Letter

Clique densification in networks

Haochen Pi,[1] Keith Burghardt,[2,*] Allon G. Percus,[3,2] and Kristina Lerman[2]

[1]*Department of Computer Science, University of Southern California, Los Angeles, California 90007, USA*
[2]*Information Sciences Institute, University of Southern California, Marina del Rey, California 90292, USA*
[3]*Institute of Mathematical Sciences, Claremont Graduate University, Claremont, California 91711, USA*

(Received 26 December 2022; accepted 5 April 2023; published 26 April 2023)

Real-world networks are rarely static. Recently, there has been increasing interest in both network growth and network densification, in which the number of edges scales superlinearly with the number of nodes. Less studied but equally important, however, are scaling laws of higher-order cliques, which can drive clustering and network redundancy. In this paper, we study how cliques grow with network size, by analyzing several empirical networks from emails to Wikipedia interactions. Our results show superlinear scaling laws whose exponents increase with clique size, in contrast to predictions from a previous model. We then show that these results are in qualitative agreement with a model that we propose, the local preferential attachment model, where an incoming node links not only to a target node, but also to its higher-degree neighbors. Our results provide insights into how networks grow and where network redundancy occurs.

DOI: 10.1103/PhysRevE.107.L042301

I. INTRODUCTION

Networks underlie a wide variety of social phenomena, from the spread of disease and information [1] to the formation of collaborations [2,3]. The evolution of networks has been a popular research topic since the Barabasi-Albert model demonstrated that growth through preferential attachment can explain a fundamental property of networks—their heavy-tailed degree distributions [4,5]. More recent research has studied another fundamental aspect of network growth, known as densification, where the number of links increases superlinearly with the number of nodes [3]. Densification can create advantages for larger systems: For instance, in collaboration networks, it provides more opportunities for researchers at larger institutions over smaller ones [2]. Several network growth models have been developed to help explain mechanisms of specific networks, such as gene regulatory networks [6,7], or provide general mechanisms of patterns seen in empirical data, such as fitness [8], graph spectra [9], or copying mechanisms [2,10,11], among others [12–15].

Growth of higher-order structures in networks is a less studied aspect of network growth, but is critical to our understanding of a range of phenomena, including disease spread [16]. Although some higher-order structures, such as triangles [17], have long been known to play an important role in network phenomena, less attention has been devoted to how these and higher-order motifs form in growing networks. Recent research, notably by Bhat *et al.* [10] and Lambiotte *et al.* [11], has offered potential mechanisms that predict how edges and larger cliques will scale as a function of network size. [For clarity, a clique of size k is a fully connected subgraph, with k nodes and $k(k-1)/2$ edges]. The mechanism of clique formation proposed by Lambiotte *et al.*, however, has not been tested empirically before.

In this paper, we study clique formation in growing networks. Figure 1(a) considers the case of an empirical network of user interactions on the question-answer website known as Math Overflow [14] (answers to questions, comments to questions, and comments to answers). Plotting the number of cliques of differing size k as a function of network size (measured by the number of nodes), we see that the number of edges ($k = 2$) grows superlinearly with network size. Network degree therefore increases with network size, consistent with previous results [2,3,10,11]. But crucially, we observe that the number of triangles ($k = 3$) and larger cliques grows *even faster*, leading to an increased level of redundant connections in the network. This effect, which we call *clique densification*, is found in many empirical networks (see also Fig. S1 in the Supplemental Material [18], which shows results for other networks [14,19–22]). We also find that these networks form links locally, i.e., between nearby nodes, and preferentially connect to high-degree nodes. Furthermore, the effect of 2-cliques being overtaken by increasingly large clique sizes in Fig. 1(a) gives rise to an intriguing envelope structure that itself appears to follow a power law.

