Degree Relations of Triangles in Real-world Networks and Graph Models

Nurcan Durak, Ali Pinar, Tamara G. Kolda, and C. Seshadhri
Sandia National Laboratories
Livermore, CA
{ndurak, apinar, tgkolda, scomand}@sandia.gov

ABSTRACT

Triangles are an important building block and distinguishing feature of real-world networks, but their structure is still poorly understood. Despite numerous reports on the abundance of triangles, there is very little information on what these triangles look like. We initiate the study of *degree-labeled triangles* — specifically, degree homogeneity versus heterogeneity in triangles. This yields new insight into the structure of real-world graphs. We observe that networks coming from social and collaborative situations are dominated by homogeneous triangles, i.e., degrees of vertices in a triangle are quite similar to each other. On the other hand, information networks (e.g., web graphs) are dominated by heterogeneous triangles, i.e., the degrees in triangles are quite disparate. Surprisingly, nodes within the top 1% of degrees participate in the *vast majority* of triangles in heterogeneous graphs. We investigate whether current graph models reproduce the types of triangles that are observed in real data and observe that most models fail to accurately capture these salient features.

1. INTRODUCTION

There is a growing interest in understanding the structure, dynamics, and evolution of large scale networks. Observing the similarities and differences among real-world networks improves graph mining in many aspects ranging from community detection to generation of more realistic random graphs.

A triangle is a set of three vertices that are pairwise connected and is arguably one of the most important patterns in terms of understanding the inter-connectivity of nodes in real graphs [18]. Note that the community structure is closely tied to triangles, and the degree behavior of triangles is an integral part of this structure [16]. Whether these graphs come from communication networks, social interaction, or the Internet, the presence of triangles is an indication of community behavior. In social networks, it is highly probable that friends of friends will themselves be friends, thus forming many triangles.

In this paper, we take a closer look at the structure of triangles, specifically, the degrees of the triangle vertices. How are the degrees of the vertices related? Do different degrees represent fundamentally different types of relationships and so appear in different sorts of networks? When we look at real-world networks, we may ask if there is a high incidence of *degree homogeneity*, wherein vertices of similar degree come together to form triangles? Or do triangles tend to show *degree heterogeneity*, i.e., connecting vertices of disparate degree?

1.1 Background and Previous Work

The notion of describing graph structure based on the frequency of small patterns such as triangles has been proposed under different names such as motifs [12], graphlets [14], and structural signatures [6]. Triangle counts form the basis for community detection algorithms in [2]. Triangles have also served as the driving force for generative models [16, 18].

The frequency of triangles is often measured using the *clustering coefficient*, as defined by Watts and Strogatz [18]. We first establish some notation. Consider an undirected graph G with n vertices. Let d_j denote the degree of node j and t_j denote the number of triangles incident to node j. If we define a wedge to be a path of length 2, then the number of wedges centered at node j is $\binom{d_j}{2}$. Now we can define various clustering coefficients. The clustering coefficient of vertex j, C_j, is defined as the number of triangles incident to j divided by the number of wedges centered at j, i.e., $C_j = t_j/\binom{d_j}{2}$. The average of clustering coefficients across all vertices (called the local clustering coefficient) is defined as $\overline{C} = \frac{1}{n}\sum_j C_j$. The (global) clustering coefficient, also known as the *transitivity*, is

$$C = \frac{3 \times \text{total number of triangles}}{\text{total number of wedges}} = \frac{\sum_j t_j}{\sum_j \binom{d_j}{2}}.$$

Most of the studies on degree-based similarity are based on *assortativity*, which was introduced by Newman [13]. Various studies have been conducted on the assortativity (or lack thereof) of real graphs [8, 19]. However, Newman's assortativity measure is misleading to classify networks with heavy-tailed degree distributions because it produces either neutral or negative assortativity (disassortativity) values for most of the large scale networks as shown in Table 1 (see r values).

Most relevant to our work is that of Tsourakakis [17], which observes that the average number of triangles per de-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

gree follows a power-law distribution and the slope of the degree-triangle plot has the negative slope of the degree distribution plot of the corresponding graph. It is argued that low degree nodes form fewer triangles than higher degree nodes. Our analysis shows that while this is certainly true for social networks, it does not hold for information networks, such as the autonomous systems networks.

