# Same Stats, Different Graphs

## (Graph Statistics and Why We Need Graph Drawings)

Hang Chen[1], Utkarsh Soni[2], Yafeng Lu[2],
Ross Maciejewski[2], Stephen Kobourov[1]

[1]University of Arizona, Tucson, AZ, USA
[2]Arizona State University, Tempe, AZ, USA

**Abstract.** Data analysts commonly utilize statistics to summarize large datasets. While it is often sufficient to explore only the summary statistics of a dataset (e.g., min/mean/max), Anscombe's Quartet demonstrates how such statistics can be misleading. Graph mining has a similar problem in that graph statistics (e.g., density, connectivity, clustering coefficient) may not capture all of the critical properties of a given graph. To study the relationships between different graph properties and statistics, we examine all small non-isomorphic graphs and provide a simple visual analytics system to explore correlations across multiple graph properties. However, for graphs with more than ten nodes, generating the entire space of graphs becomes quickly intractable. We use different random graph generation methods to further look into the distribution of graph statistics for higher order graphs and investigate the impact of various sampling methodologies. We also describe a method for generating many graphs that are identical over a number of graph properties and statistics yet are clearly different and identifiably distinct.

**Keywords:** Graph mining, graph properties, graph generators

## 1 Introduction

Statistics are often used to summarize a large dataset. In a way, one hopes to find the "most important" statistics that capture one's data. For example, when comparing two countries, we often specify the population size, GDP, employment rate, etc. The idea is that if two countries have a "similar" statistical profile, they are similar (e.g., France and Germany have a more similar demographic profile than France and USA). However, Anscombe's quartet [3] convincingly illustrates that datasets with the same values over a limited number of statistical properties, can be fundamentally different – a great argument for the need to visualize the underlying data; see Fig. 1.

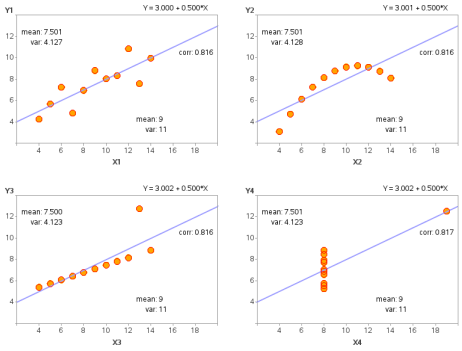


**Fig. 1.** Anscombe's quartet: all four datasets have the same mean and st. deviation in $x$ and $y$ and $(x, y)$-correlation.

Similarly, in the graph analytics community, a variety of statistics are being used to summarize graphs, such as graph density, average path length, global clustering coefficient, etc. However, summarizing a graph with a fixed set of graph statistics leads to the problem illustrated by Anscombe. It is easy to construct several graphs that have the same basis statistics (e.g., number of vertices, number of edges, number of triangles, girth, clustering coefficient) while the underlying graphs are clearly different and identifiably distinct; see Fig. 2. From a graph theoretical point of view these graphs are very different: they differ in connectivity (from 2-connected, to 1-connected, to disconnected), in planarity, symmetry, and in other structural properties.

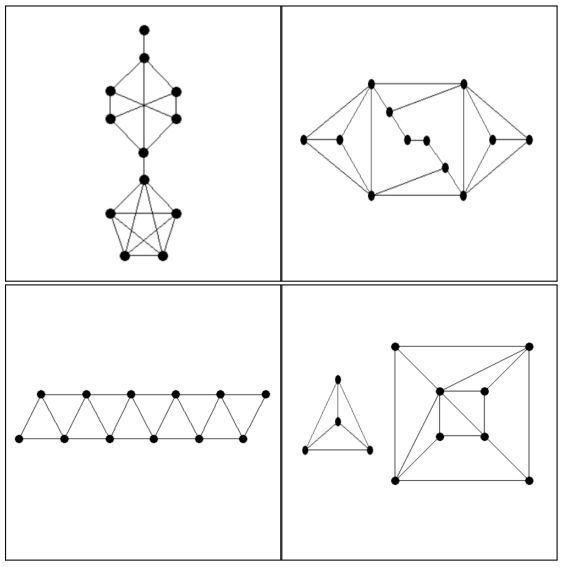

**Fig. 2.** These four graphs share the same 5 common statistics: number of vertices $|V| = 12$, number of edges $|E| = 21$, number of triangles $|\triangle| = 10$, girth $= 3$ and global clustering coefficient $GCC = 0.5$. However, structurally the graphs are very different: three of the graphs are planar and the fourth one is not, some show regular patterns and are symmetric, others are not. Finally, one of the graphs is disconnected, another is 1-connected and the rest are 2-connected.

Recently, Matejka and Fitzmaurice [30] proposed a dataset generation method that can modify a given 2-dimensional point set (like the ones in Anscombe's quartet) while preserving its summary statistics but significantly changing its visualization (what they call "graph"). Given the graphs in Fig. 2, we consider whether it is also possible to modify a given graph and preserve a given set of summary statistics while significantly change other graph properties and statistics. Note that the problem is much easier for 2D point sets and basic statistics, such as mean, deviation and correlation, than for graphs where many graph properties are structurally correlated (e.g., diameter and average path length). With this in mind, we first consider how can we fix a few graph statistics (such as number of nodes, number of edges, number of triangles) and vary another statistic (such as clustering coefficient or connectivity). We find that there is a spectrum of possibilities: sometimes the "unrestricted" statistic can vary dramatically, sometimes not, and the outcome depends on two issues: (1) the inherent correlation between some statistics (e.g., density and number of triangles), and; (2) the bias in graph generators.

We begin by studying the correlation between graph summary statistics across the set of all non-isomorphic graphs with up to 10 vertices. The statistical properties derived for all graphs for a fixed number of vertices provide further information about certain "restrictions." In other words, the range of one statistic may be restricted if another statistical property is fixed. However, we cannot explore the entire space of graph statistics and correlations. As the number of vertices grows, the number of different non-isomorphic graphs grows super-exponentially. For $|V| = 1, 2 \ldots 9$ the numbers are $1, 2, 4, 11, 34, 156, 1044, 12346$, 274668, but already for $|V| = 16$ we have $6 \times 10^{22}$ non-isomoprhic graphs.