In order to explain our findings, we propose the local preferential attachment model (LPAM) that combines two prevalent mechanisms in networks: Copying (linking not only to a target but also to some of its neighbors) [23], and preferential attachment (linking preferentially to higher-degree nodes). While copying alone can explain some network densification, it does not explain why the representation of large cliques grows so rapidly in networks. Similarly, preferential attachment cannot explain densification at all. The two mechanisms together, however, are key to understanding how such dense substructures arise in networks. These substructures can be useful, for example, when links are removed because they

[*] keithab@isi.edu

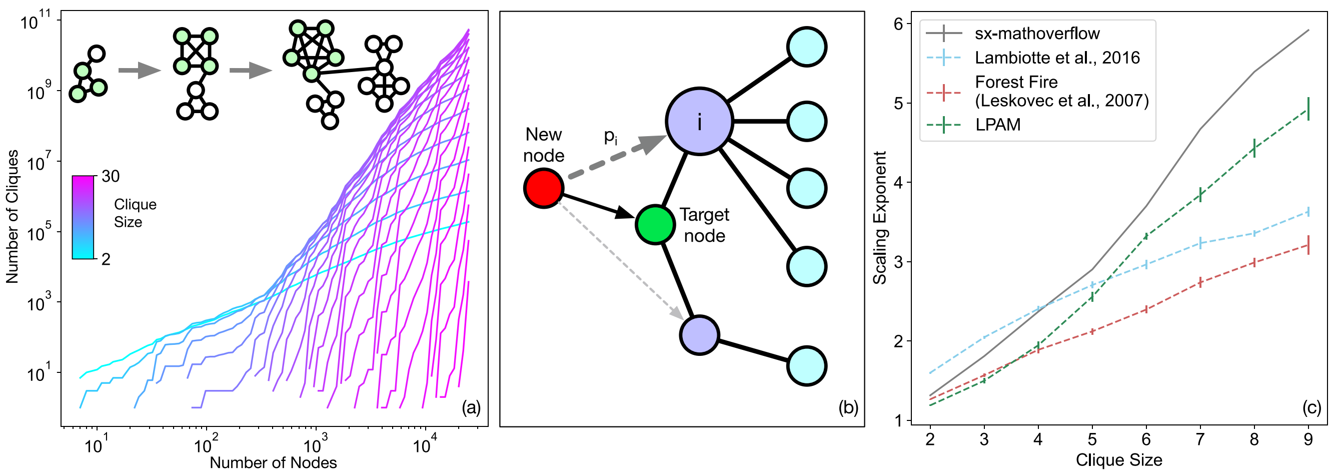

FIG. 1. (a) Number of cliques of a given size vs the number of nodes in the network for the Math Overflow question-answer website [14]. See also Fig. S1 in the Supplemental Material [18] for results on other networks [14,19–22]. (b) Local preferential attachment model (LPAM) preferentially attaches to higher-degree neighbors of the target node. A new node (red) connects to a random target node (green), as well as to the target node's higher-degree existing neighbors (purple). (c) Scaling laws vs clique size for LPAM, node copying mechanism of Lambiotte *et al.* [11], Forest fire model [3], and Math Overflow data.

provide redundancy that maintains network connectivity. This may help us understand seeming inefficiencies in network formation, as the density of these subgraphs may preserve the giant connected component of a network. Moreover, the copying and preferential attachment mechanisms could assist in explaining the formation of dense subcommunities in networks [24]. Our work provides a new understanding of how network structure evolves and can help account for these behaviors.

II. METHODS

In this section, we describe how cliques in a network are counted. We define our model that better explains how cliques scale super-linearly with network size, and we discuss how we fit this and other models to data.

A. Empirical Networks

The empirical datasets we use are freely available from the SNAP library [25]. We take 11 graphs that contain temporal information, ignoring weights and edge direction: College Messages [19] (nodes are users, and edges are messages between individuals); an email network at a large European institution [14] (nodes are users and edges are emails between users); Reddit hyperlinks within the body and within the title of posts [20] (nodes are users and edges are links to comments between users); Bitcoin Alpha and Bitcoin OTC trust weighted signed networks [21,22] (nodes are users and edges represent degree of trust, where we ignore the edge sign); conversations on Ask Ubuntu, Math Overflow, Stack Overflow, Stack Exchange Super User boards (nodes are users and edges represent comments to questions or answers, or answers to questions between users) [14]; and Wikipedia's talk pages (nodes are users and edges are comments between users) [14]. Data are captured cumulatively, such that links and nodes will appear but not disappear from the first to the last timestamp.

