Same Stats, Different Graphs

(Graph Statistics and Why We Need Graph Drawings)

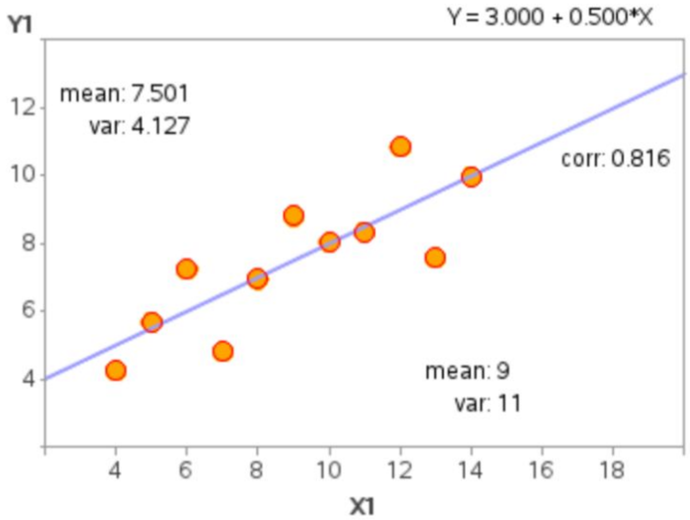
H. Chen, U. Soni, Y. Lu, R. Maciejewski, **S. Kobourov**
University of Arizona and Arizona State University

Motivation

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

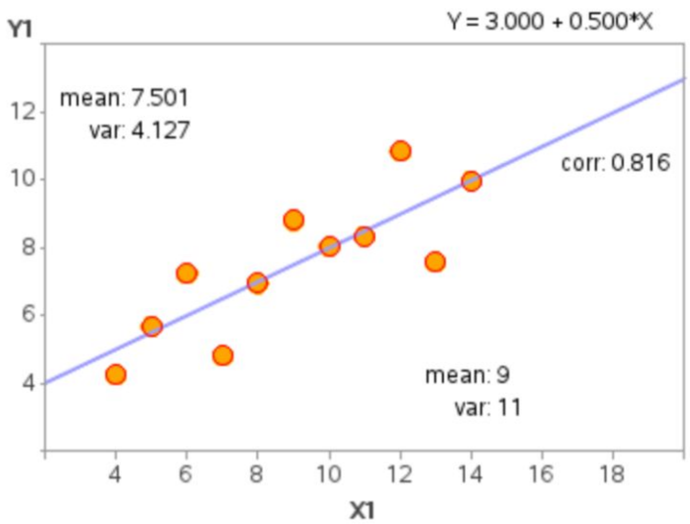
- Mean value of $x = 9$
- Variance of $x = 11$
- Mean value of $y = 7.5$
- Variance of $y = 4.1$
- Correlation between x and $y = 0.8$
- Linear regression line $y = 3 + 0.5x$

