Same Stats, Different Graphs

(Graph Statistics and Why We Need Graph Drawings)

University of Arizona and Arizona State University
Motivation

Imagine a set of 11 points in 2D with the following summary statistics:

- Mean value of $x = 9$
- Variance of $x = 11$
- Mean value of $y = 7.5$
- Variance of $y = 4.1$
- Correlation between $x$ and $y = 0.8$
- Linear regression line $y = 3 + 0.5x$
Motivation

mean $x = 9$
var($x$) = 11
mean $y = 7.5$
var($y$) = 4.1
corr($x,y$) = 0.8
reg: $y = 3 + 0.5x$

Same Stats, Different Graphs
Motivation

\[
\begin{align*}
\text{mean } x &= 9 \\
\text{var}(x) &= 11 \\
\text{mean } y &= 7.5 \\
\text{var}(y) &= 4.1 \\
\text{corr}(x, y) &= 0.8 \\
\text{reg: } y &= 3 + 0.5x
\end{align*}
\]
Motivation

mean $x = 9$
var($x$) = 11
mean $y = 7.5$
var($y$) = 4.1
corr($x,y$) = 0.8
reg: $y = 3 + 0.5x$

Same Stats, Different Graphs
Motivation

Same Stats, Different Graphs

$\text{mean } x = 9$

$\text{var}(x) = 11$

$\text{mean } y = 7.5$

$\text{var}(y) = 4.1$

$\text{corr}(x, y) = 0.8$

reg: $y = 3 + 0.5x$
Anscombe’s Quartet

Anscombe, 1973

mean $x = 9$
var($x$) = 11
mean $y = 7.5$
var($y$) = 4.1
corr($x,y$) = 0.8
reg: $y = 3 + 0.5x$
Moral of the story: Summary statistics of a dataset are great, but we should nevertheless look at the data!
Moral of the story: Summary statistics of a dataset are great, but we should nevertheless look at the data!

Or in fortune cookie language

I hear and I forget. I see and I remember.
Imagine a graph with the following properties (statistics):

- 12 vertices
- 21 edges
- girth $\gamma = 3$
- number of triangles $\Delta = 10$
- global clustering coefficient $= 0.5$
Stephen’s Quartet

|V| = 12
|E| = 21
γ = 3
Δ = 10
GCC = 0.5
Stephen’s Quartet

|V| = 12
|E| = 21
\(\gamma\) = 3
\(\Delta\) = 10
GCC = 0.5

Same Stats, Different Graphs
Stephen’s Quartet

\[ |V| = 12 \]
\[ |E| = 21 \]
\[ \gamma = 3 \]
\[ \Delta = 10 \]

GCC = 0.5
Stephen’s Quartet

\[ |V| = 12 \]
\[ |E| = 21 \]
\[ \gamma = 3 \]
\[ \Delta = 10 \]
\[ GCC = 0.5 \]
These four graphs have the same 5 statistics but they differ in structure, planarity, connectivity, symmetry, etc.
Stephen’s Quartet

Moral of the story: every graph drawing paper could begin with these 4 graphs as the motivation behind “Why We Still Need to Draw our Graphs”

Same Stats, Different Graphs
So, can we modify a given graph and preserve a given set of summary statistics while significantly changing other graph properties and statistics?
Question

So, can we modify a given graph and preserve a given set of summary statistics while significantly changing other graph properties and statistics?

This is significantly harder to do with graphs than with the 2D pointsets in Anscombe’s quartet, as some graph properties are correlated…
So, can we modify a given graph and preserve a given set of summary statistics while significantly changing other graph properties and statistics?

This is significantly harder to do with graphs than with the 2D pointsets in Anscombe’s quartet, as some graph properties are correlated...

Why? graph anonymization, to measure graph property perception in layouts, ...
So, can we modify a given graph and preserve a given set of summary statistics while significantly changing other graph properties and statistics?