B. Counting Cliques

It is typically a challenge to analyze high-order network properties, such as cliques, in part because finding the largest clique in a network is NP-hard [26]. Pivoter [27], however, helps speed up clique counting, allowing clique densification to be studied. Pivoter is based on the succinct clique tree, which efficiently stores a representation of all cliques in the network. This is built via an algorithm called pivoting, which reduces the recursion tree used to find the cliques. We use this method to study all empirical networks. Code used to model and analyze data is available at Ref. [30].

C. Local Preferential Attachment Model

We find three attributes of growing networks that we aim to capture within a single mechanistic model: (a) the number of cliques scaling superlinearly with the network size, (b) nodes forming new links with nearby nodes, and (c) nodes preferentially connecting to high-degree nodes. One theoretically grounded mechanistic model is by Lambiotte *et al.* [11], in which nodes enter the network, find a random target node to connect with, and then also connect to random neighbors of that target node. Their model provides theoretical predictions on the scaling laws of edges and higher-order cliques versus network size, but does not assume any preferential attachment mechanism.

We therefore expand on this model with LPAM, shown in Fig. 1(b). Consider a process where, at each time step, a new node (red node) enters the network. It connects to an existing target node (green node) chosen uniformly at random, and also connects to some number of neighbors (purple nodes) of the target, with preference given to higher-degree nodes [larger-sized nodes in Fig. 1(b)].

LPAM is characterized by two parameters, p and r. For a target node of degree k, the marginal probability of establishing a connection to a given one of its neighbors is p, such that the expected number of new connections is pk. However, conditional on the neighbor's own degree, this probability depends on r. The parameter r interpolates linearly between the case of no preferential attachment at all ($r = 0$), corresponding to the Lambiotte *et al.* model [11], and the case of strong preferential attachment ($r = 1$). Specifically, for the ith neighbor of the target node, we define the initial scaled probability

$$p_i = p \frac{k_i}{\sum_{j=1}^{k} k_j/k}, \tag{1}$$

where k_i is the degree of the ith neighbor. If p_i exceeds a threshold level $p + (1 - p)r$, then the "excess" probability $p_i - (p + (1 - p)r)$ is spread over the probabilities of connecting to other neighbors of the target node, giving new probabilities $p'_j = p_j + (p_i - (p + (1 - p)r))/(k - 1)$. As this may result in certain probabilities exceeding $p + (1 - p)r$, the process is iterated until all probabilities fall below that threshold. The end result is an expectation value independent of r; we continue to connect to pk nodes on average, but with a preferential attachment to higher-degree nodes. The threshold level allows us to smoothly transition between strictly preferential attachment ($r = 1$) and the Lambiotte *et al.* node copying model [11].

For a network with N nodes and $L(N)$ links, the network growth mechanism implies, as in [11],

$$L(N + 1) = L(N) + 1 + 2p\frac{L(N)}{N}. \tag{2}$$

Following the same theory as in Bhat *et al.* [10], this results in

$$L(N) = \begin{cases} N/(1 - 2p) & p < 1/2 \\ N \ln N & p = 1/2, \\ A(p)N^{2p} & p > 1/2 \end{cases} \tag{3}$$

i.e., the number of links scales superlinearly for $p > 1/2$, where $A(p) = [(2p - 1)\Gamma(1 + 2p)]^{-1}$. Sadly, because p_i depends on the other neighbor degrees k_j, higher-order dependencies are not solvable, such as the number of triangles as a function of N. We instead calculate scaling laws numerically by taking a linear fit of the log of the number of cliques versus log of the network size [see Fig. 1(a)] for different realizations of this model.

D. Fitting Models

Another methodological contribution of our work is fitting a clique densification model to empirical data of clique scaling. We measure the distribution of clique sizes for a given network size and compare this distribution to our model's prediction (see an example of this distribution in Fig. S2 in the Supplemental Material [18] for our model and competing models [3,11]). We find the parameters that fit the empirical distributions best across several network sizes, which can be characterized by maximizing the likelihood function averaged over the network sizes, N. We call this metric MeanMLE. Each N are log-spaced steps between which the network grows 10% until we reach the maximum network size. MeanMLE allows us to find parameters and models with the best overall fit to data, rather than the best at an arbitrary time point.