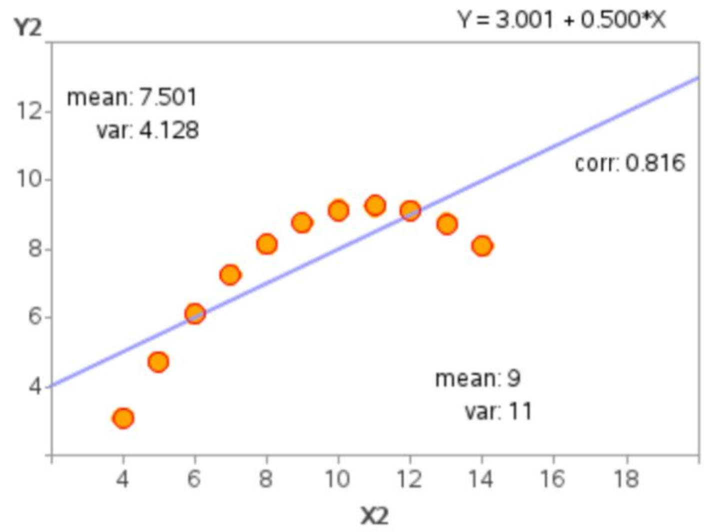
Motivation



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

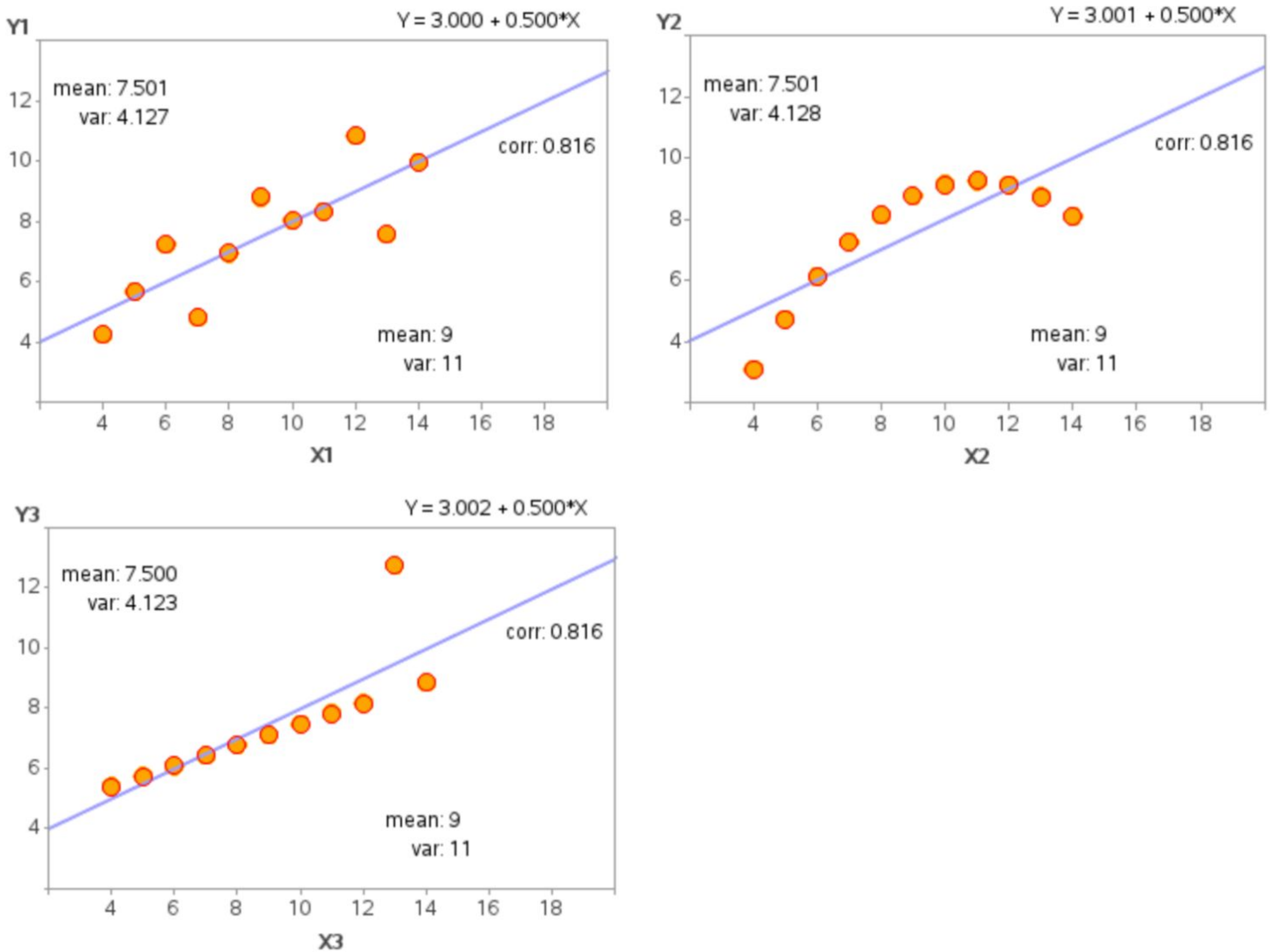
Motivation





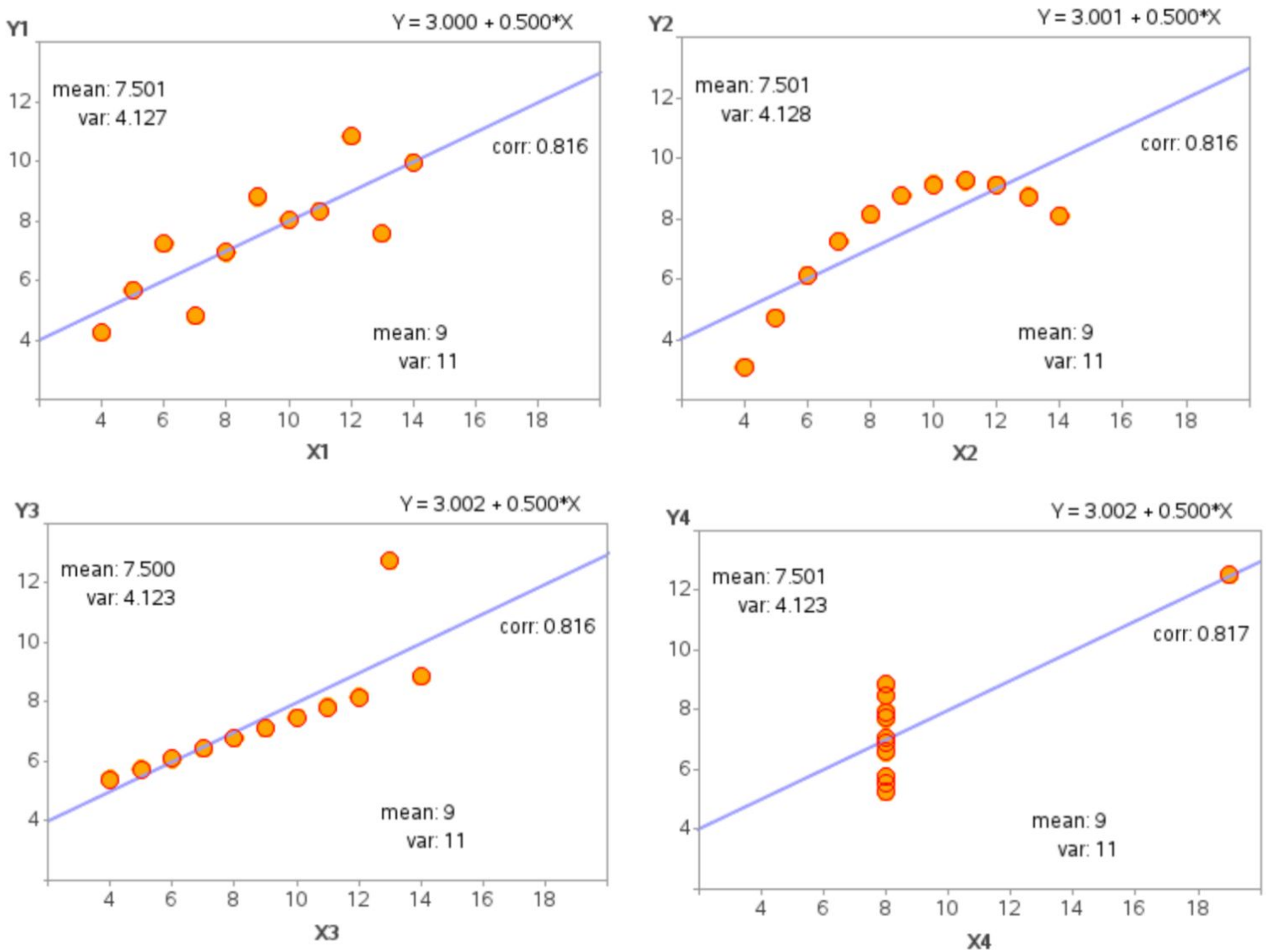
mean x = *9*
var(x) = *11*
mean y = *7.5*
var(y) = *4.1*
corr(x,y) = *0.8*
reg: $y=3+0.5x$

Motivation



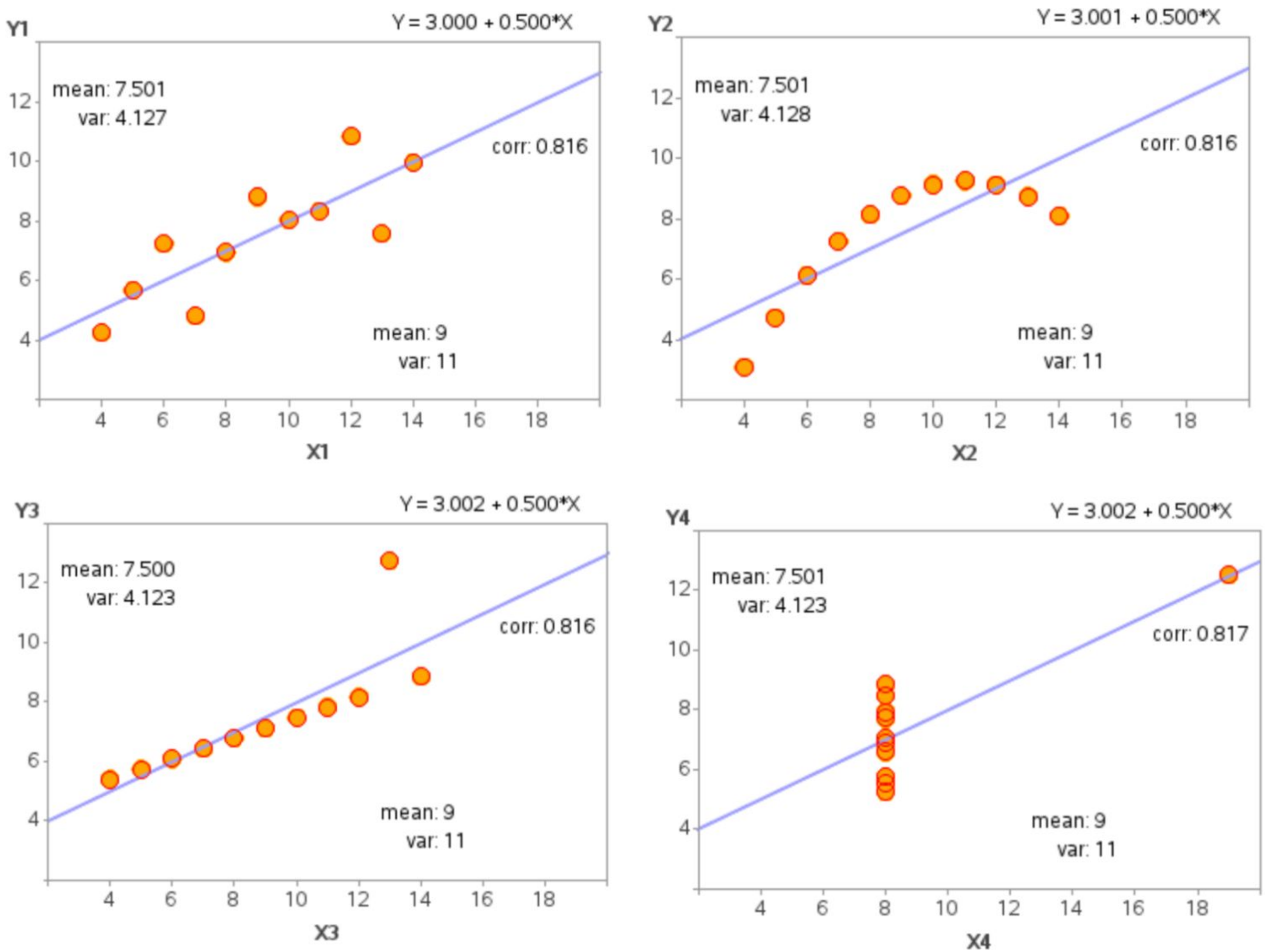
mean x = *9*
var(x) = *11*
mean y = *7.5*
var(y) = *4.1*
corr(x,y) = *0.8*
reg: $y=3+0.5x$

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*

Anscombe's Quartet



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

Anscombe, 1973

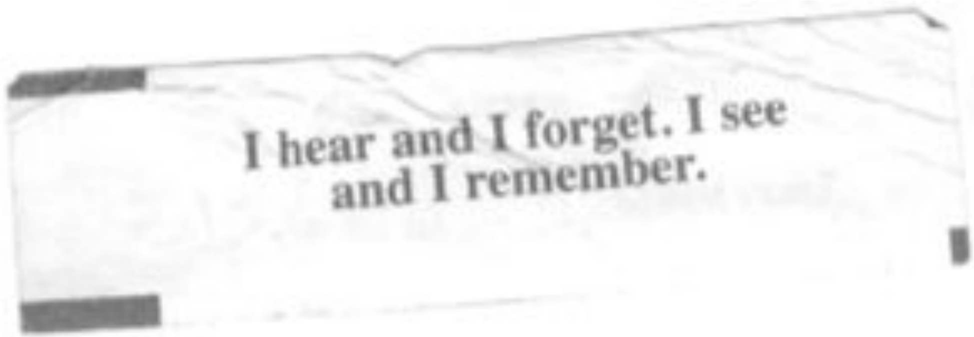
Anscombe's Quartet

Moral of the story: Summary statistics of a dataset are great, but we should nevertheless look at the data!

