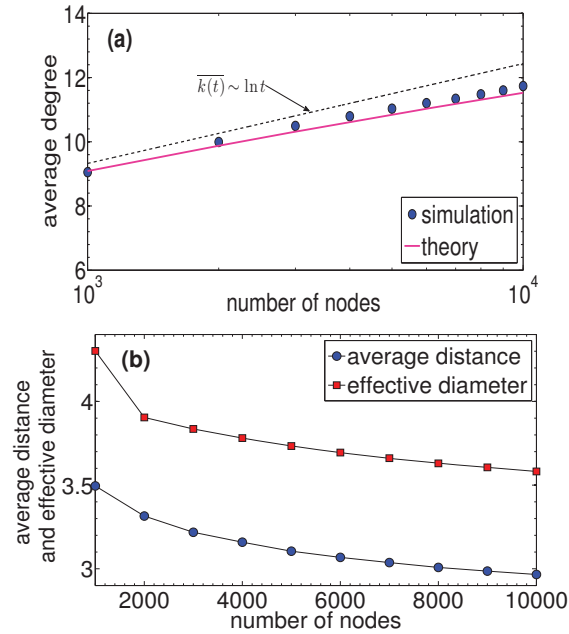
Fig. S9. Densification effects. Plot (a) shows the average degree $\overline{k(t)}$ as a function of size $t \in [10^3, 10^4]$ of networks grown according to Model$_{2'}$, with $T = 0.5, \gamma = 2.1$, and $R_t = r_t$. Plot (b) shows the average shortest path distance between nodes as a function of size $t \in [10^3, 10^4]$ of networks grown as in Plot (a) but with $R_t = \delta r_t$, and $\delta = 1.2$. The plot also shows the *effective diameter* defined as the 90th percentile of the shortest path distance distribution [54].
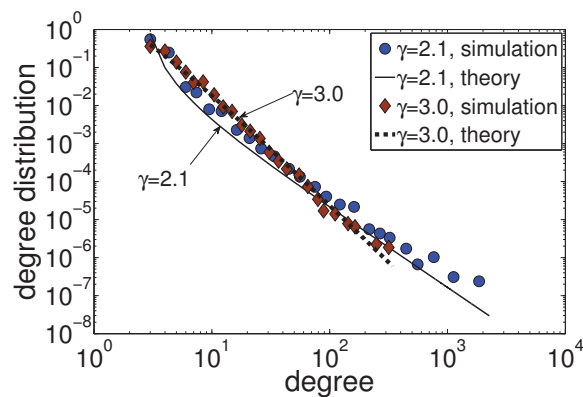
Fig. S10. Distribution $P(k)$ of node degrees in networks growing according to the standard PA limit with $T = \frac{\zeta}{2} \rightarrow \infty$, $m = 3$, and $\gamma = 2.1, 3.0$. The theoretical predictions are given by Equation (18). For $\gamma = 2.1, 3.0$ the average clustering in the simulated networks is $\bar{c} = 0.068, 0.004$.