When fitting data, we discard model instances that will yield low likelihoods and remove models that time out computationally (take more than a few hours to run). We show in Fig. S8 in the Supplemental Material [18] that each realization can have clique frequencies vary wildly for LPAM, and the wide variance can, in turn, sometimes make calculating cliques computationally infeasible. This occurs rarely, however. For example, out of 150 000 instances across the three models used to fit Math Overflow, only 135 instances (0.09%) are discarded.

The LPAM, forest fire [3], and copying (Lambiotte *et al.*, [11]) models all have parameters constrained to lie between zero and one. The entire parameter range is taken when models are fit and the parameters are randomly realized and rounded to the nearest 0.01, with five realizations on average for each parameter value. For the forest fire and LPAM, there are two parameters whose range is between zero and one, therefore there are 5 × 101 × 101 or approximately 50 000 realizations for each dataset. In contrast, for the copying model, there is only one parameter and therefore 5 × 101 = 505 realizations.

III. RESULTS

We compare the statistics of several empirical graphs against all candidate models: The forest fire model [3], which was the first of two models to explain densification; the copying model [11], which provides theoretical predictions for clique scaling; and LPAM. While there are many other potential models one could compare against [12–15], our results show that LPAM captures basic aspects of network growth with a simple theoretically-grounded mechanism.

To test the importance of preferential attachment [4], we measure the mean degree of the target node's neighbors to which a new node connects, divided by the mean degree of all the target node's neighbors, averaged over all network sizes sampled. Preferential attachment would imply that this ratio is greater than one. In Fig. 2, we show our findings for all networks studied. While the copying model has a ratio of nearly one, implying no significant preferential attachment, the empirical data show a ratio significantly greater than one (strong preferential attachment) which is better captured with LPAM. See Fig. S4 in the Supplemental Material [18] for further support of the consistency of these results across different empirical datasets [14,19–22].

We also plot the mean distance between nodes before they connect to each other, and compare this distance to a null model (connecting between random nodes), as well as to the candidate models shown in Fig. 3 (similar plots are seen for other datasets [14,19–22] in Fig. S5 in the Supplemental Material [18]). We find that nodes form links to nearby nodes (the distance is smaller than the null model), while the models assume even closer distances—neighbors of neighbors, implying a distance of two. We therefore qualitatively capture the closeness of link formation, although the models tested do not fully address the links that are formed at a distance greater than two. Capturing these nuances are left for future work.

Furthermore, we explore how the different mechanisms capture the scaling exponents of different clique sizes. We show in Fig. 1(c) that exponents increase significantly with clique size, which is qualitatively captured from the copying