Anscombe's Quartet

Moral of the story: Summary statistics of a dataset are great, but we should nevertheless look at the data!

Or in fortune cookie language

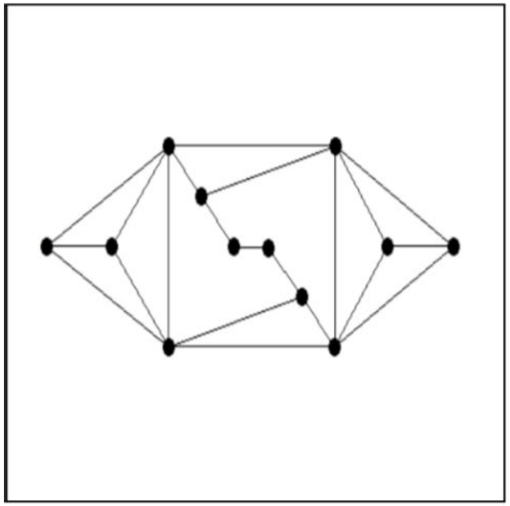


Stephen’s Quartet

Imagine a graph with the following properties (statistics):

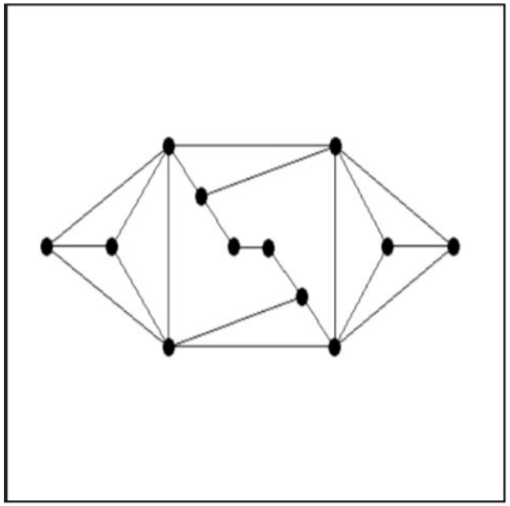
- 12 vertices
- 21 edges
- girth $\gamma = 3$
- number of triangles $\Delta = 10$
- global clustering coefficient = 0.5

