LETTER

doi:10.1038/nature11459

Popularity versus similarity in growing networks

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

The principle[1] that 'popularity is attractive' underlies preferential attachment[2], which is a common explanation for the emergence of scaling in growing networks. If new connections are made preferentially to more popular nodes, then the resulting distribution of the number of connections possessed by nodes follows power laws[3,4], as observed in many real networks[5,6]. Preferential attachment has been directly validated for some real networks (including the Internet[7,8]), and can be a consequence of different underlying processes based on node fitness, ranking, optimization, random walks or duplication[9–16]. Here we show that popularity is just one dimension of attractiveness; another dimension is similarity[17–24]. We develop a framework in which new connections optimize certain trade-offs between popularity and similarity, instead of simply preferring popular nodes. The framework has a geometric interpretation in which popularity preference emerges from local optimization. As opposed to preferential attachment, our optimization framework accurately describes the large-scale evolution of technological (the Internet), social (trust relationships between people) and biological (*Escherichia coli* metabolic) networks, predicting the probability of new links with high precision. The framework that we have developed can thus be used for predicting new links in evolving networks, and provides a different perspective on preferential attachment as an emergent phenomenon.

Nodes that are similar have a higher chance of getting connected, even if they are not popular. This effect is known as homophily in social sciences[17,18], and it has been observed in many real networks[19–24]. In the web[23,24], for example, an individual creating her new homepage tends to link it not only to popular sites such as Google or Facebook, but also to not-so-popular sites that are close to her special interests—for example, sites devoted to the composer Tartini or to free solo climbing. These observations suggest the introduction of a measure of attractiveness that would somehow balance popularity and similarity.

The simplest proxy for popularity is the node birth time. All other things being equal, older nodes have more chances to become popular and attract connections[3,4]. If nodes join the network one by one, then the node birth time is simply the node number $t = 1, 2, \ldots$. To model similarity, we randomly place nodes on a circle that represents the simplest similarity space. That is, the angular distances between nodes model their similarity distances, such as the cosine similarity or any other measure[22–24]. The simplest way to model a balance between popularity and similarity is then to establish new connections that optimize the product between popularity and similarity. In other words, the model is simply as follows: (1) initially the network is empty; (2) at time $t \geq 1$, new node t appears at a random angular position θ_t on the circle; and (3) new node t connects to a subset of existing nodes s, $s < t$, consisting of the m nodes with the m smallest values of product $s\theta_{st}$, where m is a parameter controlling the average node degree $\bar{k} = 2m$, and θ_{st} is the angular distance between nodes s and t (Fig. 1a, b). At early times $t \leq m$, node t connects to all the existing nodes.

This model has an interesting geometric interpretation, shown in Fig. 1c. Specifically, after mapping birth time t of a node to its radial coordinate r_t via $r_t = \ln t$, all nodes lie not on a circle but on a plane—their polar coordinates are (r_t, θ_t). It then turns out that new nodes connect simply to the closest m nodes on the plane, except that distances are not Euclidean but hyperbolic[25]. The hyperbolic distance between two nodes at polar coordinates (r_s, θ_s) and (r_t, θ_t) is approximately $x_{st} = r_s + r_t + \ln(\theta_{st}/2) = \ln(st\theta_{st}/2)$. Therefore the sets of nodes s minimizing x_{st} or $s\theta_{st}$ for each t are identical. The hyperbolic