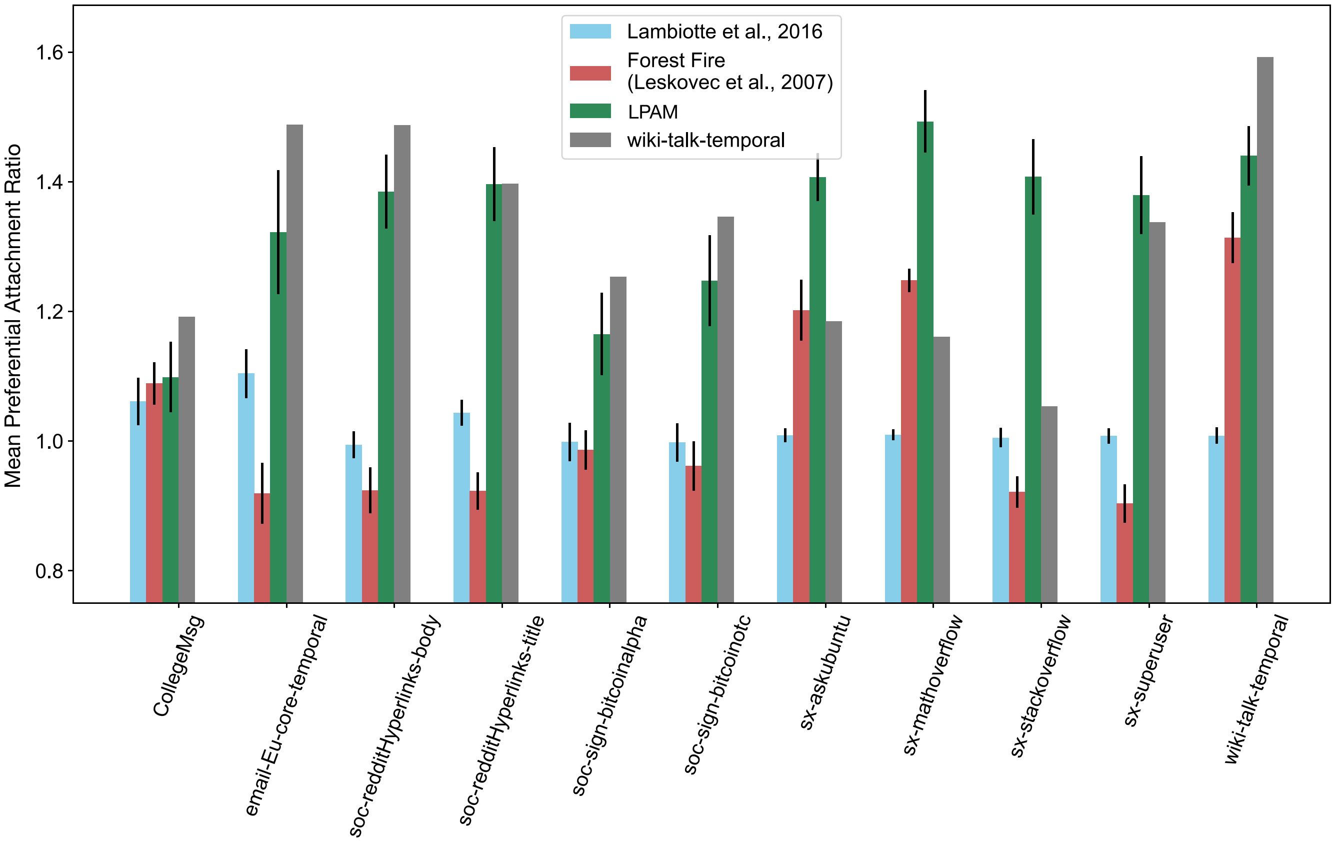

FIG. 2. Effect of preferential attachment. The y axis shows the mean degree of a target's neighbors that a new node connects to, averaged across all nodes at a given timestep, divided by the mean degree of all of the target's neighbors, again averaged over all nodes at a given timestep. When this ratio is greater than one, nodes preferentially connect to higher-degree neighbors. Empirical data (gray bars) are compared against the node copying mechanism (light blue) [10], the forest fire model [3], and LPAM. Datasets are College messages (CollegeMsg) [19]; emails at a large European institution (email-Eu-core-temporal) [14]; Reddit hyperlinks within the body of a Reddit post (soc-redditHyperlinks-body), or in the title (soc-redditHyperlinks-title) [20]; Bitcoin Alpha and Bitcoin OTC trust networks (soc-sign-bitcoinalpha and soc-sign-bitcoinotc) [21,22]; conversations on Ask Ubuntu (sx-askubuntu), Math Overflow (sx-mathoverflow), Stack Overflow (sx-stackoverflow), and Stack Exchange Super User (sx-superuser) boards [14]; and Wikipedia's talk pages (wiki-talk-temporal) [14]. Error bars are standard errors of this ratio across all sampled network sizes.

mechanism [10,11] but this model has a lower exponent than what we find empirically. (This is also consistent with what we find in other datasets [14,19–22], shown in Fig. S6 in the Supplemental Material [18].) LPAM, however, can better capture the scaling law exponents, and therefore help us understand why extremely dense cliques are unusually common in large networks. While the performance is comparable with the forest fire model, LPAM provides a clearer mechanism to explain this behavior. We also show in Fig. S3 in the Supplemental Material [18] that LPAM captures the mean clique size better than the competing models [3,11] for many datasets [14,19–22].