To go beyond ten vertices we use graph generators based on classic models, such as Erdös-Rényi and Watts-Strogatz. However, different graph generators have different biases and these can significantly impact the results. We study the extent to which sampling using random generators can represent the whole graph set for an arbitrary number of vertices with respect to their coverage of the graph statistics. One way to evaluate the performance of random generators is based on the ground-truth graph sets that are available: all non-isomorphic graphs for $|V| \leq 10$ vertices. If we randomly generate a small set of graphs (also for $|V| \leq 10$ vertices) using a given graph generator, we can explore how well the sample and generator cover the space of graph statistics. In this way, we can begin exploring the issues of "same stats, different graphs" for larger graphs.

All data and tools are available at http://vader.lab.asu.edu/GraphAnalytics/. In particular, we have a basic visual analytics system and basic exploration tools for the space of all small non-isomorphic graphs and sampled larger graphs. We also include a generator for "same stats, different graphs," i.e., multiple graphs that are identical over a number of graph statistics, yet are clearly different.

## 2 Related Work

We briefly review the graph mining literature, paying special attention to the commonly collected graph statistics. We also consider different graph generators.

**Graph Statistics:** Graph mining is applied in different domains from bioinformatics and chemistry, to software engineering and social science. Essential to graph mining is the efficient collection of various graph properties and statistics that can provide useful insight about the structural properties of a graph. A review of recent graph mining systems identified some of the most frequently extracted statistics. We list those, along with their definitions, in Table 1. These properties range from basic, e.g., vertex count and edge count, to complex, e.g., clustering coefficients and average path length. Many of them can be used to derive further properties and statistics. For example, graph density can be determined directly as the ratio of the number of edges $|E|$ to the maximum number of edges possible $|V| \times (|V| - 1)/2$, and real-world networks are often found to have a low graph density [32]. Node connectivity and edge connectivity measures may be used to describe the resilience of a network [8, 28], and graph diameter [22] captures the maximum among all pairs of shortest paths [2, 7].

Other graph statistics measure how tightly nodes are grouped in a graph. For example, clustering coefficients have been used to describe many real-world networks, and can be measured locally and globally. Nodes in a highly connected

**Table 1.** The set of graph statistics considered in this paper.

| Name | Formula | Reference |
|---|---|---|
| Average Clustering Coefficient | $ACC(G) = \frac{1}{n}\sum_{i=1}^{n} c(u_i), u_i \in V, n = \lvert V\rvert$<br>$c(v) = \frac{\lvert\{(u,w)\lvert u,w\in\Gamma(v),(u,w)\in E\}\rvert}{\lvert\Gamma(v)\rvert(\lvert\Gamma(v)\rvert-1)/2}, v,u,w \in V$ | [10, 26, 24, 9, 33] |
| Global Clustering Coefficient | $GCC(G) = \frac{3\times\lvert triangles\rvert}{\lvert connected\ triples\rvert\ in\ the\ graph}$ | [9, 24] |
| Square Clustering | $SCC(G) = \frac{\sum_{u=1}^{k_v}\sum_{w=u+1}^{k_v} q_v(u,w)}{\sum_{u=1}^{k_v}\sum_{w=u+1}^{k_v}[a_v(u,w)+q_v(u,w)]}$ | [27] |
| Average Path Length | $APL = ave\{\frac{n-1}{\sum_{v\in V} d(u,v), u\neq v}\}$ | [10, 26, 9, 33] |
| Degree Assortativity | $r = \frac{\sum_{xy} xy(e_{xy}-a_x b_y)}{\sigma_a\sigma_b}$ | [35, 33] |
| Diameter | $diam(G) = max\{dist(v,w), v,w \in V\}$ | [10, 31, 24, 33] |
| Density | $den = \frac{2\lvert E\rvert}{\lvert V\rvert(\lvert V\rvert-1)}$ | |
| Ratio of Triangles | $\mathrm{Rt} = \frac{\lvert triangles\rvert}{\lvert V\rvert(\lvert V\rvert-1)/2}$ | |
| Node Connectivity | Cv: the minimum number of nodes to remove to disconnect the graph | [16] |
| Edge Connectivity | Ce: the minimum number of edges to remove to disconnect the graph | [16] |

clique tend to have a high local clustering coefficient, and a graph with clear clustering patterns will have a high global clustering coefficient [17, 18, 25, 36]. Studies have shown that the global clustering coefficient has been found to be nearly always larger in real-world graphs than in Erdös-Rényi graphs with the same number of vertices and edges [9, 36, 41], and a small-world network should have a relatively large average clustering coefficient [11, 14, 43]. The average path length (APL) is also of interest; small-world networks have APL that is logarithmic in the number of vertices, while real-world networks have small (often constant) APL [11, 14, 36, 41–43].

Degree distribution is one frequently used property describing the graph degree statistics. Many real-world networks, including communication, citation, biological and social networks, have been found to follow a power-law shaped degree distribution [5, 9, 36]. Other real world networks have been found to follow an exponential degree distribution [20, 39, 44]. Degree assortativity is of particular interest in the study of social networks and is calculated based on the Pearson correlation between the vertex degrees of connected pairs [34]. A random graph generated by Erdös-Rényi model has an expected assortative coefficient of 0. Newman [34] extensively studied assortativity in real-world networks and found that social networks are often assortative (positive Assortativity), i.e., vertices with a similar degree preferentially connect together, whereas technological and biological networks tend to be disassortative (negative Assortativity) implying that vertices with a smaller degree tend to connect to high degree ver-

tices. Assortativity has been shown to affect clustering [29], resilience [34], and epidemic-spread [6] in networks.