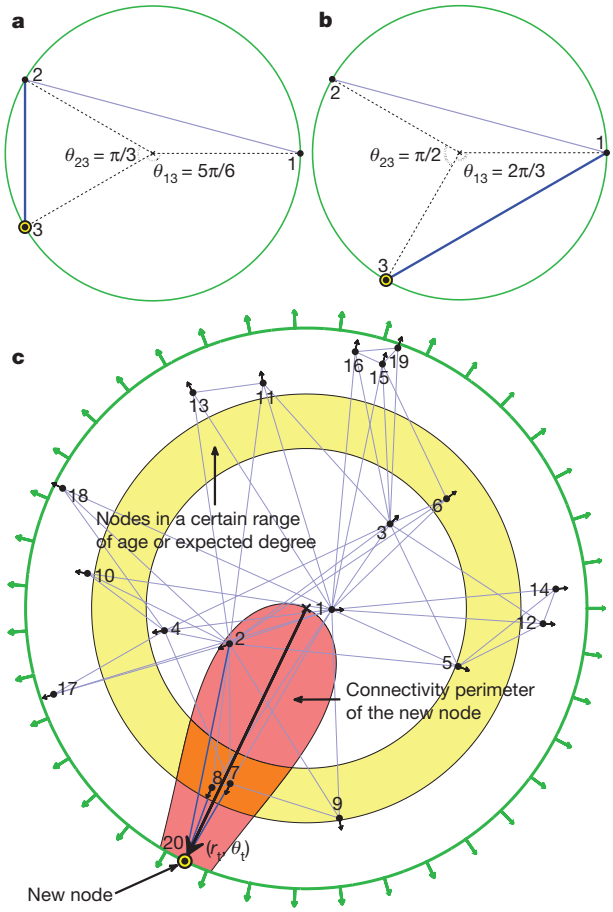

Figure 1 | Geometric interpretation of popularity × similarity optimization. The nodes (dots) are numbered by their birth times, and located at random angular (similarity) coordinates. On its birth, the new circled node t in the yellow annulus connects to m old nodes s minimizing $s\theta_{st}$. The new connections are shown by the thicker blue links. In **a** and **b**, $t = 3$ and $m = 1$. In **a**, node 3 connects to node 2 because $2\theta_{23} = 2\pi/3 < 1\theta_{13} = 5\pi/6$. In **b**, node 3 connects to node 1 because $1\theta_{13} = 2\pi/3 < 2\theta_{23} = \pi$. In **c**, an optimization-driven network with $m = 3$ is simulated for up to 20 nodes. The radial (popularity) coordinate of new node $t = 20$ is $r_t = \ln t$, shown by the long thick arrow. This node connects to the three hyperbolically closest nodes. The red shape marks the set of points located at hyperbolic distances less than r_t from the new node. Arrows on dots show all nodes drifting away from the crossed origin, emulating popularity fading as explained in the text. The drift speed in the network shown corresponds to the degree distribution exponent $\gamma = 2.1$. The outer green circle shows the current network boundary of radius $r_t = \ln t$ expanding with time t as indicated by green arrows.

[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, California 92093, USA. [3]Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.

distance is then nothing but a convenient single-metric representation of a combination of the two attractiveness attributes, radial popularity and angular similarity. We will use this metric extensively below.

The networks grown as described may seem to have nothing in common with preferential attachment (PA)[2–4]. Yet we show in Fig. 2a that the probability $\Pi(k)$ that an existing node of degree k attracts a connection from a new node is the same linear function of k in the described model and in PA. It is not surprising then that the degree distributions in PA and in our model are the same power laws. In Supplementary Information section IV, we prove that the exponent γ of this power law approaches 2. Preferential attachment thus emerges as a process originating from optimization trade-offs between popularity and similarity.

However, there are crucial differences between such optimization and PA. In the latter, new nodes connect with the same probability $\Pi(k)$ to any nodes of degree k in the network. In the former, new nodes connect only to specific subsets of such k-degree nodes that are closest to the new node along the similarity dimension θ (Fig. 1c). To quantify, we compare in Fig. 2b the probability of connection between a pair of nodes as a function of their hyperbolic distance in the two cases. We see that close nodes are almost always connected in the optimization model, whereas in PA the probability of their connection is lower by an order of magnitude. On the other hand, nodes that are far apart are never connected in the optimization model, whereas they can be connected in PA. These differences manifest themselves in the strength of clustering, which is the probability that two nodes connected to the same node are also connected to each other. In PA, clustering is asymptotically zero[26], whereas it is strong in many real networks[5,6].