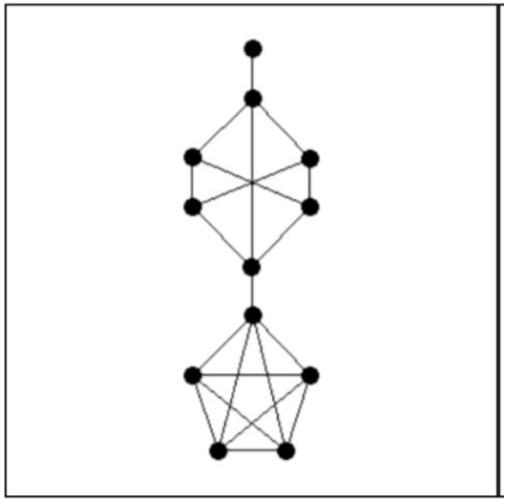
Stephen's Quartet



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

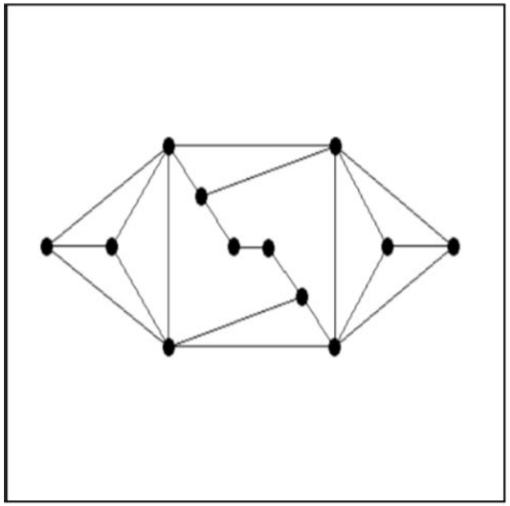
Stephen's Quartet

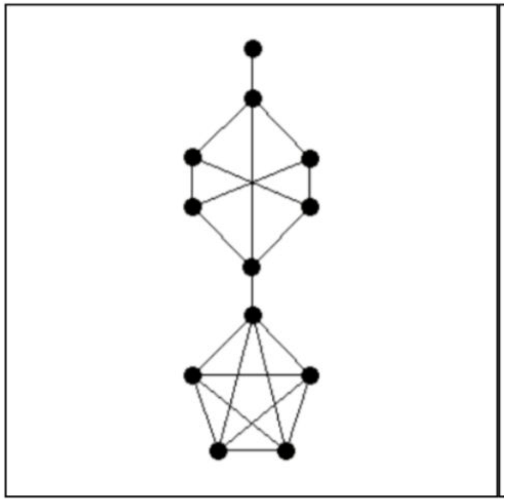


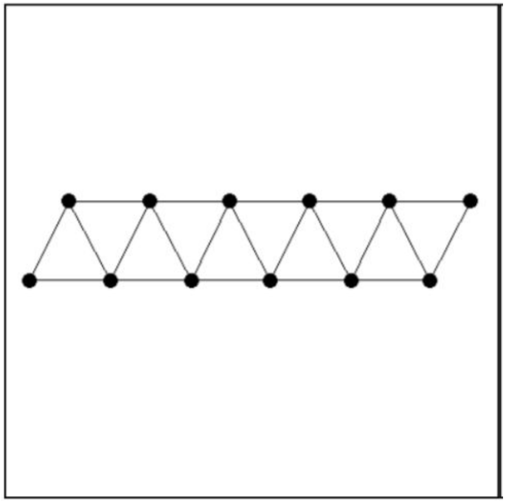


$|V| = 12$
$|E| = 21$
$\gamma = 3$
$\Delta = 10$
$GCC = 0.5$

Stephen’s Quartet

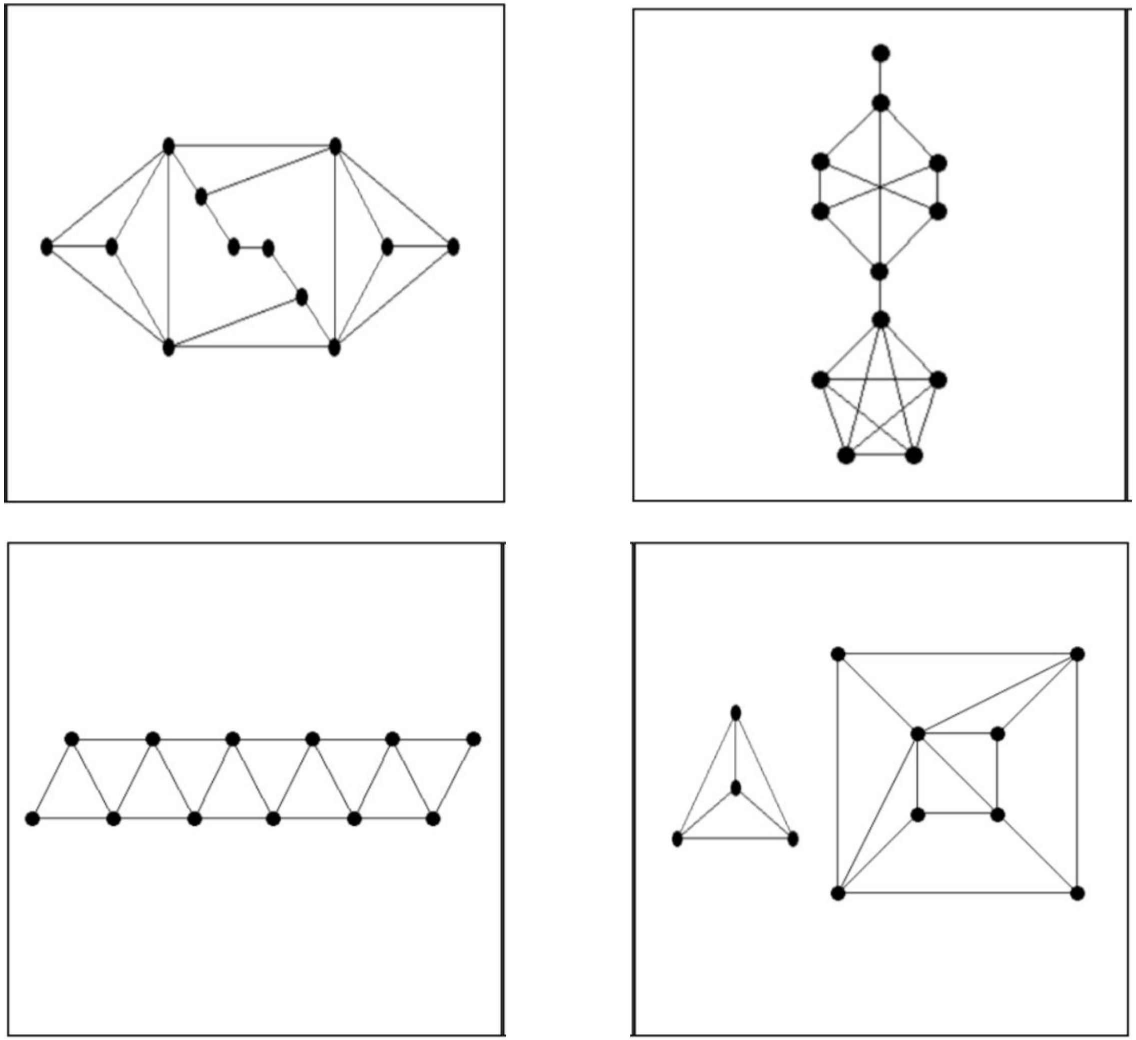






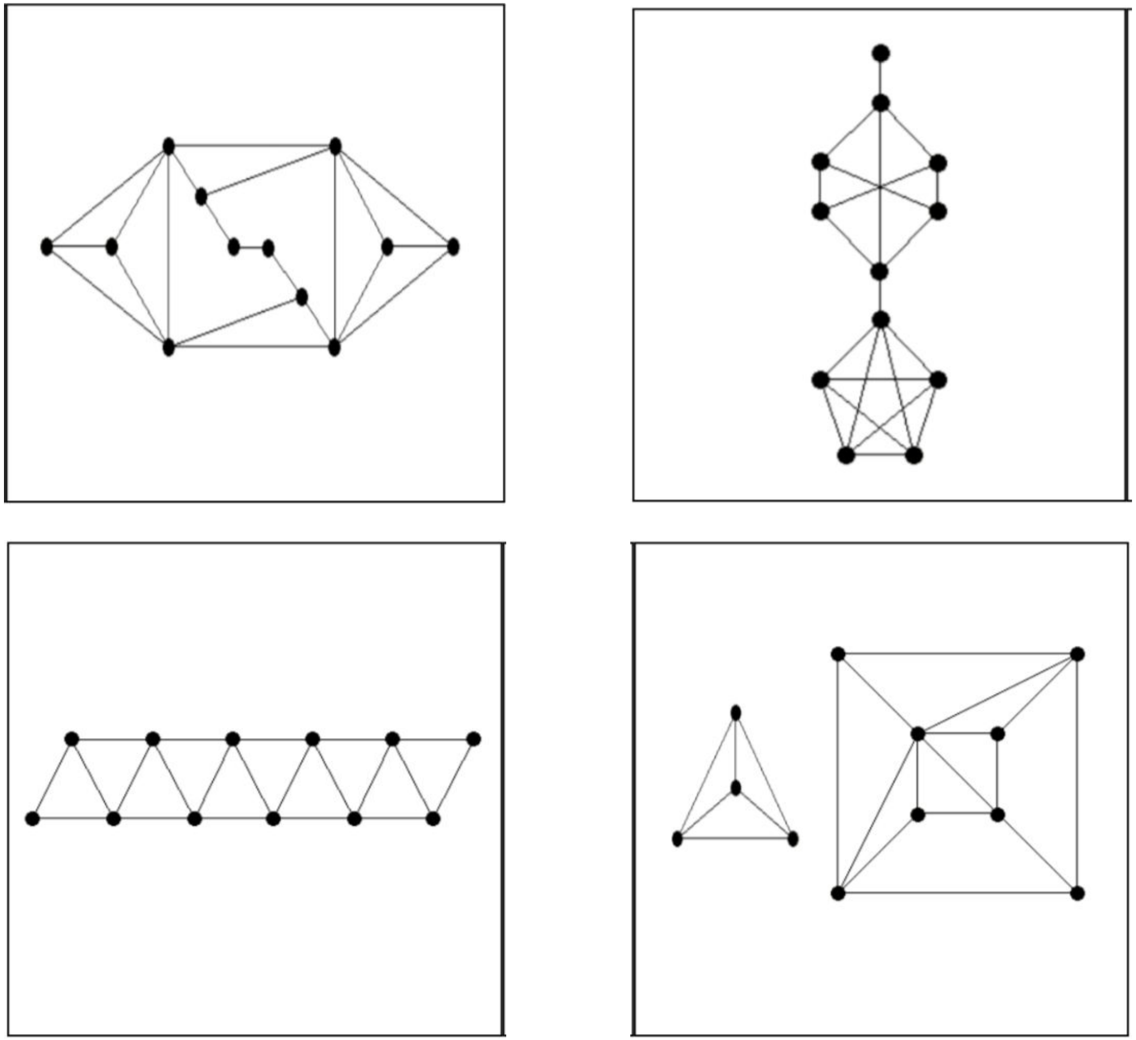
$|V| = 12$
$|E| = 21$
$\gamma = 3$
$\Delta = 10$
$GCC = 0.5$

Stephen's Quartet



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

Stephen's Quartet

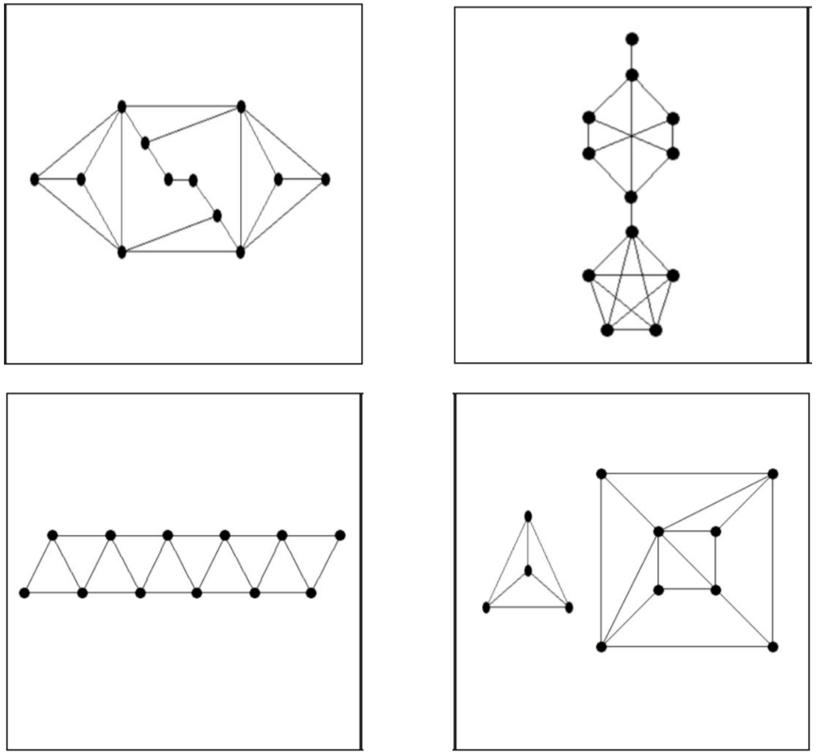


$|V| = 12$
$|E| = 21$
$\gamma = 3$
$\Delta = 10$
$GCC = 0.5$

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"



Question

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…

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…

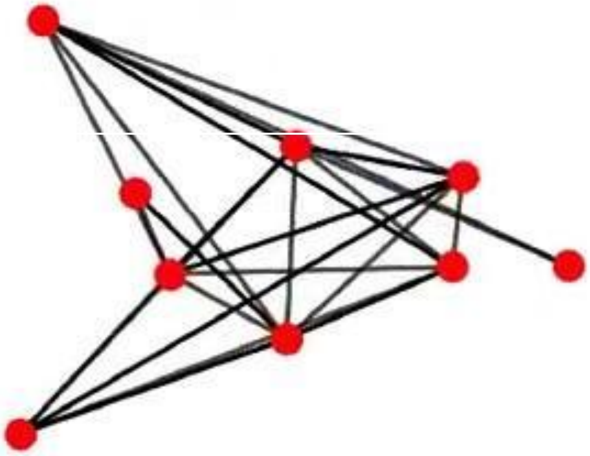
Why? graph anonymization, to measure graph property perception in layouts, ...

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…

Why? graph anonymization, to measure graph property perception in layouts, ...