**Graph Generators:** Basic graph statistics have been used to describe various classes of graphs (e.g., geometric, small-world, scale-free) and a variety of algorithms have been developed to automatically generate graphs that mimic these various properties. Charkabati et al. [10] divide graph models and generators into four broad categories:

1. Random Graph Models: The graphs are generated by a random process.
2. Preferential Attachment Models: In these models, the "rich get richer," as the network grows, leading to power law effects.
3. Optimization-Based Models: Here, power laws are shown to evolve when risks are minimized using limited resources.
4. Geographical Models: These models consider the effects of geography (i.e., the positions of the nodes) on the topology of the network. This is relevant for modeling router or power grid networks.

The Erdös-Rényi (ER) network model is a simple graph generation model [9] that generates graphs either by choosing a network randomly with equal probability from a set of all possible networks of size $|V|$ with $|E|$ edges [19] or by creating each possible edge of a network with $|V|$ vertices with a given probability $p$ [15]. The latter process gives a binomial degree distribution and it can be approximated with a Poisson distribution. It is also possible to create networks where the degree follows other common probability distributions, such as exponential [12] or Gaussian [23]. Networks with any given degree sequence can be generated using the configuration model [36].

Watts and Strogatz [43] developed the classical approach for generating small-world graphs. This model could create disconnected graphs, but the variation suggested by Newman and Watts [37] generates connected graphs. Models have also been proposed for generating synthetic scale-free networks with a varying scaling exponent($\gamma$). The first scale-free directed network model was given by de Solla Price [38]. Barabasi and Albert (BA) [4] described another popular network model for generating undirected networks. It is a network growth model in which each added vertex has a fixed number of edges $|E|$, and the probability of each edge connecting to an existing vertex $v$ is proportional to the degree of $v$. Dorogovtsev et al. [13] and Albert and Barabasi [1] also developed a variation of the BA model with a tunable scaling exponent.

## 3 Preliminary Experiments and Findings

In a recent study of the ability to perceive different graph properties (e.g., edge density, clustering coefficient) in different types of graph layouts (e.g., force-directed, circular) we generated a large number of graphs with 100 vertices. Specifically, we generated graphs that vary in a controlled way in edge density and graphs that vary in a controlled way in the average clustering coefficient [40]. A post-hoc analysis of this data (http://vader.lab.asu.edu/GraphAnalytics/), reveals some interesting patterns among the statistics described in Table 1.

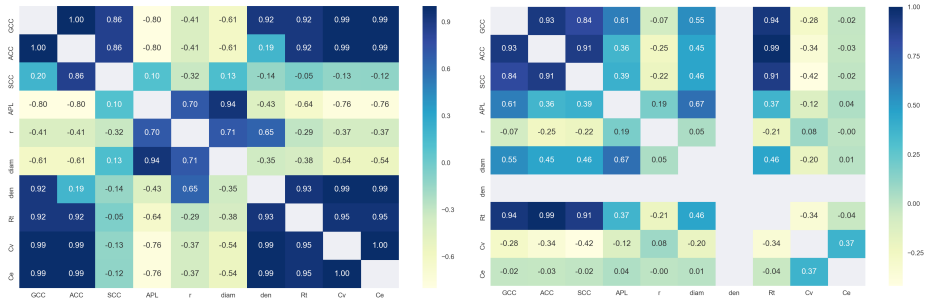


**Fig. 3.** Graph property correlation matrix plots for the edge density dataset (left) and the average clustering coefficient dataset (right).

The edge density dataset has 4,950 graphs and the average clustering coefficient dataset has 3,450 graphs. We compute all ten statistics from Table 1 and compute Pearson correlation coefficients; see Fig 3. We observed high correlations either positive (blue) or negative (yellow) in many property pairs. For example, the average clustering coefficient is highly correlated with the global clustering coefficient, the number of triangles, the edge density, and graph connectivity. The average clustering coefficient dataset also shows some strong correlations. For example the number of triangles is strongly correlated with local and global clustering coefficients, but also with the square clustering coefficient, that last of which is somewhat surprising.

However, these graphs were developed for a very specific purpose and cover only limited space of all graphs with $|V| = 100$. Two factors likely impact these correlations: the generator used, and the statistical properties that were controlled (e.g., insuring the number of edges remained constant). For the edge density graph set, we used a random graph generation algorithm that varies the number of edges while keeping a uniform probability of connecting two nodes. For the average clustering coefficient graph set, we controlled graph density and degree distribution (power-law). This motivated us to conduct the following experiments:

1. Generate all isomorphic lower order graphs ($|V| \leq 10$) and analyze the relationships between statistical properties. We consider this type of data as ground truth due to its completeness.
2. Use popular graph generators to create a sample of graphs and compare the graph property distributions of these outputs to the ground truth. This depicts the characteristics of different graph generators in terms of graph property coverage and helps us to think about which generator we should use in different scenarios.

## 4 Analysis of Graph Statistics

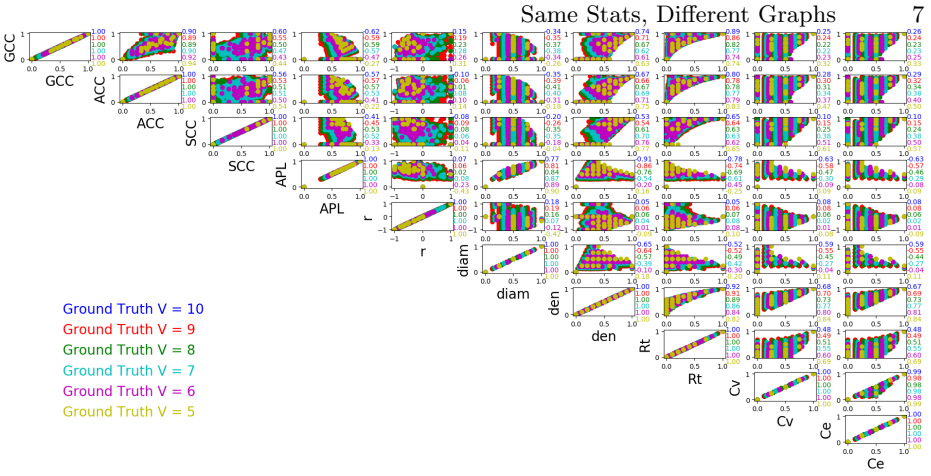


**Fig. 4.** Correlations between graph statistics in the ground truth for $|V| = 5, 6, 7, 8, 9$. Note that for $|V| = 9$ there are already 274,668 points.Points are plotted to overlap, with the largest sets plotted first (i.e., $|V| = 9, ...|V| = 5$) to enable us to identify the range of statistics that can be covered with a given number of vertices.