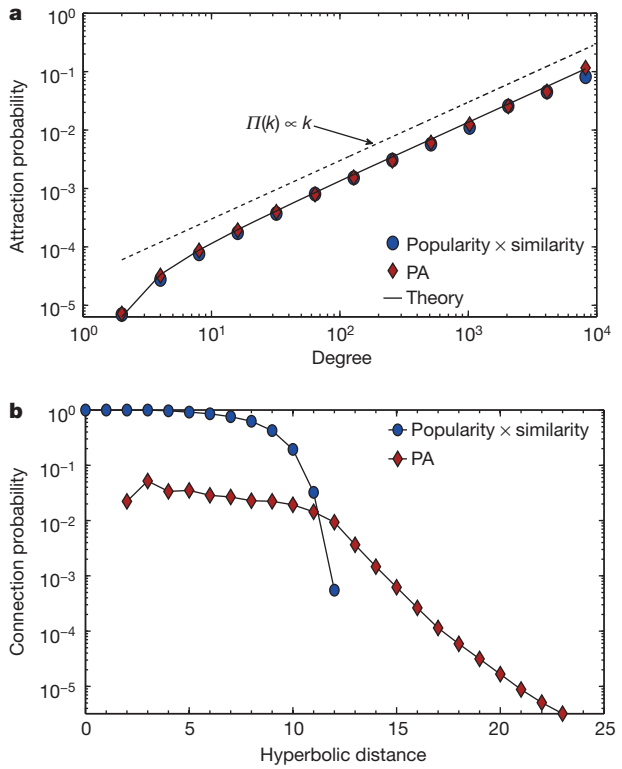

Figure 2 | Emergence of PA from popularity × similarity optimization. Two growing networks have been simulated up to $t = 10^5$ nodes, one growing according to the described optimization model, and the other according to PA. In both networks, each new node connects to $m = 2$ existing nodes. The $\gamma \to 2$ limit is not well-defined in PA, so that $\gamma = 2.1$ is used instead as described in the text. **a**, The probability $\Pi(k)$ that an existing node of degree k attracts a new link. The solid line is the theoretical prediction, while the dashed line is a linear function, $\Pi(k) \propto k$. **b**, The probability $p(x)$ that a pair of nodes located at hyperbolic distance x are connected. The average clustering (over all nodes) in the optimization and PA networks is $\bar{c} = 0.83$ and $\bar{c} = 0.12$, respectively.

We show in Supplementary Information section IV that the described optimization model leads to clustering that is the strongest possible for networks with a given degree distribution.

The strength of clustering and the power-law exponent can both be adjusted to arbitrary values via the following model modifications. We first consider the effect of popularity fading, observed in many real networks[27,28]. We note that the closer the node to the centre in Fig. 1c, the more popular it is: the higher its degree, and the more new connections it attracts, which explains why PA emerges in the model. Therefore to model popularity fading, we let all nodes drift away from the centre such that the radial coordinate of node s at time $t > s$ is increasing as $r_s(t) = \beta r_s + (1 - \beta) r_t$, where $r_s = \ln s$ and $r_t = \ln t$, and parameter $\beta \in [0, 1]$. This modification is identical to minimizing $s^{\beta}\theta_{st}$ (or $s^{b}\theta^{a}_{st}$ with $\beta = b/a$) instead of $s\theta_{st}$. It changes the power-law exponent to $\gamma = 1 + 1/\beta \geq 2$. If $\beta = 1$, the nodes do not move and $\gamma = 2$. If $\beta = 0$, all nodes move with the maximum speed, always lying on the circle of radius r_t, while the network degenerates to a random geometric graph growing on the circle. PA emerges at any $\gamma = 1 + 1/\beta$ since the attraction probability $\Pi(k)$ is a linear function of degree k, $\Pi(k) \propto k + m(\gamma - 2)$, the same as in PA[4]. We prove these statements in Supplementary Information sections IV–VII, where we also show that the popular fitness model[10] can be mapped to our geometric optimization framework by letting different nodes drift away with different speeds (Supplementary Information section V).