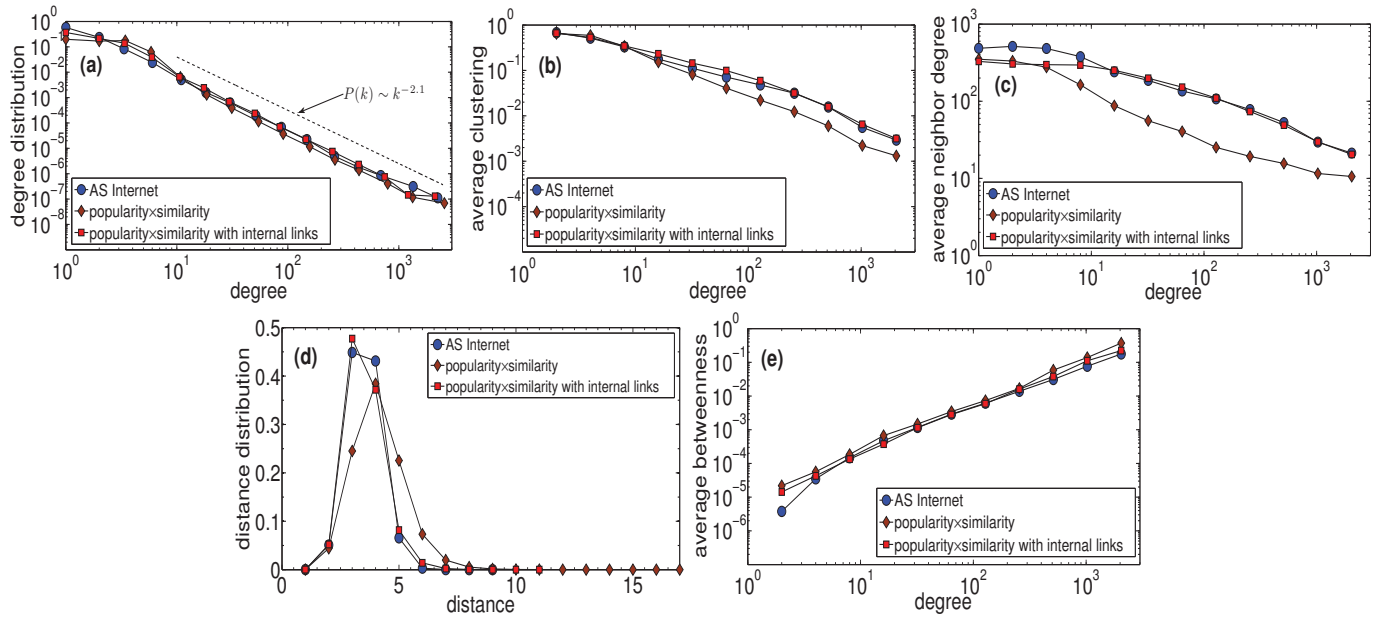
Fig. S11. Properties of the AS Internet vs. networks grown according to popularity×similarity optimization. The plots show: (a) the degree distribution $P(k)$; (b) the average clustering $\bar{c}(k)$ of k-degree nodes; (c) the average neighbor degree $\bar{k}_{nn}(k)$ of k-degree nodes; (d) the distance distribution $d(l)$; and (e) the average node betweenness $\bar{B}(k)$ of k-degree nodes.
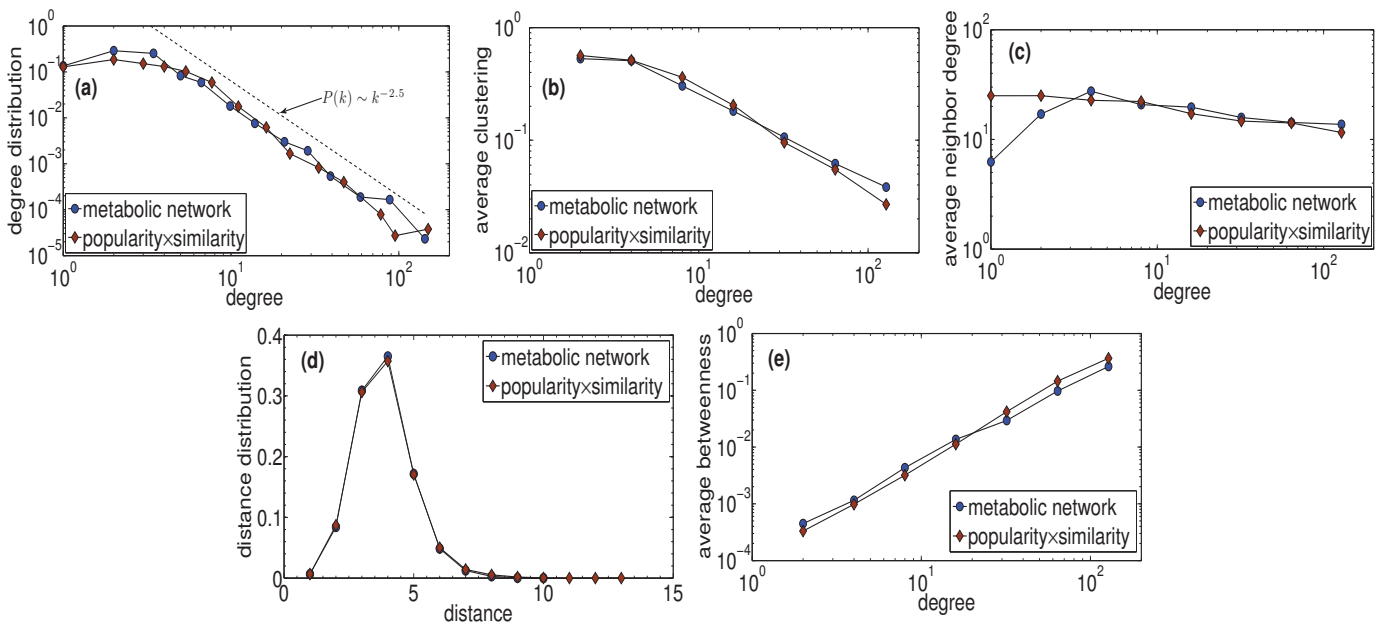

Fig. S12. Properties of the *E.coli* metabolic network vs. networks grown according to popularity×similarity optimization. The plots show: (a) the degree distribution $P(k)$; (b) the average clustering $\bar{c}(k)$ of k-degree nodes; (c) the average neighbor degree $\bar{k}_{nn}(k)$ of k-degree nodes; (d) the distance distribution $d(l)$; and (e) the average node betweenness $\bar{B}(k)$ of k-degree nodes.