We compute all ten statistics for all non-isomorphic graphs for $|V| = 4, 5, \ldots, 10$ (note that for $|V| = 1, 2, 3$ many of the statistics are not well defined and there are only a handful of graphs). We then consider the pairwise correlations between the different statistics as the number of vertices in the ground truth dataset increases; see Fig 4. To compare the coverage of statistics with different $|V|$, we scale the statistic values into the same range. The clustering coefficients (ACC, GCC, SCC) are in $[0, 1]$ by definition and degree assortativity is in $[-1, 1]$. We keep their values and ranges without scaling. For the following five statistics, edge density, number of triangles, diameter and the connectivity measures ($C_v$ and $C_e$), we normalize them into $[0, 1]$ by dividing them by the corresponding maximum value. The last statistic, APL, is more challenging. We compute the exact value for our ground truth

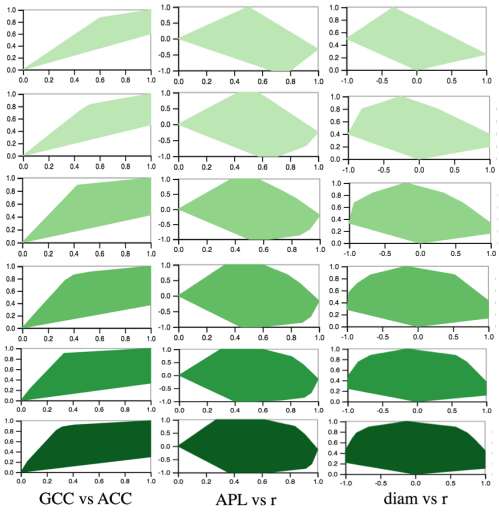


**Fig. 5.** The convex hull of graph coverage across several statistical properties. Each row (starting from the top) represents all graphs for a fixed number of vertices ($|V| = 5...|V| = 10$). Columns are pairs of graph properties.

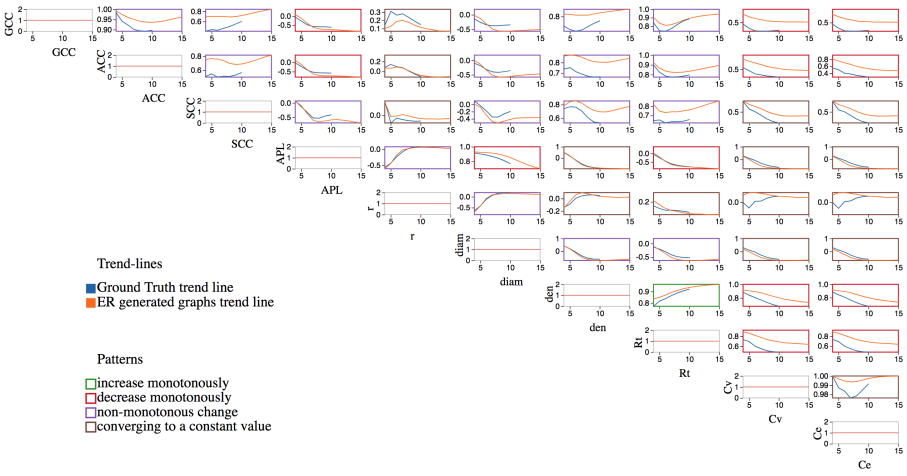


**Fig. 6.** Trends in the correlations with increasing $|V|$: the $x$-axis shows the number of vertices and the $y$-axis shows the correlation value for the pair of graph statistics.

datasets, but not when we use the generators, where we use the maximal value encountered (which may not be the same as the maximum).

It is easy to see that the coverage of values expands with increasing $|V|$. Figure 5 shows this pattern for three pairs of properties. This indicates that we are more likely to find larger ranges of different statistics (when a set of other statistics are fixed) for graphs with more vertices. With this in mind, we consider graphs with more than 10 vertices, but this time relying on random graph generators. Figure 6 shows how correlation values between all pairs of statistics change when the number of vertices increases. The blue trend lines for the ground truth data show the correlation values calculated using the set of all possible graphs for a given number of nodes. The orange trend lines show the correlation values calculated from graphs generated with the ER model.

For most the cells in the matrix shown in Figure 6, the correlation values seem to converge as $|V|$ becomes larger than 8 (both in the ground truth and the ER-model generated graph sets). Moreover, for most of the cells, the pattern of the change in correlation values appears to be the same for both sets. Analyzing the trend lines of the ER-model, we observe four patterns of change in the correlation values (also annotated in Figure 6 by an enclosing colored box): convergence to a constant value, monotonic decrease, monotonic increase, and non-monotonic change. There are exceptions that do not fit these patterns, e.g., ($S_c$, r) and in two cases, (r, $C_v$) and (r, $C_e$), the trend lines show different patterns.

## 5 Graph Statistics and Graph Generators

We select four different random generators that cover most fundamental types: the ER random graph model, the WS small-world model, the BA preferential attachment model, and the geometric random graph model. We employ a general strategy: start by fixing one statistic $s_1$ and check what other statistics can vary.