Graph Properties Considered

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	$GCC(G) = \frac{3 \times \|triangles\|}{\|connected\ triples\|\ in\ the\ graph}$
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	$APL = ave\{\frac{n-1}{\sum_{v \in V} d(u,v), u \neq v}\}$
Degree Assortativity	$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}$
Diameter	$diam(G) = max\{dist(v,w), v, w \in V\}$
Density	$den = \frac{2\|E\|}{\|V\|(\|V\|-1)}$
Ratio of Triangles	$\mathrm{Rt} = \frac{\|triangles\|}{\|V\|(\|V\|-1)/2}$
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

- normalized to *[0,1]*
- assortativity: *[-1,1]*

Correlations between Graph Properties

	GCC	ACC	SCC	APL	r	diam	den	Rt	Cv	Ce
GCC		1.00	0.86	-0.80	-0.41	-0.61	0.92	0.92	0.99	0.99
ACC	1.00		0.86	-0.80	-0.41	-0.61	0.19	0.92	0.99	0.99
SCC	0.20	0.86		0.10	-0.32	0.13	-0.14	-0.05	-0.13	-0.12
APL	-0.80	-0.80	0.10		0.70	0.94	-0.43	-0.64	-0.76	-0.76
r	-0.41	-0.41	-0.32	0.70		0.71	0.65	-0.29	-0.37	-0.37
diam	-0.61	-0.61	0.13	0.94	0.71		-0.35	-0.38	-0.54	-0.54
den	0.92	0.19	-0.14	-0.43	0.65	-0.35		0.93	0.99	0.99
Rt	0.92	0.92	-0.05	-0.64	-0.29	-0.38	0.93		0.95	0.95
Cv	0.99	0.99	-0.13	-0.76	-0.37	-0.54	0.99	0.95		1.00
Ce	0.99	0.99	-0.12	-0.76	-0.37	-0.54	0.99	0.95	1.00	

Colour scale: 0.9, 0.6, 0.3, 0.0, −0.3, −0.6

* 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

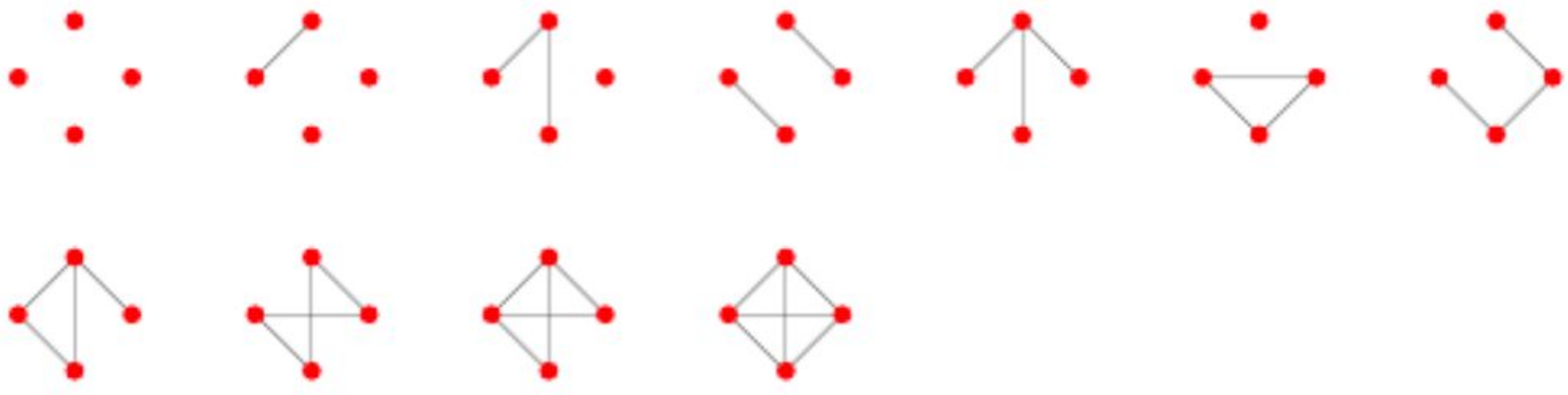
Can we trust the numbers from the previous table?

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

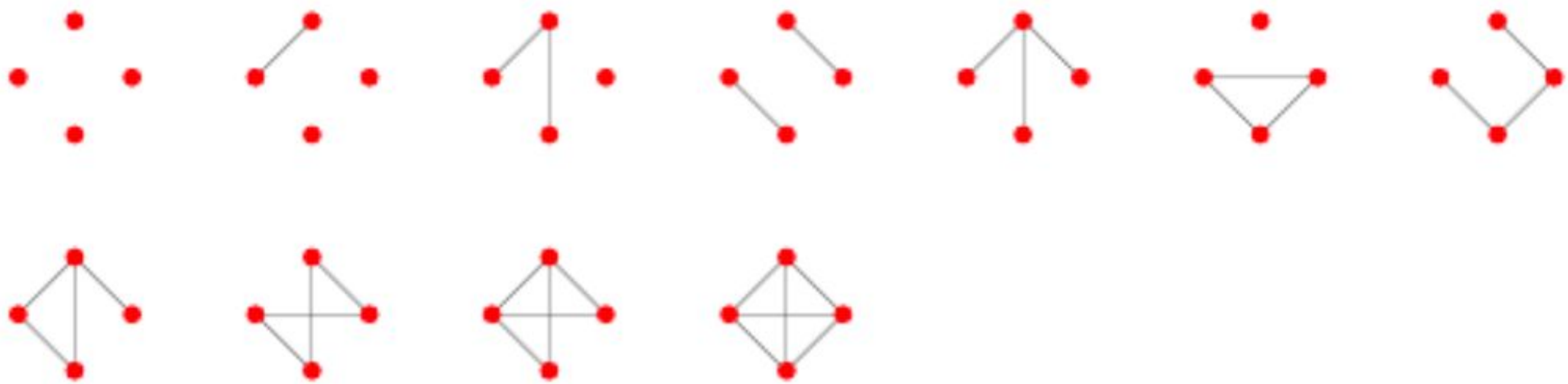


Correlations between Graph Properties

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:



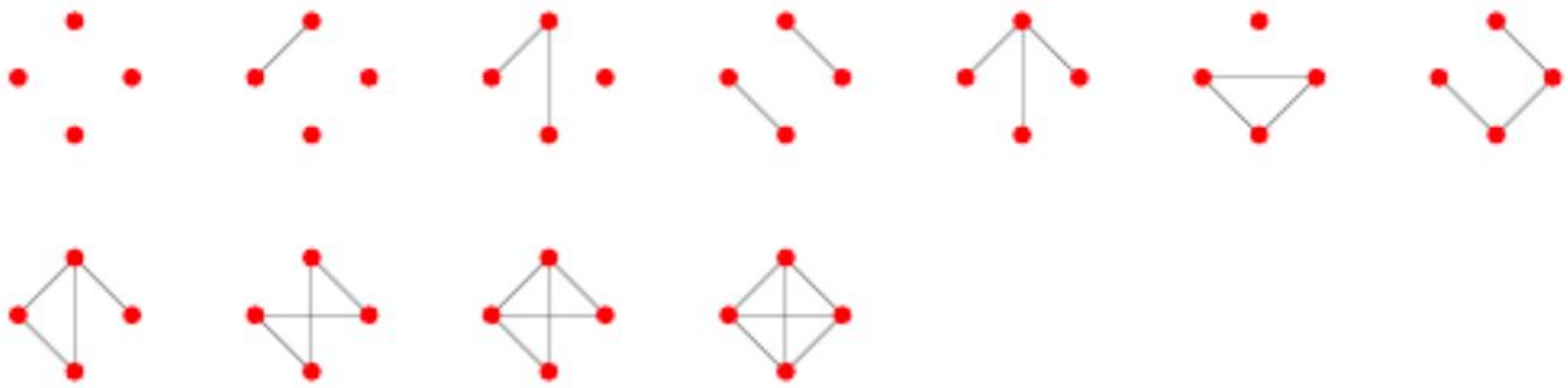
Correlations between Graph Properties

Can we trust the numbers from the previous table?