1.2 Contributions

Our contributions fall into two categories. Our first set of results comes from empirical studies of degree relations in triangles of real graphs. Then, we perform experiments on a variety of graph models to show their (in)ability to reproduce the behavior of real graphs.

Triangle homogeneity vs heterogeneity: We take a collection of graphs from diverse scenarios (collaboration, social networking, web, infrastructure) and measure triangle degree relationships. We compute various correlations between degrees of vertices in triangles to understand the homogeneity of these triangles.

Our experiments show that graphs coming from social or collaborative scenarios are completely dominated by homogeneous triangles. There are a few heterogeneous triangles. This may not be surprising from a sociological viewpoint, since like should attract like. But graphs coming from web, routing, or communication are dominated by heterogeneous triangles. It is interesting that in communication or routing networks, the majority of triangles are formed by the vertices within the top 1% of degrees.

We observed that there is a high correlation between the global clustering coefficient, C, and the homogeneity tendency of triangles. Higher C values imply stronger homogeneity of triangles.

The triangle behavior of graph models: Our result can be stated quite succinctly. No existing graph model reflects the homogeneous and heterogeneous triangle behavior together. Many standard graph models like Preferential Attachment [1], Edge Copying [9], Stochastic Kronecker [10], etc. do not generate enough triangles and they cannot approximate the clustering coefficients of the real graphs [15]. The Chung-Lu [3] model cannot generate homogeneous graphs and cannot approximate the clustering coefficient of the high clustering coefficient networks. However, Chung-Lu is the only model that imitates the networks with low clustering coefficients and heterogeneous triangles.

The Forest Fire [11] and BTER [16] models generate a reasonable number of triangles (especially incident to low degree vertices) but these triangles are extremely homogeneous. Low degree vertices, when they participate in triangles, exclusively form triangles with other low degree vertices. This happens regardless of parameter choices, and is a fundamental property of these models. This shows that while they can qualitatively look like social and interaction networks, the behavior of heterogeneous networks cannot be reproduced by these models.

2. REAL-WORLD TRIANGLE BEHAVIOR

2.1 Data

We analyze the degree relations among vertices of triangles on a diverse set of graphs obtained from the SNAP database [20] and listed in Table 1. We have symmetrized the graphs by treating all edges as undirected, made each graph simple by removing self loops and parallel edges, and did not use edge weights. Cohen's algorithm [5] was used to enumerate all triangles.

In Table 1, we provide the following properties of the graphs we analyzed: N = number of nodes; E = number of edges; $\rho = E/N$ (density); C = global clustering coefficient; $\overline{C}$ = local clustering coefficient; T = number of triangles; α = power-law exponent, which is computed by fitting power-law distribution to degree distribution plots of the graphs [4]; κ_{90} and κ_{99}, respectively, are the 90th and 99th percentiles of degree of all nodes participating in triangles (i.e., we obtain all nodes participating in any triangle (each node is only counted once), put their degrees in a list, and then pick the 99th percentile of the degree list); $d_{\max}$ = maximum degree; and r = assortativity value.

In this paper, a network whose global clustering coefficient, C is greater than 0.01, is referred to as a *high-C* network; otherwise, it is a *low-C* network.

2.2 Analysis

We analyze the degree similarity of triangle vertices by grouping the triangles according to their minimum degree. We first present the notation. For $t = 1, \ldots, T$, let $d_{\min}(t)$, $d_{\mathrm{mid}}(t)$, and $d_{\max}(t)$ denote the minimum, middle, and maximum degree of the t-th triangle. For instance, if the t-th triangle has vertices of degrees 5, 10, and 4, then $d_{\min}(t) = 4$, $d_{\mathrm{mid}}(t) = 5$, and $d_{\max}(t) = 10$. Define $\mathcal{B}(i)$ to be the set of all triangles whose minimum degree is i, i.e.,

$$\mathcal{B}(i) = \{\, t \in T \mid d_{\min}(t) = i \,\}.$$

We then define some average statistics for each set $\mathcal{B}(i)$. Define $d_1(i)$, $d_2(i)$, and $d_3(i)$ to be the median of minimum, middle, and maximum degree, respectively, of triangles in $\mathcal{B}(i)$. In other words,

$$\begin{aligned} d_1(i) &= \text{median}\,\{\, d_{\min}(t) \mid t \in \mathcal{B}(i) \,\} = i, \\ d_2(i) &= \text{median}\,\{\, d_{\mathrm{mid}}(t) \mid t \in \mathcal{B}(i) \,\}, \\ d_3(i) &= \text{median}\,\{\, d_{\max}(t) \mid t \in \mathcal{B}(i) \,\}. \end{aligned}$$