If the generators give broad distribution ranges for the other statistics, then we fix $s_1$. The first statistic we considered is density. We fix $|V| = 50$ and fix density at different values starting from 0.01 and progressively increasing it; for each density value we generate 100,000 graphs with that density using all four random graph generators. Given a fixed number of vertices and fixed number of edges, we consider the GCC statistic. We did not anticipate that fixing density will restrict GCC as they are not strongly correlated, but our results were surprising. We failed to find a fixed value for edge density that allows GCC to vary across its range from 0 to 1 for any of the four graph generators, although in the ground truth we have many such examples.

The result above implies that our generators (or the way we are using them) are not sampling the ground-truth graph set well. Thus, before we fix one statistic we should find out why our graph generators are not matching what we find in the ground truth. It is easy to compare the ground-truth graph set with a set of graphs created by one of the generators when we have the ground truth. However, we only have the ground-truth graph set for small values of $|V|$ (recall that the number of non-isomorphic graphs grows very fast with increasing number of vertices). Thus it is important to find a way to evaluate how well a graph generator captures the ground truth, in the absence of ground truth when the number of vertices is large. To illustrate this point, let us appeal to the ground-truth graph set with $|V| = 9$ (the one used in the previous section). We now consider how to get the most "different" types of graphs out of each random generator, i.e., we consider how to obtain the best "coverage" of different values for each graph statistic. We use implementations of all four generators (ER, WS, BA, geometric) from NetworkX [21], with more details provided in the appendix.

**Coverage for Ground-Truth Graph Set:** For each generator, we generate 1%, 0.1% and 0.01% of the total number of graphs in ground-truth graph set (as even these rates are difficult to obtain for large values of $|V|$). We evaluate coverage in two different ways. Intuitively, we can measure the coverage of random generators by plotting each of the graphs in ground-truth graph set as dots in the 2D matrix of correlations and then draw the generated graph set on top of the first plot to see how well the generator set covers the ground-truth graph set. We color the ground-truth graph set in blue and the generated data in red. Because the ground-truth graph set includes all possible graphs for a fixed $|V|$, there is at least one blue point under each red point. Thus, the more blue we see (as in WS and BA) the worse the coverage; see Fig 13-16.

More than just looking at the coverage, we want to test the impact of the increase of $|V|$ under the ER model. Specifically, we consider selecting the edge probability $p$ uniformly at random in the range from 0 to 1 versus selecting values of $p$ that correspond to the edge distribution in the ground truth (which is much more like a normal distribution with a peak at edge density equal to 1/2). Generating graphs with uniform distribution of $p$ is easy, whereas following the population edge distribution is not. We do this by using the population edge distribution as weight and then generating weighted samples between 0 and 1.

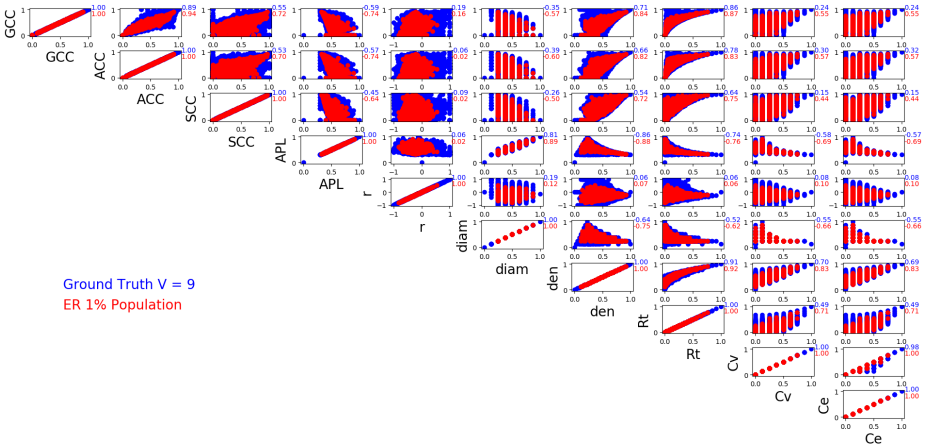


**Fig. 7.** Ground truth (blue) and population edge distribution for ER (red).

From the correlation matrix for ER using the population distribution we can see that correlations are close to the ground truth; see Fig. 7. The correlations are worse when we use uniform distribution, but the coverage of the extreme cases is much better; see Fig. 8. This is especially evident from the columns corresponding to APL, diameter and Rt where the leftmost and/or rightmost points in the plots are blue in the population experiment and red in the uniform.

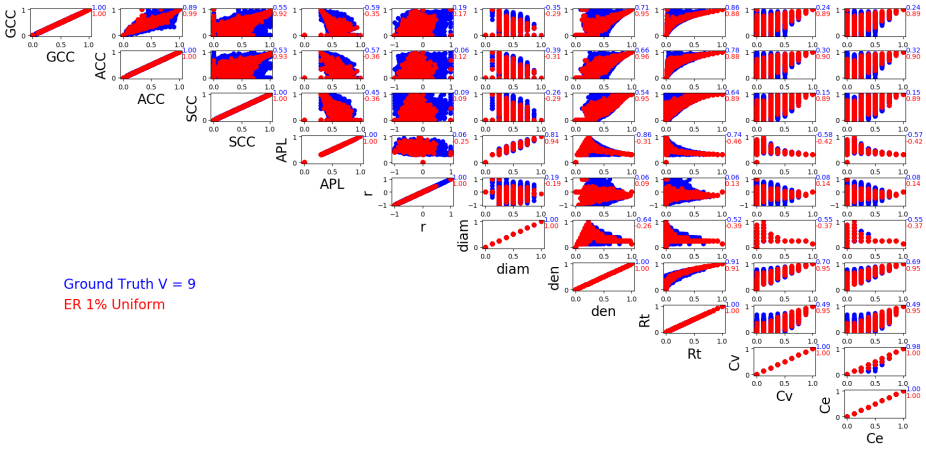


**Fig. 8.** Ground truth (blue) and uniform at random edge distribution for ER (red).

## 6 Finding Different Graphs with Identical Statistics

To find graphs that are identical over a number of graph statistics and yet are different, we use the ground truth data for small non-isomorphic graphs. For larger graphs, we use the graph generators together with some filters.