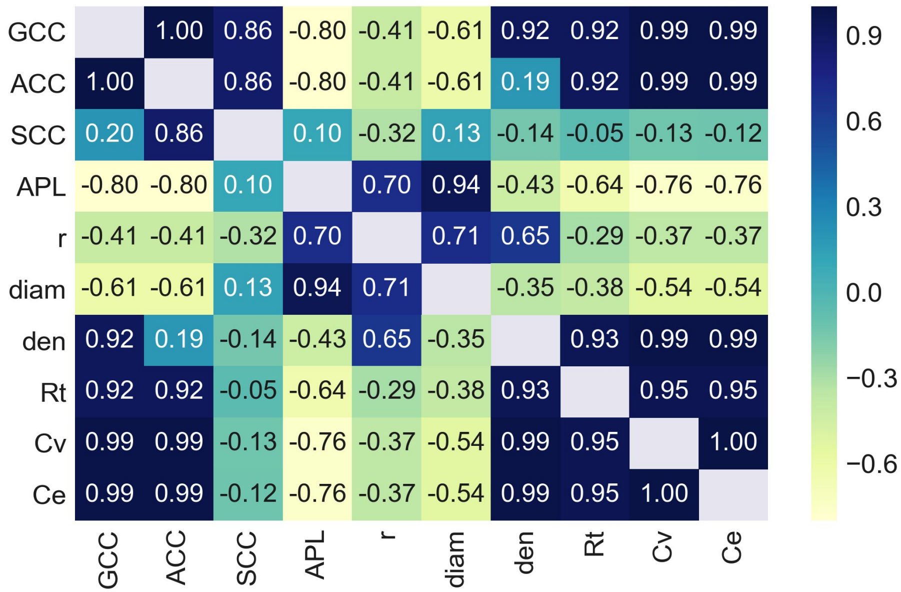
Let's look at the set of all (non-isomorphic) graphs on 100 vertices and compute the correlations again

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

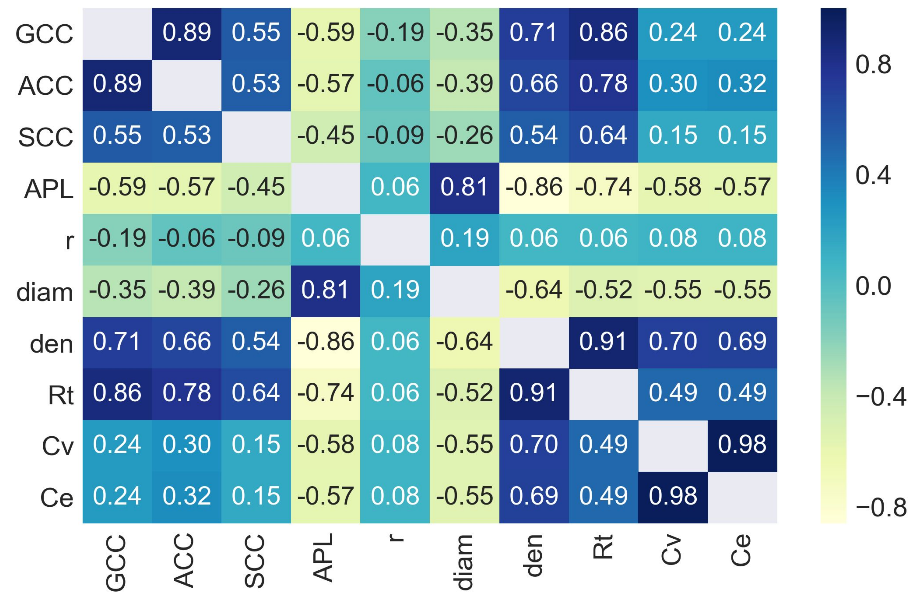
For $|V| = 1, 2 \ldots 9$ the numbers are 1, 2, 4, 11, 34, 156, 1044, 12346, 274668, and for $|V| = 16$ we have 6×10^{22}