This is significantly harder to do with graphs than with the 2D pointsets in Anscombe’s quartet, as some graph properties are correlated…

Why? graph anonymization, to measure graph property perception in layouts, …
Graph Properties Considered

- Normalized to $[0,1]$
- Assortativity: $[-1,1]$

- **Average Clustering Coefficient**
  \[
  ACC(G) = \frac{1}{n} \sum_{i=1}^{n} c(u_i), u_i \in V, n = |V|
  \]
  \[
  c(v) = \frac{|\{(u,w)|u,w \in \Gamma(v), (u,w) \in E\}|}{|\Gamma(v)|(|\Gamma(v)|-1)/2}, v, u, w \in V
  \]

- **Global Clustering Coefficient**
  \[
  GCC(G) = \frac{3 \times |\text{triangles}|}{|\text{connected triples}| \text{ in the graph}}
  \]

- **Square Clustering**
  \[
  SCC(G) = \frac{\sum_{u=1}^{k_v} \sum_{w=u+1}^{k_v} q_v(u,w)}{\sum_{u=1}^{k_v} \sum_{w=u+1}^{k_v} [a_v(u,w)+q_v(u,w)]}
  \]

- **Average Path Length**
  \[
  APL = \text{ave}\left\{ \frac{n-1}{\sum_{v \in V} d(u,v), u \neq v} \right\}
  \]

- **Degree Assortativity**
  \[
  r = \frac{\sum_{x,y} xy(e_{xy}-a_x b_y)}{\sigma_a \sigma_b}
  \]

- **Diameter**
  \[
  \text{diam}(G) = \max\{\text{dist}(v, w), v, w \in V\}
  \]

- **Density**
  \[
  \text{den} = \frac{2|E|}{|V|(|V|-1)}
  \]

- **Ratio of Triangles**
  \[
  \text{Rt} = \frac{|\text{triangles}|}{|V|(|V|-1)/2}
  \]

- **Node Connectivity**
  Cv: the minimum number of nodes to remove to disconnect the graph

- **Edge Connectivity**
  Ce: the minimum number of edges to remove to disconnect the graph
Correlations between Graph Properties

* Data from EuroVis’18 paper where we generated 4950 graphs with 100 vertices
Correlations between Graph Properties

Can we trust the numbers from the previous table?
Correlations between Graph Properties

Can we trust the numbers from the previous table?

Let’s look at the set of all (non-isomorphic) graphs on 100 vertices and compute the correlations again.
Correlations between Graph Properties

Can we trust the numbers from the previous table?

Let’s look at the set of all (non-isomorphic) graphs on 100 vertices and compute the correlations again.

Statistical Diagram: Diagram illustrating various graph structures.
Correlations between Graph Properties

Can we trust the numbers from the previous table?

Let’s look at the set of all (non-isomorphic) graphs on 100 vertices and compute the correlations again.

Good idea, but the number of (non-isomorphic) graphs grows very quickly:

Same Stats, Different Graphs
Can we trust the numbers from the previous table? Let's look at the set of all (non-isomorphic) graphs on 100 vertices and compute the correlations again.

Good idea, but the number of (non-isomorphic) graphs grows very quickly:

For $|V| = 1, 2 \ldots 9$ the numbers are 1, 2, 4, 11, 34, 156, 1044, 12346, 274668, and for $|V| = 16$ we have $6 \times 10^{22}$
Correlations between Graph Properties