Because the strongest clustering is due to connections to the closest nodes, to weaken clustering we allow connections to more-distant nodes. Connecting to the m closest nodes is approximately the same as connecting to nodes lying within distance $R_t \approx r_t$ (see Fig. 1c and Supplementary Information section IV, where we derive the exact expression for R_t, which controls the average degree in the network). If new nodes t establish connections to existing nodes s at hyperbolic distance x_{st} with probability $p(x_{st}) = 1/\{1 + \exp[(x_{st} - R_t)/T]\}$, where parameter $T \geq 0$ is the network temperature (see Supplementary Information sections IV and VI), then clustering is a decreasing function of temperature. That is, temperature is the parameter controlling clustering in the network. At zero temperature, the connection probability $p(x_{st})$ is either 1 or 0 depending on whether distance x_{st} is less or greater than R_t, so that we recover the strongest clustering case above, where new nodes connect only to the closest existing nodes. Clustering gradually decreases to zero at $T = 1$, and remains asymptotically zero for any $T \geq 1$ (Supplementary Information sections IV, VI). At high temperatures $T \to \infty$, the model degenerates either to growing random graphs, or, remarkably, to standard PA (Supplementary Information section VII).

To investigate if similarity shapes the structure and dynamics of real networks as our model predicts, we consider a series of historical snapshots of the Internet, the *E. coli* metabolic network, and the social network of trust relationships between people, also known as the web of trust (WoT). The first two networks are disassortative (nodes of dissimilar degrees are connected with a higher probability), while the third is assortative (nodes of similar degrees are connected with a higher probability), and its degree distribution deviates from a power law. We map these networks to their popularity × similarity spaces (Methods Summary). The mapping infers the radial (popularity) and angular (similarity) coordinates for all nodes, so that we can compute the hyperbolic distances between all node pairs, and the probabilities of new connections as functions of the hyperbolic distance between corresponding nodes. These probabilities are shown in Fig. 3. In all the three networks, they are close to the theoretical predictions of our model.

This finding is important for several reasons. First, it shows that real-world networks evolve as our framework predicts. Specifically, given the popularity and similarity coordinates of two nodes, they link with probability close to the theoretical value predicted by the model. The framework may thus be used for link prediction, a notoriously difficult and important problem in many disciplines[29], with applications