Correlations between Graph Properties

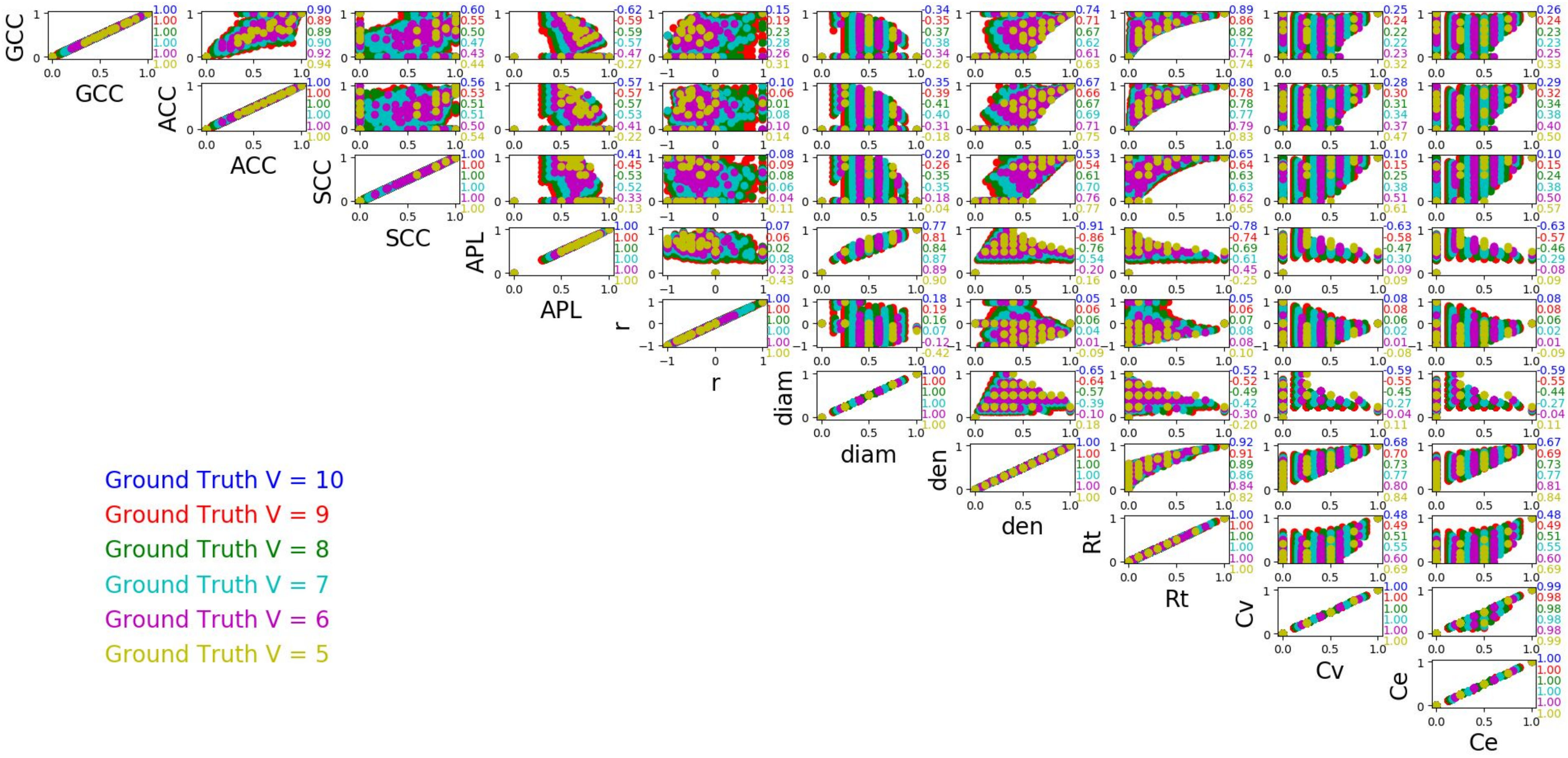


4950 graphs with 100 vertices from EuroVis'17 experiment

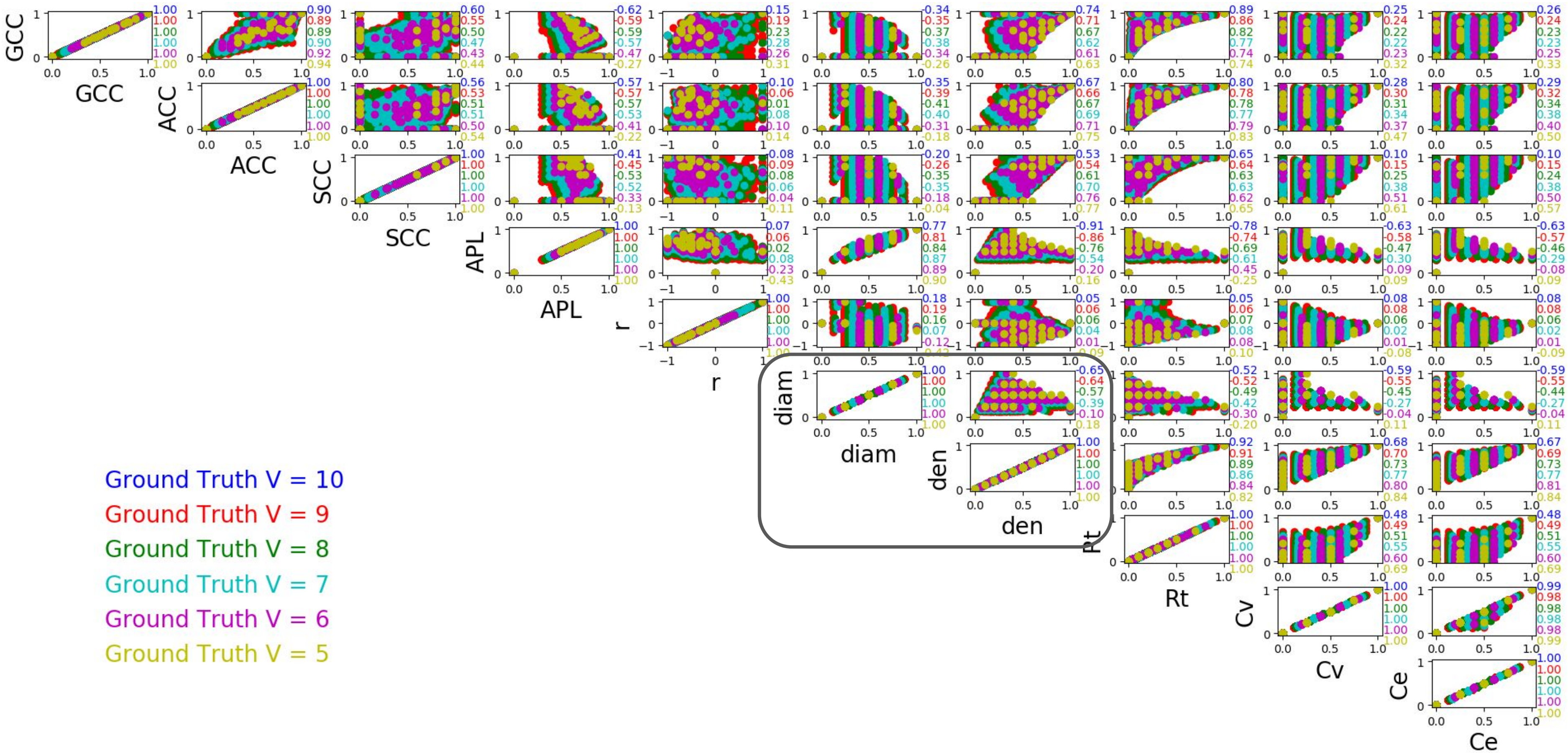


ground truth for |V|=9 and the results are different...

Ground Truth Data for Small |V|



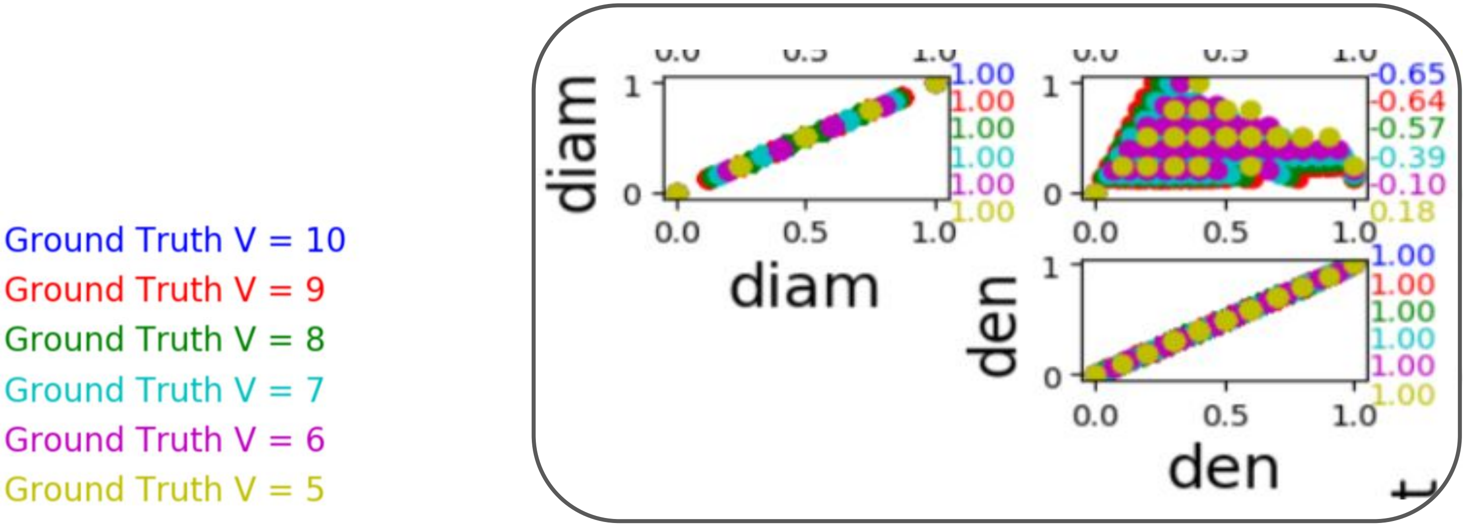
Ground Truth Data for Small |V|



Ground Truth Data for Small |V|

Let’s look at one of these more carefully

As |V| grows the correlation changes!



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?

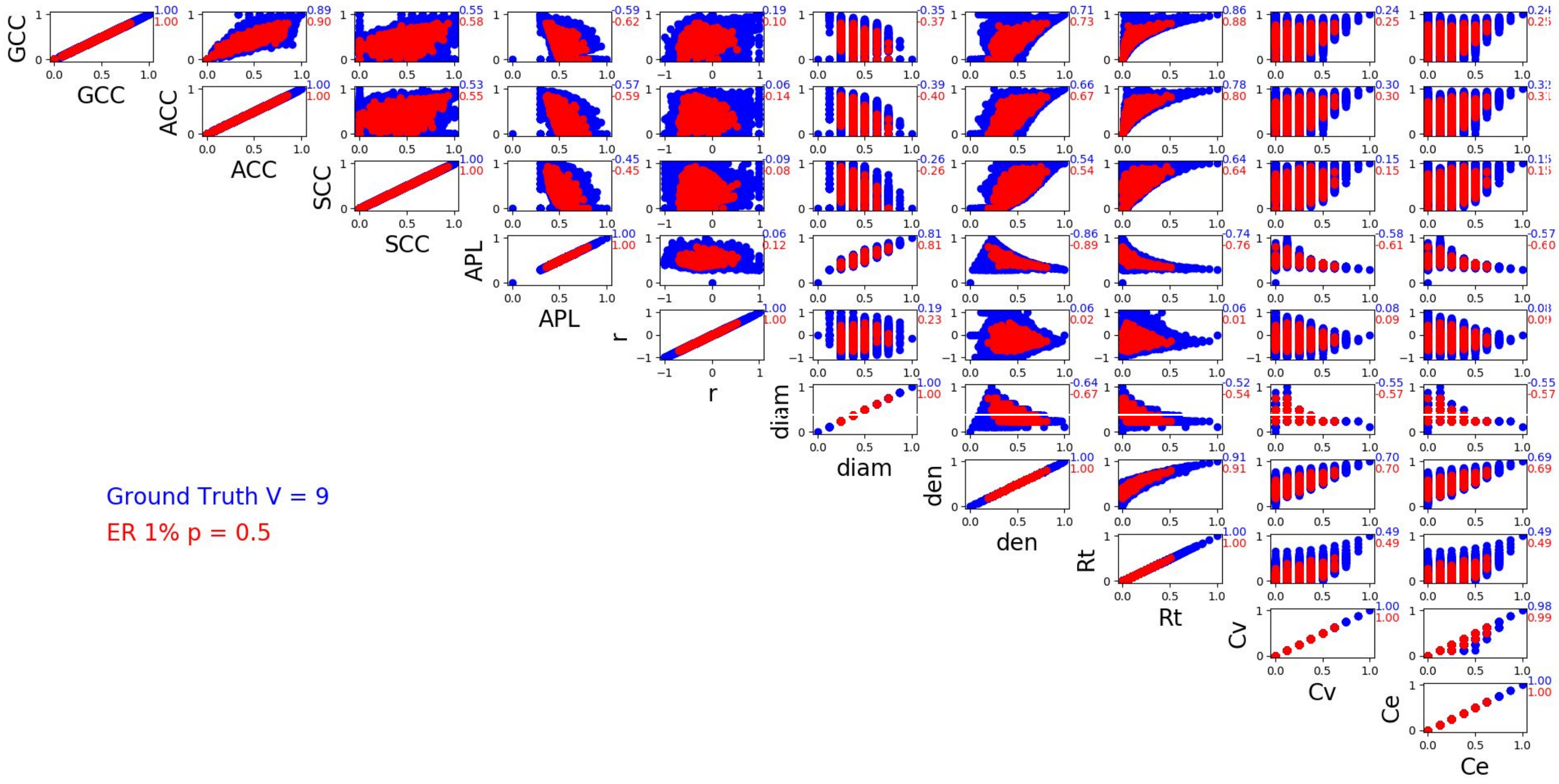
Desirable Generator Properties

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?