For instance, if the triangles in $\mathcal{B}(2)$ were given by [2 2 3], [2 4 5], and [2 3 3]. Then, the $d_1(2) = 2$, $d_2(2) = 3$ and $d_3(2) = 3$.

To compare the relations among triangle degrees, we plot $d_2(i)$ and $d_3(i)$ versus $d_1(i)$ in Figure 1 and call them degree-comparison plots. Note that in these and all other log-log and semi-log plots, we use the exponential binning which is a standard procedure to de-noise the data when plotting on logarithmic scale. The degree-comparison plots of the rest of the graphs can be found in our extended paper [7].

2.3 Observations

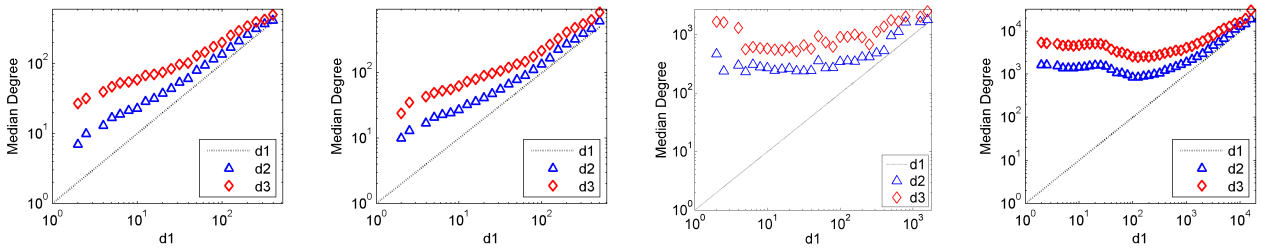
By considering the degree relations of the triangle vertices, we make the following observations.

Observation 1: *The global clustering coefficient is an indicator for the triangle degree relations.*

In Figure 1, we can see a clear relation between global clustering coefficient C and the type of triangles. In *high-C* networks, minimum, middle, and maximum degrees of triangle vertices are close in value. While, in *low-C* networks, triangles are highly heterogeneous. Observe how very small values of $d_1(i)$ connect to quite large $d_2(i)$ or $d_3(i)$ in Figure 1c and Figure 1d.

Table 1: Properties of networks we analyzed.

	Graph Name	N	E	ρ	C	$\overline{C}$	T	α	κ_{90}	κ_{99}	$d_{\max}$	r
high-C	amazon0312	400K	2,349K	5.9	0.260	0.41	3,686K	3.1	19	55	2747	-0.02
	ca-AstroPh	18K	198K	11	0.318	0.63	1,351K	1.52	56	145	504	0.2
	cit-HepPh	34K	420K	12	0.146	0.30	1,276K	1.53	56	147	846	0
low-C	as-caida20071105	26K	53K	2	0.007	0.21	36K	1.52	12	99	2628	-0.19
	oregon1_010331	10K	22K	2.1	0.009	0.45	17K	1.5	10	839	2312	-0.18
	wiki-Talk	2,394K	4,659K	1.9	0.002	0.20	9,203K	1.67	21	401	100029	-0.06



(a) ca-AstroPh (b) cit-HepPh (c) as-caida20071105 (d) wiki-Talk

Figure 1: Triangle degree-comparison plots which compare the minimum degree, $d_1(i)$, and the medians of the middle degree, $d_2(i)$, and the maximum degree, $d_3(i)$

The average clustering coefficient $\overline{C}$ is not a very distinguishing metric for our study. The global clustering coefficient C shows wide variance and is a better indicative of the triangle behavior.