<table>
<thead>
<tr>
<th></th>
<th>GCC</th>
<th>ACC</th>
<th>SCC</th>
<th>APL</th>
<th>diam</th>
<th>den</th>
<th>Rt</th>
<th>Cv</th>
<th>Ce</th>
<th>GCC</th>
<th>ACC</th>
<th>SCC</th>
<th>APL</th>
<th>diam</th>
<th>den</th>
<th>Rt</th>
<th>Cv</th>
<th>Ce</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCC</td>
<td>1.00</td>
<td>0.86</td>
<td>0.20</td>
<td>-0.80</td>
<td>-0.61</td>
<td>0.92</td>
<td>0.92</td>
<td>0.99</td>
<td>0.99</td>
<td>0.89</td>
<td>0.55</td>
<td>-0.59</td>
<td>-0.19</td>
<td>-0.35</td>
<td>0.71</td>
<td>0.86</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>ACC</td>
<td>1.00</td>
<td>0.86</td>
<td>0.10</td>
<td>-0.80</td>
<td>-0.61</td>
<td>0.19</td>
<td>0.92</td>
<td>0.99</td>
<td>0.99</td>
<td>0.89</td>
<td>0.53</td>
<td>-0.57</td>
<td>-0.06</td>
<td>-0.39</td>
<td>0.66</td>
<td>0.78</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td>SCC</td>
<td>0.20</td>
<td>0.86</td>
<td>0.10</td>
<td>-0.32</td>
<td>0.13</td>
<td>-0.14</td>
<td>-0.05</td>
<td>-0.13</td>
<td>-0.12</td>
<td>0.55</td>
<td>0.53</td>
<td>-0.45</td>
<td>-0.09</td>
<td>-0.26</td>
<td>0.54</td>
<td>0.64</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>APL</td>
<td>-0.80</td>
<td>-0.80</td>
<td>0.10</td>
<td>0.70</td>
<td>0.94</td>
<td>-0.43</td>
<td>-0.64</td>
<td>-0.76</td>
<td>-0.76</td>
<td>-0.59</td>
<td>-0.57</td>
<td>-0.45</td>
<td>0.06</td>
<td>0.19</td>
<td>0.06</td>
<td>0.06</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>diam</td>
<td>-0.61</td>
<td>-0.61</td>
<td>-0.13</td>
<td>0.94</td>
<td>0.71</td>
<td>-0.35</td>
<td>-0.38</td>
<td>-0.54</td>
<td>-0.54</td>
<td>-0.35</td>
<td>-0.39</td>
<td>-0.26</td>
<td>0.81</td>
<td>0.19</td>
<td>-0.64</td>
<td>-0.52</td>
<td>-0.55</td>
<td>-0.55</td>
</tr>
<tr>
<td>den</td>
<td>0.92</td>
<td>0.19</td>
<td>-0.14</td>
<td>-0.43</td>
<td>0.65</td>
<td>-0.35</td>
<td>0.93</td>
<td>0.99</td>
<td>0.99</td>
<td>0.71</td>
<td>0.66</td>
<td>-0.86</td>
<td>0.06</td>
<td>0.64</td>
<td>0.91</td>
<td>0.70</td>
<td>0.69</td>
<td>-0.49</td>
</tr>
<tr>
<td>Rt</td>
<td>0.92</td>
<td>0.92</td>
<td>-0.05</td>
<td>-0.64</td>
<td>-0.29</td>
<td>-0.38</td>
<td>0.93</td>
<td>0.95</td>
<td>0.95</td>
<td>0.86</td>
<td>0.78</td>
<td>-0.74</td>
<td>-0.52</td>
<td>0.91</td>
<td>0.49</td>
<td>0.49</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Cv</td>
<td>0.99</td>
<td>0.99</td>
<td>-0.13</td>
<td>-0.76</td>
<td>-0.37</td>
<td>-0.54</td>
<td>0.99</td>
<td>0.95</td>
<td>1.00</td>
<td>0.24</td>
<td>0.30</td>
<td>0.15</td>
<td>-0.58</td>
<td>0.08</td>
<td>0.55</td>
<td>0.69</td>
<td>0.49</td>
<td>0.98</td>
</tr>
<tr>
<td>Ce</td>
<td>0.99</td>
<td>0.99</td>
<td>-0.12</td>
<td>-0.76</td>
<td>-0.37</td>
<td>-0.54</td>
<td>0.99</td>
<td>0.95</td>
<td>1.00</td>
<td>0.24</td>
<td>0.32</td>
<td>0.15</td>
<td>-0.57</td>
<td>0.08</td>
<td>-0.55</td>
<td>0.69</td>
<td>0.49</td>
<td>0.98</td>
</tr>
</tbody>
</table>

4950 graphs with 100 vertices from EuroVis’17 experiment

ground truth for |V|=9 and the results are different...
Ground Truth Data for Small $|V|$
Ground Truth Data for Small $|V|$
Ground Truth Data for Small $|V|$ 

Let's look at one of these more carefully

As $|V|$ grows the correlation changes!