In order to determine the best overall model among these three, we take the mean Kullback-Leibler (KL) divergence [28] between model and empirical clique size distributions (details in Fig. S7 in the Supplemental Material [18]). We find that LPAM and the forest fire model have less error (lower KL divergence) than the copying model of Lambiotte *et al.* [11], which suggests that the Lambiotte *et al.* model may not fully capture how networks grow. Although LPAM can sometimes outperform the other models, we do not claim that another model cannot fit data even better. The main goal of our paper is to instead provide a theoretically-motivated mechanism beyond the copying model.

Finally, we can study ablation of LPAM either by removing node copying or removing preferential attachment. Setting $r = 0$, we remove traditional preferential attachment, and the model simplifies to the node copying model of Lambiotte *et al.* [11], a poorer-fitting model. Alternatively, we can remove node copying and have nodes connect to other nodes preferentially based on degree. This simplifies LPAM to the Barabasi-Albert model [4], whose degree is fixed independent of network size. Because neither simplification fits data as well, LPAM is an effective mechanism to reproduce the results we observe.

IV. CONCLUSION

We observe that cliques scale superlinearly with network size, therefore we observe strong patterns in the higher-order structure of networks. Moreover, we observe that scaling exponents vary significantly for large and small cliques in a growing network. We further observe nodes connect locally (e.g., to neighbors of neighbors) and confirm previous analysis that nodes have preferential attachment. We develop a mecha-

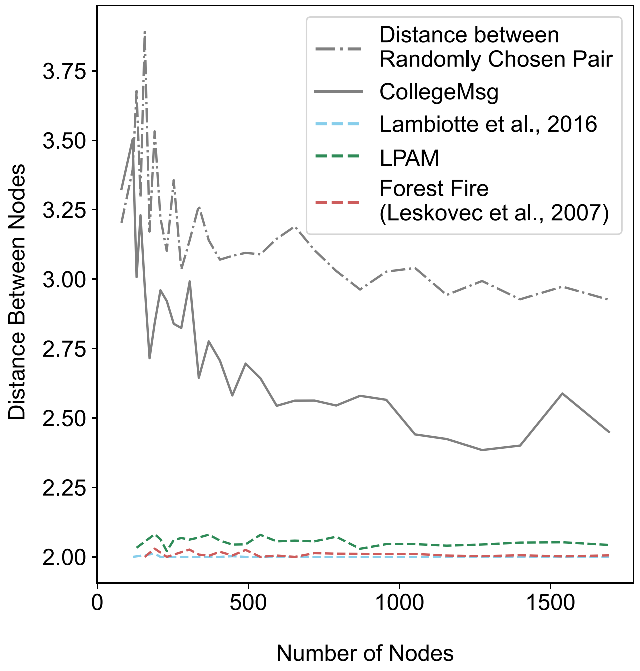

FIG. 3. Nodes make local connections. The distance between pairs nodes prior to forming a mutual connection with a new node. Distance between randomly chosen pairs in the College Messaging dataset [19], and the empirical distance between nodes prior to connecting to a mutual new node. Other examples show similar results, see Fig. S5 in the Supplemental Material [18] for other networks [14,19–22].

nism, LPAM, to explain these patterns. LPAM is an extension of previous mechanisms in which a new node attaches to a target node and preferentially to the target node's higher-degree neighbors. We carried out an ablation study to show this is one of the simplest mechanisms to explain the empirical patterns we measure.

There are a number of ways this method could be improved in future work. First, the mechanism is not theoretically grounded for cliques of order $k > 2$. Next, LPAM does not fully match empirical data, which is both a disadvantage and an advantage in that it greatly simplifies the rich complex patterns that each observational network encodes. We notice in Fig. S7 in the Supplemental Material [18], for example, that LPAM performs worse than or similarly to the competing forest fire model for small networks, such as the College Messages or cryptocurrency networks. This points to finite size effects that our model overlooks. Even when the model performs well, LPAM's exponents are often lower than the empirical data (Fig. S6 in the Supplemental Material [18] for other networks [14,19–22]), and the simulated nodes connect to closer neighbors than in empirical data. This motivates extensions of LPAM to address finite size effects and the strong relation between clique size and scaling exponent. One way to improve this model, which might address some of its limitations, includes having new edges connect between two old nodes in the network with some probability, which is similar to the Newman-Watts small world model [29]. Another way to improve the model could be to seed the model with a real network as an initial condition. Finally, we assume that the fitted scaling laws are asymptotic, but this needs to be tested with more networks, especially with sizes in the hundreds of millions to billions (which our current computing power cannot tolerate).