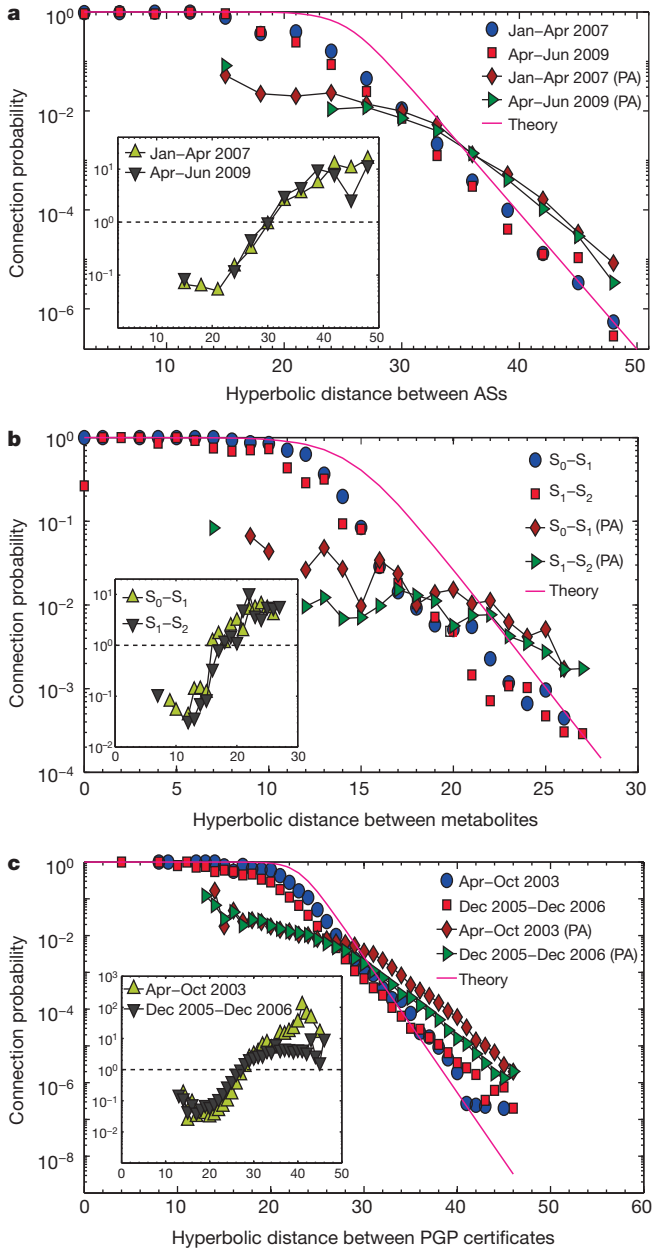

Figure 3 | Popularity × similarity optimization for three different networks. **a**, The growing Internet; **b**, *E. coli* metabolic network; and **c**, pretty-good-privacy (PGP) web of trust (WoT) between people. Each plot shows the probability of connections between new and old nodes, as a function of the hyperbolic (popularity × similarity) distance x between them in the real networks (circles and squares) and in PA emulations (diamonds and triangles). To emulate PA, new links are disconnected from old nodes to which these links are connected in the real networks, and reconnected to old nodes according to PA. For a pair of historical network snapshots S_0 (older) and S_1 (newer), new nodes are the nodes present in S_1 but not in S_0, and old nodes are the nodes present both in S_1 and S_0. Each plot shows the data for two pairs of such historical snapshots. The solid curve in each plot is the theoretical connection probability in the optimization model with the parameters corresponding to a given real network. Because the probability of new connections in the real networks is close to the theoretical curves, the shown data demonstrate that these networks grow as the popularity × similarity optimization model predicts, whereas PA, accounting only for popularity, is off by orders of magnitude in predicting the connections between similar (small x) or dissimilar (large x) nodes. To quantify this inaccuracy, the insets show the ratio between the connection probabilities in PA emulations and in the real networks, that is, the ratios of the values shown by diamonds and circles, and by triangles and squares in the main plots. The x-axes in the insets are the same as in the main plots.

ranging from predicting protein interactions or terrorist connections to designing recommender and collaborative filtering systems[30]. Second, Fig. 3 directly validates our framework and its core mechanism. It is not surprising then that, as a consequence, the synthetic graphs that the model generates are remarkably similar to real networks across a range of metrics (Supplementary Information section IX), implying that the framework can be also used for accurate modelling of real network topologies. We review related work in Supplementary Information section X, and to the best of our knowledge, there is no other model that would simultaneously: (1) be simple and universal, that is, applicable to many different networks, (2) have a similarity space as its core component, (3) cast PA as an emergent phenomenon, (4) generate graphs similar to real networks across a wide range of metrics, and (5) validate the proposed growth mechanism directly. Validation is usually limited to comparing certain graph metrics, such as degree distribution, between modelled and real networks; however, this 'validates' a consequence of the mechanism, not the mechanism itself. Direct validation is usually difficult, because proposed mechanisms tend to incorporate many unmeasurable factors—economic or political factors in Internet evolution, for example. Our approach is no different in that it cannot measure all the factors or node attributes contributing to node similarity in any of the considered real networks. Yet, the angular distances between nodes in our approach can be considered as projections of properly weighted combinations of all such similarity factors affecting network evolution, and we can infer these distances using statistical inference methods, directly validating the growth mechanism.

To summarize, popularity is attractive, but so is similarity. Neglecting the latter would lead to severe aberrations. Within the Internet, for example, a local network in Nebraska would connect directly to a local network in Tibet, in the same way as on the web, a person not even knowing about Tartini or free solo climbing would suddenly link her page to these subjects. The probability of such dissimilar connections is very low in reality, and the stronger the similarity forces, the smaller this probability is. Neglecting the network similarity structure leads to overestimations or underestimations of the probability of dissimilar or similar connections by orders of magnitude (Fig. 3). However, one cannot tell the difference between our framework and PA by examining node degrees only. The probability that an existing node of degree k attracts a new link optimizing popularity × similarity is exactly the same linear function of k as in PA (Fig. 2a). Supplementary Fig. 1 shows that this function is indeed realized in the considered real networks, re-validating effective PA for these networks. Therefore the popularity × similarity optimization approach provides a natural geometric explanation for the following 'dilemmas' characteristic of PA. On the one hand, PA has been validated for many real networks, while on the other hand, it requires exogenous mechanisms to explain not only strong clustering, but also linear popularity preference, and how such preference can emerge in real networks, where nodes do not have any global information about the network structure. As PA appears as an emergent phenomenon in the framework developed here, our framework provides a simple and natural resolution to these dilemmas, and this resolution is directly validated against large-scale evolution of very different real networks.