Observation 2: *The ratios among degrees of triangle vertices are small in* high-C *networks and large in* low-C *networks.*

The ratios of triangle degrees provide valuable information to see the distinction between networks. For the t-th triangle, three degree ratios are defined as follows.

$$r_{21}(t) = \frac{d_{\text{mid}}(t)}{d_{\min}(t)},\ r_{31}(t) = \frac{d_{\max}(t)}{d_{\min}(t)},\ \text{and}\ r_{32}(t) = \frac{d_{\max}(t)}{d_{\text{mid}}(t)}$$

These ratios are computed for all the triangles separately and their averages are taken as $\bar{r}_{21}$, $\bar{r}_{31}$, and $\bar{r}_{32}$, respectively.

Based on the ratios among degrees, we define homogeneity measure $h = \frac{\bar{r}_{32}}{\bar{r}_{31}}$ to discriminate networks with homogenous triangles from the networks with heterogenous triangles. Table 2 lists the average ratios for all the networks. There

Table 2: The average of triangle degree ratios

	Graph Name	$\bar{r}_{21}$	$\bar{r}_{31}$	$\bar{r}_{32}$	h
high-C	amazon0312	1.98	4.95	2.53	0.51
	ca-AstroPh	1.88	3.46	1.89	0.54
	cit-HepPh	2.20	4.96	2.38	0.48
low-C	as-caida20071105	70.99	164.35	8.14	0.05
	oregon1_010331	54.80	175.69	9.09	0.05
	wiki-Talk	42.64	138.01	4.75	0.03

is a clear distinction between *high-C* and *low-C* networks. The average ratios are very small in *high-C* networks, which also supports the triangle homogeneity in *high-C* networks. Whereas, the average degree ratios ($\bar{r}_{21}$ and $\bar{r}_{31}$) are significantly large in *low-C* networks. Homogeneity measures h are very small for *low-C*, whereas h values are around 0.5 for *high-C* networks.

Observation 3: *In* low-C *networks, high degree vertices within the top 1% participate in the vast majority of the triangles.*

In *high-C* networks, the triangles incident to low degree vertices are mostly connecting to two low degree vertices. On the other hand, in *low-C* networks (particularly when ρ is low), a significant portion of the triangles contain at least one high degree vertex.

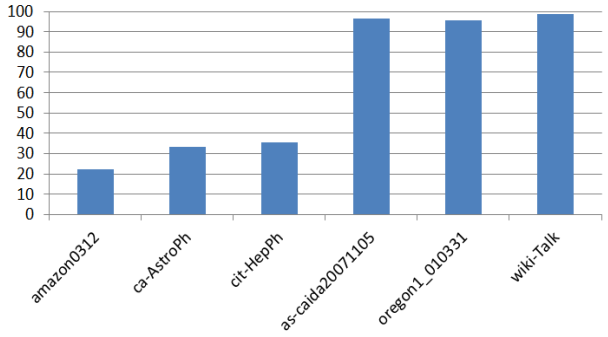


Figure 2: The percentage of triangles produced by vertices in the Top 1% degrees

To set a threshold between low degrees and high degrees, we have experimented different percentiles of vertices that participate in at least one triangle. We pick κ_{99} (see Table 1) as a threshold, since κ_{90} is still relatively low compared to

the maximum degree in most networks. A degree of a triangle vertex is considered *high*, if the degree is greater than κ_{99}, otherwise it is considered *low*.

We look at the percentages of the triangles having at least one *high* degree node in Figure 2. In *low-C* networks, we can see that high degree vertices within the top 1% participate in most of the triangles. In *high-C*, *high* degree nodes are participating in fewer triangles.

3. GRAPH-MODEL TRIANGLE BEHAVIOR

In this section, we investigate how well random graph generators match the real graphs in terms of triangle degree similarity. We concentrate on the graph models generating heavy-tailed degree distributions.

3.1 Graph Models

The Chung-Lu (CL) model [3] is a variant of the classic Erdős-Rényi model for any degree distribution. In this model, the probability of inserting an edge is proportional to the product of the degrees of its endpoints, (i.e., $Pr(e_{ij}) = \frac{d_i d_j}{(2E)}$).

The Block Two-Level Erdős-Rényi (BTER) model [16] is built on the observation of high-clustering coefficients and skewed degree distributions. This model achieves high clustering coefficients by embedding communities with an Erdős-Rényi structure, which is typically much denser compared to the rest of the graph. Additional edges are added in a subsequent phase using the CL model, to satisfy the degree distribution requirements. It has been shown that BTER graphs can match many properties of real world graphs [16].