The code is available in Ref. [30].

ACKNOWLEDGMENTS

This research was sponsored, in part, by the Air Force Office for Scientific Research under contract FA9550-20-1-0224.

[1] P. Holme and J. Saramäki, Phys. Rep. **519**, 97 (2012), temporal Networks.

[2] K. A. Burghardt, Z. He, A. G. Percus, and K. Lerman, Commun. Phys. **4**, 189 (2021).

[3] J. Leskovec, J. Kleinberg, and C. Faloutsos, ACM Trans. Knowl. Discov. Data **1**, 2 (2007).

[4] A. Barabási and R. Albert, Science **286**, 509 (1999).

[5] H. Jeong, Z. Néda, and A. L. Barabási, Europhys. Lett. **61**, 567 (2003).

[6] S. A. Teichmann and M. M. Babu, Nat. Genet. **36**, 492 (2004).

[7] D. V. Foster, S. A. Kauffman, and J. E. S. Socolar, Phys. Rev. E **73**, 031912 (2006).

[8] M. Bell, S. Perera, M. Piraveenan, M. Bliemer, T. Latty, and C. Reid, Sci. Rep. **7**, 42431 (2017).

[9] J. Kunegis, D. Fay, and C. Bauckhage, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10 (Association for Computing Machinery, New York, NY, USA, 2010), pp. 739–748.

[10] U. Bhat, P. L. Krapivsky, R. Lambiotte, and S. Redner, Phys. Rev. E **94**, 062302 (2016).

[11] R. Lambiotte, P. L. Krapivsky, U. Bhat, and S. Redner, Phys. Rev. Lett. **117**, 218301 (2016).

[12] V. Zadorozhnyi and E. Yudin, in *Information Technologies and Mathematical Modelling*, edited by A. Dudin, A. Nazarov, R. Yakupov, and A. Gortsev (Springer International Publishing, Cham, 2014), pp. 432–439.

[13] L. Zalányi, G. Csárdi, T. Kiss, M. Lengyel, R. Warner, J. Tobochnik, and P. Érdi, Phys. Rev. E **68**, 066104 (2003).

[14] A. Paranjape, A. R. Benson, and J. Leskovec, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (ACM, Cambridge United Kingdom, 2017), pp. 601–610.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08* (ACM Press, Las Vegas, Nevada, USA, 2008), p. 462.

[16] A. K. Rizi, L. A. Keating, J. P. Gleeson, D. J. P. O'Sullivan, and M. Kivelä, arXiv:2304.10405.

[17] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).

[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.107.L042301 which includes reference for robustness analyses on a range of empirical networks.

[19] P. Panzarasa, T. Opsahl, and K. M. Carley, J. Am. Soc. Inf. Sci. **60**, 911 (2009).

[20] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, in *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (International World Wide Web Conferences Steering Committee, 2018), pp. 933–943.

[21] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (ACM, 2018), pp. 333–341.

[22] S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos, in *Data Mining (ICDM), 2016 IEEE 16th International Conference on* (IEEE, 2016), pp. 221–230.

[23] P. L. Krapivsky and S. Redner, Phys. Rev. E **71**, 036118 (2005).

[24] J. Chen and Y. Saad, IEEE Trans. Knowledge Data Eng. **24**, 1216 (2012).

[25] J. Leskovec and A. Krevl, SNAP Datasets: Stanford large network dataset collection, http://snap.stanford.edu/data (2014).

[26] R. A. Rossi, D. F. Gleich, A. H. Gebremedhin, and M. M. A. Patwary, in *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion (Association for Computing Machinery, New York, NY, USA, 2014), pp. 365–366.

[27] S. Jain and C. Seshadhri, The power of pivoting for exact clique counting, in *Proceedings of the 13th International Conference on Web Search and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2020), pp. 268–276.

[28] S. Kullback and R. A. Leibler, Ann. Math. Statist. **22**, 79 (1951).

[29] M. Newman and D. Watts, Phys. Lett. A **263**, 341 (1999).

[30] https://github.com/haochenpi314/Clique-Densification.