**Finding graphs in the ground truth:** For $|V| \leq 10$, we directly use all possible non-isomorphic graphs as our dataset. In fact, we can fix different combinations of 5 statistics and still get multiple distinct graphs. We visualize this with figures that encapsulate the variability of one statistic in 10 slots, covering the ranges $[0.0, 0.1], [0.1, 0.2], \ldots [0.9, 1]$ and in each slot we show a graph (if it exists) drawn by a spring layout; see Fig. 9.

For the first experiment, we fix $|V| = 9$, $APL \in (1.42, 1.47)$, $den \in (0.52, 0.57)$, $GCC \in (0.5,0.6)$, $R_t \in (0.15, 0.25)$. Since all our statistics are normalized to $[0, 1]$ and assortativity is in $[-1, 1]$, each of the ten slots has a range of 0.2. We find graphs for seven of the ten possible slots; see Fig. 9. This figure also illustrates the output of our "same stats, different graph" generator: fix several statistics and generate graphs that vary in another statistic.

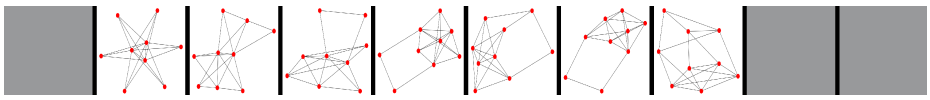

**Fig. 9.** Variability in assortativity.

Similarly, for the second experiment, we fix $|V| = 9$, $APL \in (1.47, 1.69)$, $diam = 3$, $Cv = 2$, $Ce = 2$, and $r \in (-0.22, -0.29)$ to obtain $GCC$ in the range $(0, 0.8)$. The graphs in Figure 9 and Figure 10 are different in structure even though they possess similar values for many properties.

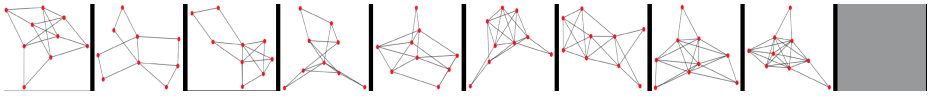

**Fig. 10.** Variability in GCC.

As a final example, we fix $|V| = 9$, $SCC \in (0.75, 0.85)$, $ACC \in (0.75, 0.8)$, $r \in$ (-0.3, -0.2), $R_t \in (0.35, 0.45)$ and find graphs with Ce from 0 to 5.

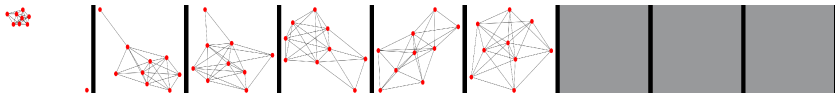

**Fig. 11.** Variability in edge connectivity.

**Finding graphs using graph generators:** This approach relies on generating many graphs and filtering graphs based on several fixed statistics. For the two most important statistics of a graph, $|V|$ and $|E|$, we generate all graphs with same $|V|$ and properly choose $|E|$ in the following two ways:

1. uniform: select $|E|$ uniformly from its range. This is equivalent to forcing the edge density in the generated set to follows a uniform distribution
2. population: select $|E|$ by forcing the edge density in the generated set to match the distribution in the ground truth (population) graph set.

Using both edge selection strategies for all four generators, we compare the statistics distribution to the ground truth for $|V| = 9$. Figure 12 illustrates how different statistics are distributed given uniform edge sampling and population-based edge sampling for the ER model. It shows that although the population-based sampling approach generates more similar distribution to the ground truth, it has a narrower coverage (larger min and smaller max) than the uniform sampling. The WS and BA models do not provide good coverage of the various statistics; see the Appendix for more details.

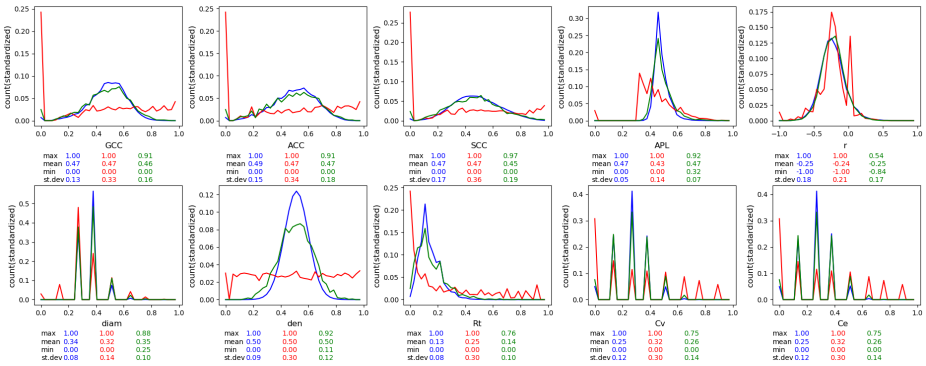


**Fig. 12.** Distribution of the ten statistics, including min/mean/max and standard deviation. Ground truth is in blue, population ER in green, uniform ER in red.

## 7 Discussion and Future Work

We considered how to explore the space of graphs and graph statistics that make it possible to have multiple graphs that are identical over a number of graph statistics, yes are clearly different. To "see" the difference it often suffices to look at the drawings of the graphs. However, as graphs get larger, some graph drawing algorithms may not allow us to distinguish differences in statistics between two graphs, purely from their drawings. We recently studied how the perception statistics such as density and ACC is affected by different graph drawing algorithms [40]. The results confirm the intuition that some drawing algorithms are more appropriate than others in aiding viewers perceive differences between underlying graph statistics. Further work in this direction might help ensure that differences between graphs are captured in the different drawings.