Ground Truth $V = 10$
Ground Truth $V = 9$
Ground Truth $V = 8$
Ground Truth $V = 7$
Ground Truth $V = 6$
Ground Truth $V = 5$

Same Stats, Different Graphs
Graph Generators to the Rescue

We cannot explore the ground data for large values of $|V|$, so let’s use generators

- Erdos-Renyi
- Watts-Strogatz
- Barabasi-Albert
- geometric

But which generator does a **good job** in this context?

What do we want from a graph generator?
Desirable Generator Properties

Does the graph generator:

- **represent** the ground truth data well, i.e., does the generator yield a sample that with similar properties as those in the ground truth?
- **cover** the complete range of values for the properties found in the ground truth data?
Graph Generator Representativeness

We measure how **representative** a graph generator is by comparing pairwise correlations in the sample and in the ground truth.

Ground Truth $V = 9$
ER 1% $p = 0.5$
Graph Generator Representativeness

We measure how **representative** a graph generator is by comparing pairwise correlations in the sample and in the ground truth.

Ground Truth $V = 9$
ER 1% $p = 0.5$
Graph Generator Representativeness

We measure how representative a graph generator is by comparing pairwise correlations in the sample and in the ground truth.

Ground Truth $V = 9$
ER 1% $p = 0.5$

Same Stats, Different Graphs
Graph Generator Coverage

We measure how well a graph generator **covers** the range of values in the ground truth data by comparing the volumes of the generated data and the ground truth.

For example, we can compare the ratios of the 10D bounding boxes of the two datasets (generator, ground truth).

<table>
<thead>
<tr>
<th>WS model</th>
<th>BA model</th>
<th>ER model $p = 0.5$</th>
<th>ER model $p^\sim$ Uniform</th>
<th>ER model $p^\sim$ Population</th>
<th>Geometric model</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.04%</td>
<td>0.10%</td>
<td>0.98%</td>
<td>90.83%</td>
<td>12.37%</td>
<td>83.87%</td>
</tr>
</tbody>
</table>

**Same Stats, Different Graphs**
Graph Generator Performance?

No graph generator is good at representing the ground truth and covering it well. Why?

It seems the answer is that all generators sample the space of isomorphic graphs whereas we are considering the space of non-isomorphic graphs.
We can generate graphs with fixed set of statistics that vary in another statistic:

\[ |V| = 9, \text{SCC} \in (0.75, 0.85), \text{ACC} \in (0.75, 0.8), r \in (-0.3, -0.2), Rt \in (0.35, 0.45) \]

and we can find graphs in 6 out of 8 buckets for connectivity

for small sizes we can simply look in the ground truth data

for large sizes we must use generators
Related Work

1. F. Anscombe, Graphs in statistical analysis, The American Statistician 27(1), 1973
2. J. Matejka and G. Fitzmaurice, Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing, CHI, 2017
3. B. Bach, A. Spritzer, E. Lutton, J. D. Fekete, Interactive random graph generation with evolutionary algorithms, GD, 2012

Same Stats, Different Graphs
Open Problems

1. Some drawing algorithms may not allow us to see differences in statistics between two graphs purely from their drawings; how can we address this?
2. Efficiently generate graphs of the “same stats, different graphs” type?
3. What are the correlations between different graph properties/statistics?
4. Generator that represents and covers the space of non-isomorphic graphs?

ACC=0.1  ACC=0.3  ACC=0.7  ACC=0.9

Circular

MDS

Same Stats, Different Graphs