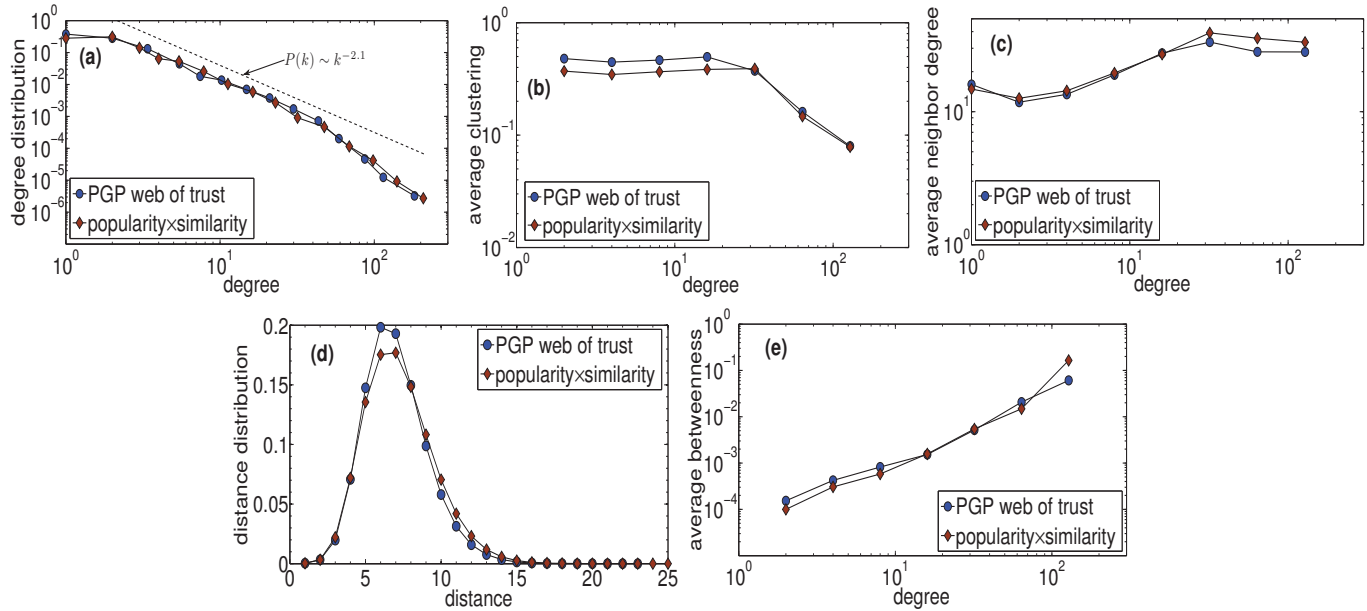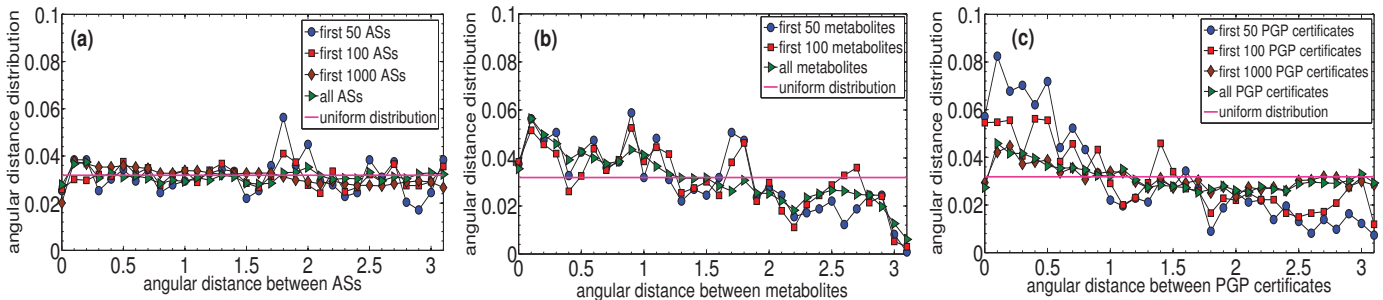
Fig. S13. Properties of the PGP web of trust vs. networks grown according to popularity×similarity optimization. The plots show: (a) the degree distribution $P(k)$; (b) the average clustering $\bar{c}(k)$ of k-degree nodes; (c) the average neighbor degree $\bar{k}_{nn}(k)$ of k-degree nodes; (d) the distance distribution $d(l)$; and (e) the average node betweenness $\bar{B}(k)$ of k-degree nodes.