The Forest Fire (FF) model [11] combines the Preferential Attachment model [1] to obtain a heavy-tailed degree distribution, the Edge Copying model [9] to obtain communities, and community guided attachment for densification.

3.2 Triangle Analysis in Graph Models

To check whether graph models can reproduce the triangle degree behavior of the real networks, we fit FF, BTER, CL, the Edge Copying (EC) [9], the Preferential Attachment (PA) [1], and the Stochastic Kronecker Graph (SKG) [10] models to the real networks listed in Table 1. The details of fitting graph models to the real networks can be found in [7]. We enumerate triangles in each randomly generated network using Cohen's algorithm [5], and we analyze the triangle behaviors in these random graphs from different aspects.

The numbers of triangles: *None of the graph models capture the triangle numbers for* both high-*C and* low-*C networks.*

Graph models behave differently in *high-C* and *low-C* networks in terms of generating triangles. The BTER is good at generating similar number of triangles for *high-C* networks, the FF and CL are good at for *low-C* networks, but none of the graph models is good at both. The number of triangles generated by different graph models for each target graph is listed in Table 3.

EC, PA, and SKG generate significantly less triangles than the original triangle numbers. These models also cannot reach the average clustering coefficient per degree for any of the networks. Therefore, we will not include them for the rest of the plots.

Degree Relations: *Models generate only one type of triangles without distinguishing* low-*C or* high-*C networks.*

In Figure 3, we show the relation between $d_1(i)$ and $d_2(i)$ for the real graphs as well as their modeled counterparts. The relation between $d_1(i)$ vs $d_3(i)$ acts similarly as shown in the long version of this paper [7].

CL produces *heterogeneous* triangles for both *high-C* and *low-C* networks in both Figure 3. For *low-C* networks, it is very intriguing that CL graphs are generating the right *type* of triangles. But we feel that this indicates that *low-C* networks have a CL flavor to them (i.e., triangles are random). BTER generates homogeneous triangles for both *high-C* and *low-C* networks. For *high-C* networks, BTER generates lower $d_2(i)$ and $d_3(i)$ values than original $d_3(i)$ values. FF behaves like BTER for *low-C* networks. Low degree $d_1(i)$ values cannot connect to high degree vertices. Distance between FF's $d_2(i)$ and original $d_2(i)$ is considerable large. FF also reaches higher $d_1(i)$ values than the original $d_1(i)$ values.

4. CONCLUSIONS

The abundance of triangles in real-word networks have been the subject of many studies, and is recognized as an important feature of real networks. In this work, we went one step further than looking at merely the number of triangles and analyzed the degree relations between the vertices of triangles in real-world networks. Our experiments showed that degrees of triangle vertices are either homogenous or heterogeneous in different networks, and the global clustering coefficient is a good indicator of the type of triangles in a network.

We have also investigated whether the current graph models can regenerate the types of triangles in the real data and showed that none of the graph models are able to capture both homogenous or heterogeneous triangles together. Our results clearly point to a deficiency in current models to create both social and communication networks. Our observations will be helpful for designing realistic graph models supporting the triangle degree behavior in community structures.

Acknowledgments

This work was funded by the Defense Advanced Research Projects Agency (DARPA), the Applied Mathematics Program at the U.S. Department of Energy (DOE), and an Early Career Award from the Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

5. REFERENCES

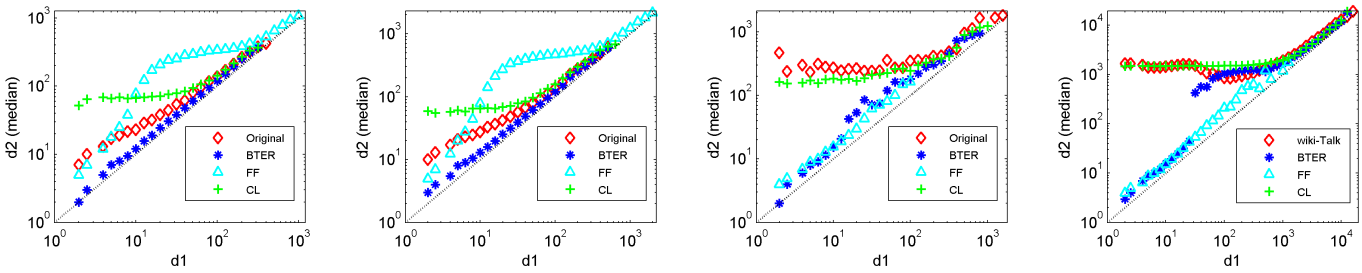
[1] Barabási, A.-L., and Albert, R. Emergence of scaling in random networks. *Science 286*, 5349 (1999), 509–512.