## References


1. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of modern physics 74(1), 47 (2002)
2. Albert, R., Jeong, H., Barabási, A.L.: Internet: Diameter of the world-wide web. nature 401(6749), 130 (1999)
3. Anscombe, F.J.: Graphs in statistical analysis. The American Statistician 27(1), 17–21 (1973), http://www.jstor.org/stable/2682899
4. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science 286(5439), 509–512 (1999)
5. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics reports 424(4-5), 175–308 (2006)
6. Boguná, M., Pastor-Satorras, R.: Epidemic spreading in correlated complex networks. Physical Review E 66(4), 047104 (2002)
7. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. Computer networks 33(1-6), 309–320 (2000)
8. Cartwright, D., Harary, F.: Structural balance: a generalization of heider's theory. Psychological review 63(5), 277 (1956)
9. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM computing surveys (CSUR) 38(1), 2 (2006)
10. Chakrabarti, D., Faloutsos, C.: Graph patterns and the r-mat generator. Mining Graph Data pp. 65–95 (2007)
11. Davis, G.F., Yoo, M., Baker, W.E.: The small world of the american corporate elite, 1982-2001. Strategic organization 1(3), 301–326 (2003)
12. Deng, W., Li, W., Cai, X., Wang, Q.A.: The exponential degree distribution in complex networks: Non-equilibrium network theory, numerical simulation and empirical data. Physica A: Statistical Mechanics and its Applications 390(8), 1481–1485 (2011)
13. Dorogovtsev, S.N., Mendes, J.F.F., Samukhin, A.N.: Structure of growing networks with preferential linking. Physical review letters 85(21), 4633 (2000)
14. Ebel, H., Mielsch, L.I., Bornholdt, S.: Scale-free topology of e-mail networks. Physical review E 66(3), 035103 (2002)
15. Erdös, P., Rényi, A.: On random graphs. Publicationes mathematicae 6, 290–297 (1959)
16. Even, S., Tarjan, R.E.: Network flow and testing graph connectivity. SIAM journal on computing 4(4), 507–518 (1975)
17. Feld, S.L.: The focused organization of social ties. American journal of sociology 86(5), 1015–1035 (1981)
18. Frank, O., Harary, F.: Cluster inference by using transitivity indices in empirical graphs. Journal of the American Statistical Association 77(380), 835–840 (1982)
19. Gilbert, E.N.: Random graphs. The Annals of Mathematical Statistics 30(4), 1141–1144 (1959)
20. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Physical review E 68(6), 065103 (2003)
21. Hagberg, A., Swart, P., S Chult, D.: Exploring network structure, dynamics, and function using networkx. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (2008)
22. Hanneman, R.A., Riddle, M.: Introduction to social network methods (2005)

23. Javarone, M.A.: Gaussian networks generated by random walks. Journal of Statistical Physics 159(1), 108–119 (2015)
24. Kairam, S., MacLean, D., Savva, M., Heer, J.: Graphprism: compact visualization of network structure. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. pp. 498–505. ACM (2012)
25. Karlberg, M.: Testing transitivity in graphs. Social Networks 19(4), 325–343 (1997)
26. Li, G., Semerci, M., Yener, B., Zaki, M.J.: Graph classification via topological and label attributes. In: Proceedings of the 9th international workshop on mining and learning with graphs (MLG), San Diego, USA. vol. 2 (2011)
27. Lind, P.G., Gonzalez, M.C., Herrmann, H.J.: Cycles and clustering in bipartite networks. Physical review E 72(5), 056127 (2005)
28. Loguinov, D., Kumar, A., Rai, V., Ganesh, S.: Graph-theoretic analysis of structured peer-to-peer systems: routing distances and fault resilience. In: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications. pp. 395–406. ACM (2003)
29. Maslov, S., Sneppen, K., Zaliznyak, A.: Detection of topological patterns in complex networks: correlation profile of the internet. Physica A: Statistical Mechanics and its Applications 333, 529–540 (2004)
30. Matejka, J., Fitzmaurice, G.: Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 1290–1294. ACM (2017)
31. McGlohon, M., Akoglu, L., Faloutsos, C.: Statistical properties of social networks. Social network data analytics pp. 17–42 (2011)
32. Melancon, G.: Just how dense are dense graphs in the real world?: a methodological note. In: Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization. pp. 1–7. ACM (2006)
33. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. pp. 29–42. ACM (2007)
34. Newman, M.E.: Assortative mixing in networks. Physical review letters 89(20), 208701 (2002)
35. Newman, M.E.: Mixing patterns in networks. Physical Review E 67(2), 026126 (2003)
36. Newman, M.E.: The structure and function of complex networks. SIAM review 45(2), 167–256 (2003)
37. Newman, M.E., Watts, D.J.: Scaling and percolation in the small-world network model. Physical review E 60(6), 7332 (1999)
38. Price, D.d.S.: A general theory of bibliometric and other cumulative advantage processes. Journal of the Association for Information Science and Technology 27(5), 292–306 (1976)
39. Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P., Mukherjee, G., Manna, S.: Small-world properties of the indian railway network. Physical Review E 67(3), 036106 (2003)
40. Soni, U., Lu, Y., Hansen, B., Purchase, H., Kobourov, S., Maciejewski, R.: The perception of graph properties in graph layouts. In: 20th IEEE Eurographics Conference on Visualization (EuroVis) (2018)
41. Uzzi, B., Spiro, J.: Collaboration and creativity: The small world problem. American journal of sociology 111(2), 447–504 (2005)

42. Van Noort, V., Snel, B., Huynen, M.A.: The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO reports 5(3), 280–284 (2004)
43. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. nature 393(6684), 440 (1998)
44. Wei-Bing, D., Long, G., Wei, L., Xu, C.: Worldwide marine transportation network: Efficiency and container throughput. Chinese Physics Letters 26(11), 118901 (2009)

# Appendix

## Generator Parameters

The ER random graph generator requires 2 parameters: the number of vertices $n$ and the probability for adding any edge, $p$. We select $n = 9$ and a random value from 0 to 1 for $p$.