Fig. S14. Distribution of the inferred angular distances (in radians) in the AS Internet (plot (a)), *E.coli* metabolic network (plot (b)), and PGP web of trust (plot (c)). In each case, we first sort the nodes in the increasing order of their inferred radial coordinates, so that the first i nodes are the nodes with the i smallest radial coordinates, and then compute the distribution of the angular distances for these first i nodes, using bins of size $\delta = 0.1$. We vary i from small values up to the total number of nodes in the network. The straight line in each plot is the uniform distribution $p(\theta) = \delta/\pi$.
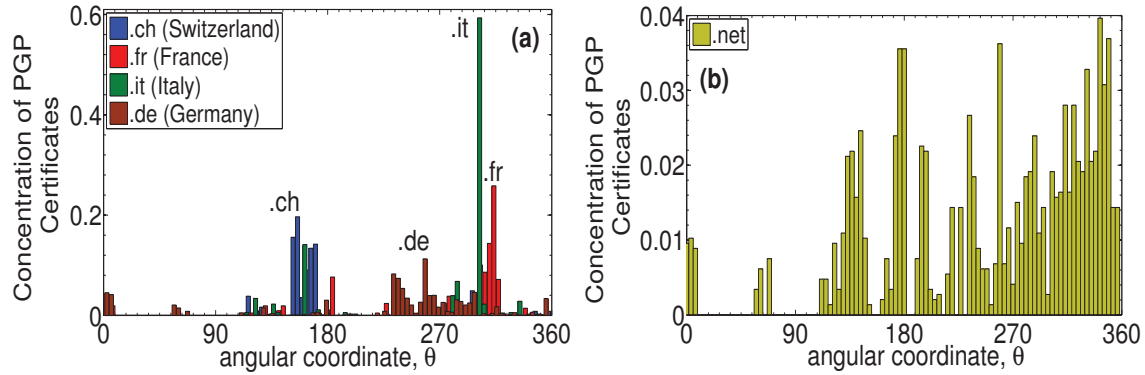
Fig. S15. The mapping of the PGP web of trust yields meaningful results since PGP certificates belonging to the same country code are mapped close to each other (plot (a)). Similar results hold for other country codes. By contrast, certificates belonging to the generic top-level domain .net are widespread (plot (b)), as expected. In [44] we show that the method also yields meaningful results for the Internet, and in [34] for the *E.coli* metabolic network.
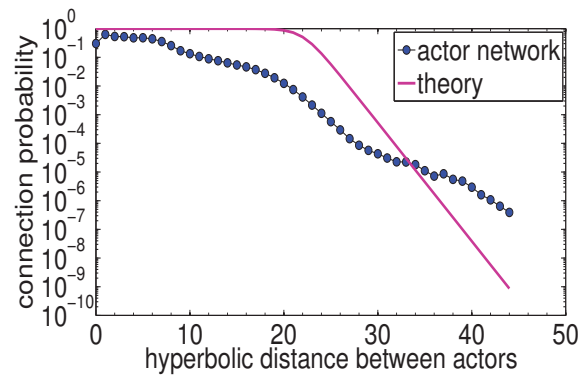

Fig. S16. Connection probability for the actor network considered in Section III C.