[2] Berry, J. W., Hendrickson, B., LaViolette, R. A., and Phillips, C. A. Tolerating the community detection resolution limit with edge weighting. *Phys. Rev. E 83* (May 2011), 056119.

[3] Chung, F., and Lu, L. The average distances in random graphs with given expected degrees. *PNAS 99* (2002), 15879–15882.

Table 3: The number of triangles generated by graph models

	Graph Name	Original	BTER	CL	FF	EC	PA	SKG
high-C	amazon0312	3,686K	3,704K	5K	4,420K	12K	10K	12K
	ca-AstroPh	1,351K	1,315K	49K	2,937K	43K	20K	4K
	cit-HepPh	1,276K	1,315K	48K	8,502K	180K	40K	34K
low-C	as-caida20071105	36K	74K	43K	38K	3K	< 1K	3K
	oregon1_010331	17K	26K	15K	17K	1K	< 1K	< 1K
	wiki-Talk	9,203K	66,740K	41,427K	2,936K	16K	< 1K	< 1K



(a) ca-AstroPh (b) cit-HepPh (c) as-caida20071105 (d) wiki-Talk

Figure 3: Triangle degree relations between $d_1(i)$ and $d_2(i)$ in the generated graph models

[4] Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review 51*, 4 (2009), 661–703.

[5] Cohen, J. Graph twiddling in a MapReduce world. *Computing in Science & Engineering 11* (2009), 29–41.

[6] Contractor, N. S., Wasserman, S., and Faust, K. Testing multitheoretical organizational networks: An analytic framework and empirical example. *Academy of Management Review 31*, 3 (2006), 681–703.

[7] Durak, N., Pinar, A., Kolda, T. G., and Seshadhri, C. Degree relations of triangles in real-world networks and graph models. arXiv:1207.7125, 2012.

[8] Holme, P., and Zhao, J. Exploring the assortativity-clustering space of a network's degree sequence. *Phys. Rev. E 75* (Apr. 2007), 046111.

[9] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. The web as a graph: measurements, models, and methods. In *Proc. COCOON'99* (1999), Springer-Verlag, pp. 1–17.

[10] Leskovec, J., and Faloutsos, C. Scalable modeling of real graphs using Kronecker multiplication. In *Proc. ICML'07* (2007), ACM, pp. 497–504.

[11] Leskovec, J., Kleinberg, J., and Faloutsos, C. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. KDD'05* (2005), ACM, pp. 177–187.

[12] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs: Simple building blocks of complex networks. *Science 298*, 5594 (2002), 824–827.

[13] Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Letter 89* (May 2002), 208701.

[14] Prdulj, N. Biological network comparison using graphlet degree distribution. *Bioinformatics 23*, 2 (2007), e177–e183.

[15] Sala, A., Cao, L., Wilson, C., Zablit, R., Zheng, H., and Zhao, B. Y. Measurement-calibrated graph models for social network experiments. In *Proc. WWW'10* (2010), ACM, pp. 861–870.

[16] Seshadhri, C., Kolda, T. G., and Pinar, A. Community structure and scale-free collections of Erdős-Rényi graphs. *Phys. Rev. E 85*, 056109 (May 2012).

[17] Tsourakakis, C. Fast counting of triangles in large real networks without counting: Algorithms and laws. In *Proc. ICDM'08* (2008), pp. 608–617.

[18] Watts, D., and Strogatz, S. Collective dynamics of 'small-world' networks. *Nature 393* (1998), 440–442.

[19] Whitney, D. E., and Alderson, D. Are technological and social networks really different? In *Unifying Themes in Complex Systems*, A. Minai, D. Braha, and Y. Bar-Yam, Eds. Springer, 2010, pp. 74–81.

[20] Stanford Network Analysis Project (SNAP). Available at `http://snap.stanford.edu/`.