The WS small-world graph generator requires 3 parameters:, the number of vertices $n$, a value $k$, where each node is joined with its $k$ nearest neighbors in a ring topology and the probability for adding any edge, $p$. We select $n = 9$, and vary $k$ from 2 to $|V| - 1$ and d a random value from 0 to 1 for $p$.

The BA preferential attachment model requires 2 parameters: the number of vertices $n$ and $m$ edges to attach from a new node to existing nodes. We select $n = 9$, and randomly choose an integer value from 1 to $|V| - 1$ for $m$.

The geometric graph generator requires 2 parameters: the number of vertices $n$ and a *radius* threshold value (if two vertices are closer than the radius an edge is placed between them and otherwise an edge is not in the graph). We select $n = 9$ and a random value from 0 to 1 for the radius.

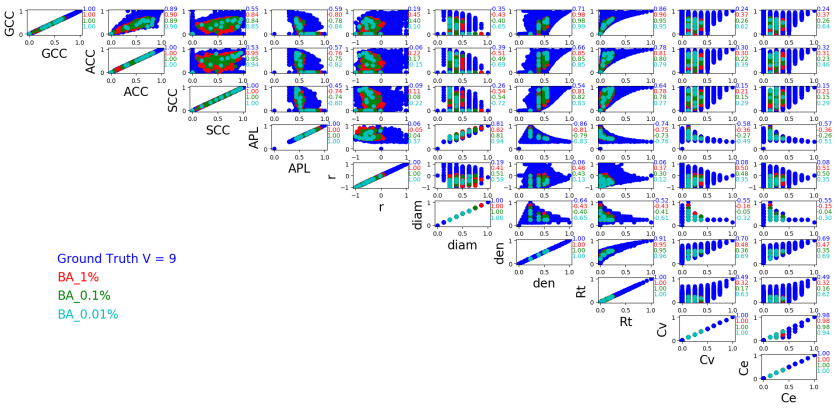


**Fig. 13.** BA model for 0.01%, 0.1% and 1% with ground truth in the back, fixed $|v| = 9$

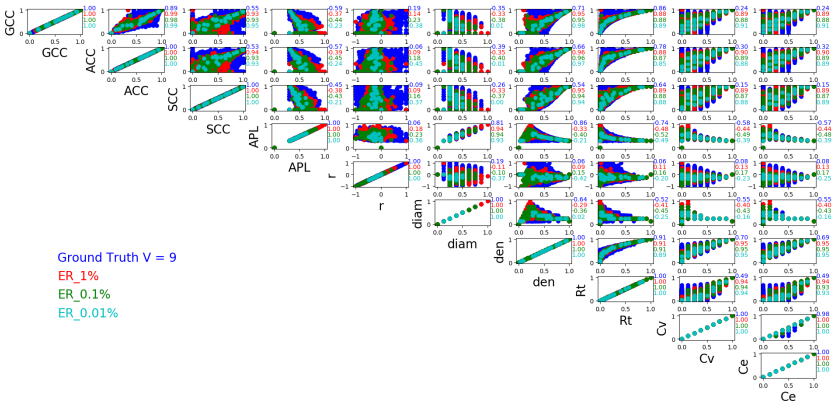


**Fig. 14.** ER model for 0.01%, 0.1% and 1% with ground truth in the back, fixed $|v| = 9$

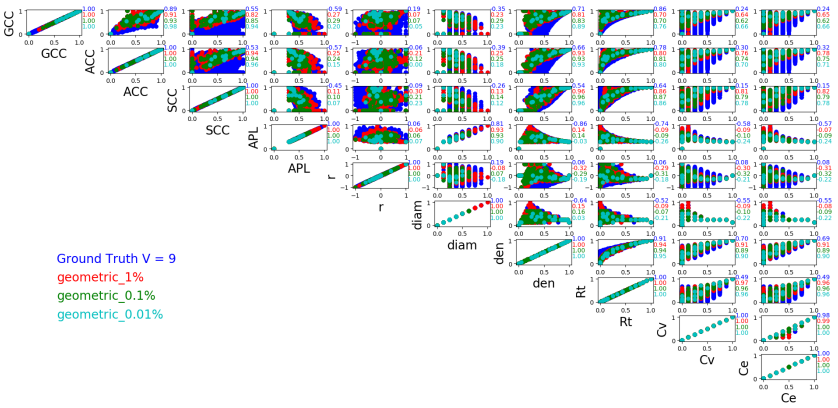


**Fig. 15.** geometric model for 0.01%, 0.1% and 1% with ground truth in the back, fixed $|v| = 9$

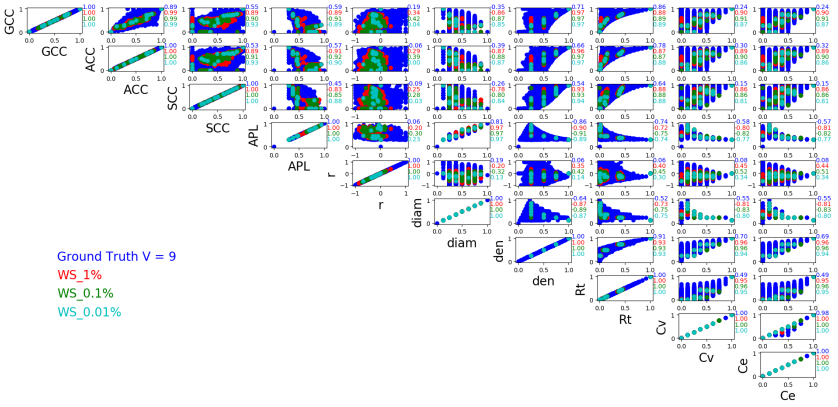


**Fig. 16.** WS model for 0.01%, 0.1% and 1% with ground truth in the back, fixed $|v| = 9$