We conclude with the observation that to know the closest nodes in the hyperbolic popularity × similarity space requires precise global information about all node locations. However, non-zero temperatures smooth out the sharp connectivity perimeter threshold in Fig. 1c, thus modelling reality where this proximity information is not precise and mixed with errors and noise. In that respect, PA is a limiting regime with similarity forces reduced to nothing but noise.

METHODS SUMMARY

To infer the radial r_i and angular θ_i coordinates for each node i in a real network snapshot with adjacency matrix a_{ij}, $i, j = 1, 2, \ldots, t$, we use the Markov Chain Monte Carlo (MCMC) method described in detail in Supplementary Information.

Specifically, we derive there the exact relation between the expected current degree $\bar{k}_i$ of node i and its current radial coordinate r_i, which scales as $\bar{k}_i \sim e^{r_t - r_i}$. To infer the radial coordinates we use the same expression substituting in it the real degrees k_i of nodes instead of their expected degrees. Having inferred the radial coordinates, we then execute the Metropolis-Hastings algorithm to find the node angular coordinates that maximize likelihood $\mathcal{L} = \Pi_{i<j} p(x_{ij})^{a_{ij}} \left[1 - p(x_{ij})\right]^{1-a_{ij}}$, where $p(x_{ij}) = 1 \Big/ \left[1 + e^{(x_{ij}-R)/T}\right]$ is the connection probability in the model, and parameters R and T are defined by the average node degree and clustering in the network via expressions in Supplementary Information section IV. Likelihood $\mathcal{L}$ is the probability that the network snapshot with node coordinates (r_i, θ_i), defining the hyperbolic distances x_{ij} between all nodes, is produced by the model. The algorithm employs an MCMC process, which finds coordinates θ_i for all i that approximately maximize $\mathcal{L}$. Further details are in Supplementary Information sections II and III, where we also show that the method yields meaningful results for the considered networks, but not for a network (movie actor collaborations) to which popularity × similarity optimization does not apply.

The nodes in Fig. 3a, b and c are respectively autonomous systems (ASs), metabolites and Pretty Good Privacy (PGP) certificates associating users' email addresses with their cryptographic keys. Parameters (R, T) used to infer the coordinates and to draw the theoretical connection probabilities are (25.2, 0.79), (14.4, 0.77) and (23, 0.59). Each panel of Fig. 3 shows data for two pairs of snapshots: Fig. 3a, January–April 2007 and April–June 2009; Fig. 3b, S_0–S_1 and S_1–S_2 defined in Supplementary Information section I; and Fig. 3c, April–October 2003 and December 2005–December 2006. The few missing data points in the empirical curves (circles and squares) indicate that there are no node pairs at the corresponding distances after the mapping, whereas extra missing points in the PA emulation curves (diamonds and triangles) indicate that all node pairs at those distances are not connected after PA emulations, meaning that the PA connection probability is zero there.

Received 25 April; accepted 26 July 2012.
Published online 12 September 2012.

1. Dorogovtsev, S., Mendes, J. & Samukhin, A. WWW and Internet models from 1955 till our days and the "popularity is attractive" principle. Preprint at http://arXiv.org/abs/cond-mat/0009090 (2000).
2. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
3. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **85,** 4629–4632 (2000).
4. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85,** 4633–4636 (2000).
5. Dorogovtsev, S. N. *Lectures on Complex Networks* (Oxford Univ. Press, 2010).
6. Newman, M. E. J. *Networks: An Introduction* (Oxford Univ. Press, 2010).
7. Pastor-Satorras, R., Vázquez, A. & Vespignani, A. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.* **87,** 258701 (2001).
8. Jeong, H., Néda, Z. & Barabási, A. L. Measuring preferential attachment in evolving networks. *Europhys. Lett.* **61,** 567–572 (2003).
9. Dorogovtsev, S. N., Mendes, J. & Samukhin, A. Size-dependent degree distribution of a scale-free growing network. *Phys. Rev. E* **63,** 062101 (2001).
10. Bianconi, G. & Barabási, A.-L. Bose-Einstein Condensation in complex networks. *Phys. Rev. Lett.* **86,** 5632–5635 (2001).
11. Caldarelli, G., Capocci, A., Rios, P. D. L., &. Muñoz, M. A. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89,** 258702 (2002).
12. Vázquez, A. Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E* **67,** 056104 (2003).
13. Pastor-Satorras, R., Smith, E. & Sole, R. V. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222,** 199–210 (2003).
14. Fortunato, S., Flammini, A. & Menczer, F. Scale-free network growth by ranking. *Phys. Rev. Lett.* **96,** 218701 (2006).
15. D'Souza, R. M., Borgs, C., Chayes, J. T., Berger, N. & Kleinberg, R. D. Emergence of tempered preferential attachment from optimization. *Proc. Natl Acad. Sci. USA* **104,** 6112–6117 (2007).
16. Motter, A. E. & Toroczkai, Z. Introduction: optimization in networks. *Chaos* **17,** 026101 (2007).
17. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27,** 415–444 (2001).
18. Simşek, O. & Jensen, D. Navigating networks by using homophily and degree. *Proc. Natl Acad. Sci. USA* **105,** 12758–12762 (2008).
19. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* **4,** 131–134 (1998).
20. Watts, D. J., Dodds, P. S. & Newman, M. E. J. Identity and search in social networks. *Science* **296,** 1302–1305 (2002).
21. Börner, K., Maru, J. T. & Goldstone, R. L. The simultaneous evolution of author and paper networks. *Proc. Natl Acad. Sci. USA* **101,** 5266–5273 (2004).
22. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. & Suri, S. in *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)* (eds Li, Y., Liu, B. & Sarawagi, S.) 160–168 (ACM, 2008).
23. Menczer, F. Growing and navigating the small world Web by local content. *Proc. Natl Acad. Sci. USA* **99,** 14014–14019 (2002).
24. Menczer, F. Evolution of document networks. *Proc. Natl Acad. Sci. USA* **101,** 5261–5265 (2004).
25. Bonahon, F. *Low-Dimensional Geometry* (AMS, 2009).
26. Bollobás, B. & Riordan, O. in *Handbook of Graphs and Networks* (eds Bornholdt, S. & Schuster, H. G.) Ch. 1 1–34 (Wiley-VCH, 2003).
27. Adamic, L. A. & Huberman, B. A. Power-law distribution of the World Wide Web. *Science* **287,** 2115 (2000).
28. van Raan, A. F. J. On growth, ageing, and fractal differentiation of science. *Scientometrics* **47,** 347–362 (2000).
29. Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453,** 98–101 (2008).
30. Menon, A. K. & Elkan, C. in *Machine Learning and Knowledge Discovery in Databases (ECML)* (eds Gunopulos, D., Hofmann, T., Malerba, D. & Vazirgiannis, M.) 437–452 (Lecture Notes in Computer Science, Vol. 6912, Springer, 2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Elkan, G. Bianconi, P. Krapivsky, S. Redner, S. Havlin, E. Stanley and A.-L. Barabási for discussions and suggestions. This work was supported by a Marie Curie International Reintegration Grant within the 7th European Community Framework Programme; MICINN Projects FIS2010-21781-C02-02 and BFU2010-21847-C02-02; Generalitat de Catalunya grant 2009SGR838; the Ramón y Cajal programme of the Spanish Ministry of Science; ICREA Academia prize 2010, funded by the Generalitat de Catalunya; NSF grants CNS-0964236, CNS-1039646, CNS-0722070; DHS grant N66001-08-C-2029; DARPA grant HR0011-12-1-0012; and Cisco Systems.

Author Contributions F.P. and D.K. planned research, performed research and wrote the paper; M.K., M.A.S. and M.B. planned and performed research. All authors discussed the results and reviewed the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.P. (f.papadopoulos@cut.ac.cy) or D.K. (dima@ucsd.edu).