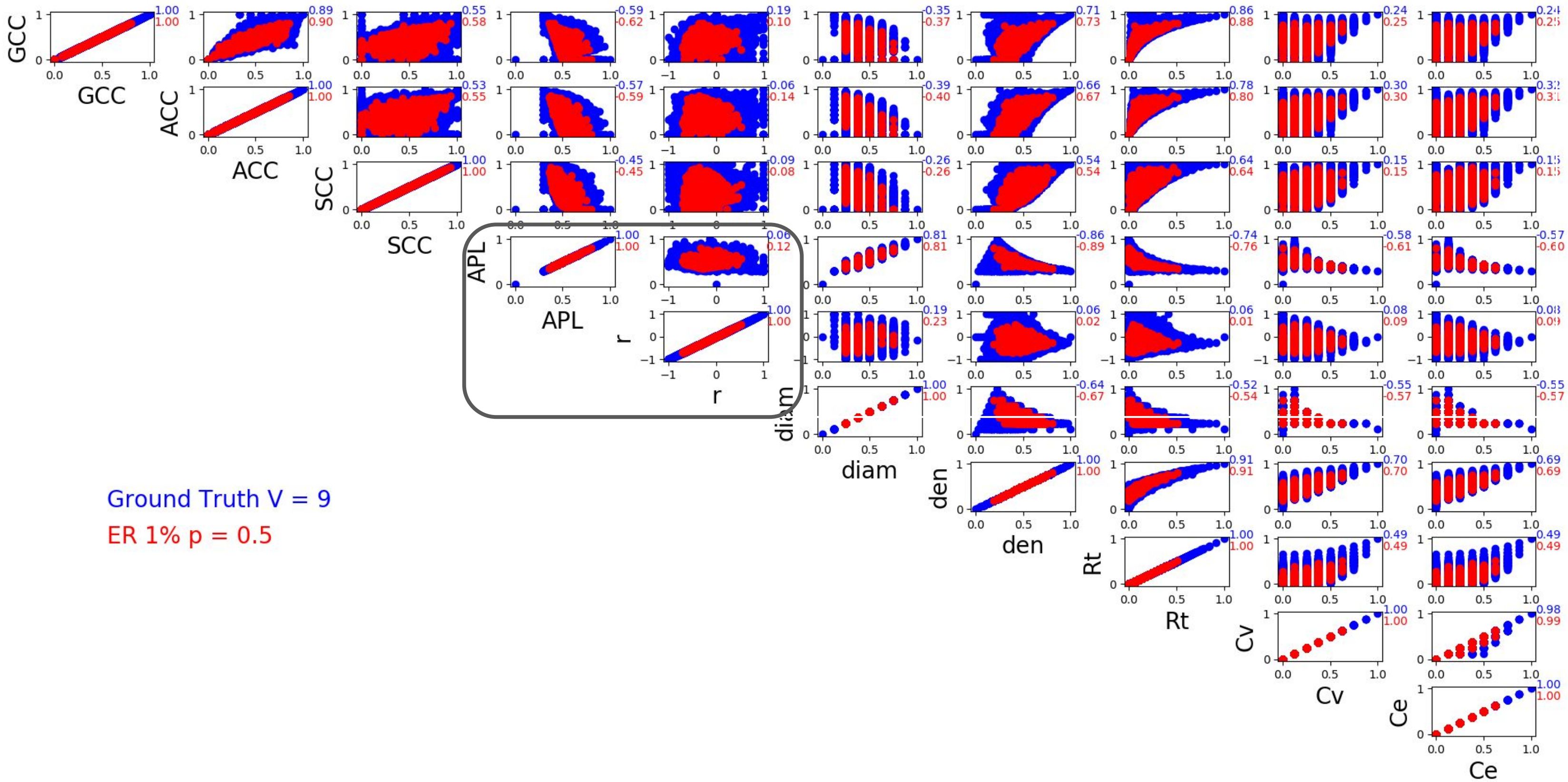
Graph Generator Representativeness

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



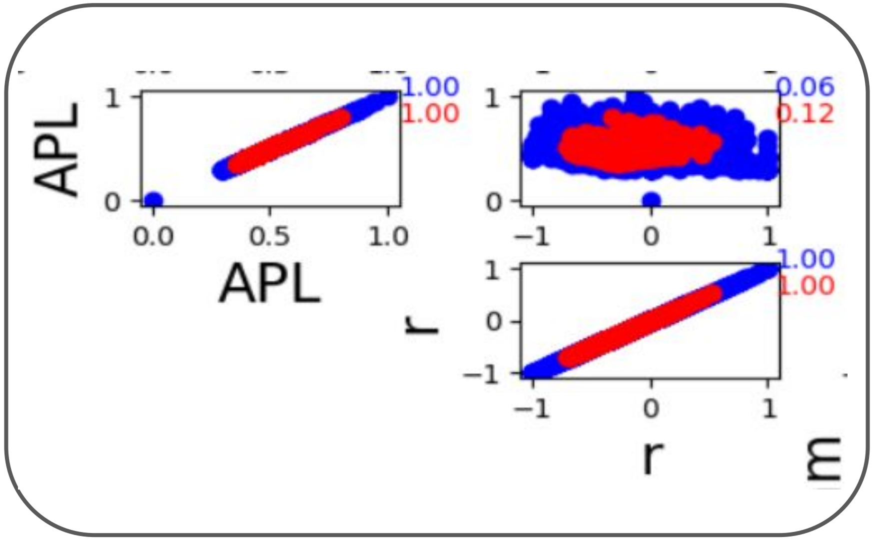
Graph Generator Representativeness

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



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 Coverage

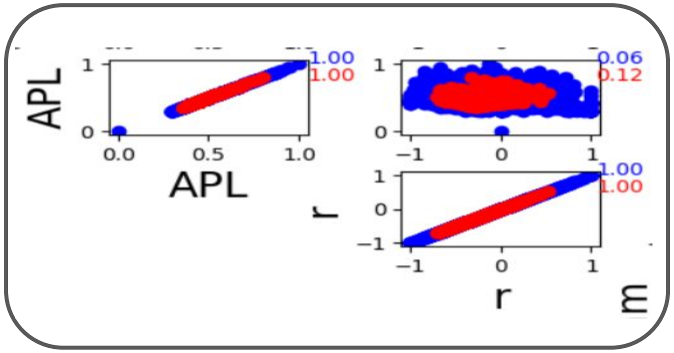
We measure how well a graph generator **covers** the range of values in the ground truth data by comparing the volumes of the generated data and the ground truth

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

WS model	BA model	ER model p = 0.5	ER model p~Uniform	ER model p~Population	Geometric model
22.04%	0.10%	0.98%	90.83%	12.37%	83.87%

Graph Generator Performance?

No graph generator is good at **representing** the ground truth and **covering** it well



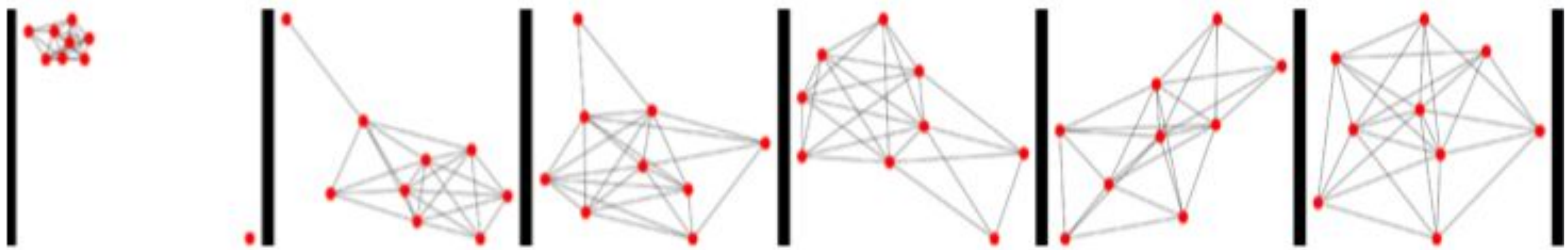
WS model	BA model	ER model p = 0.5	ER model p~Uniform	ER model p~Population	Geometric model
22.04%	0.10%	0.98%	90.83%	12.37%	83.87%

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*

Same Stats, Different Graphs

We can generate graphs with fixed set of statistics that vary in another statistic:



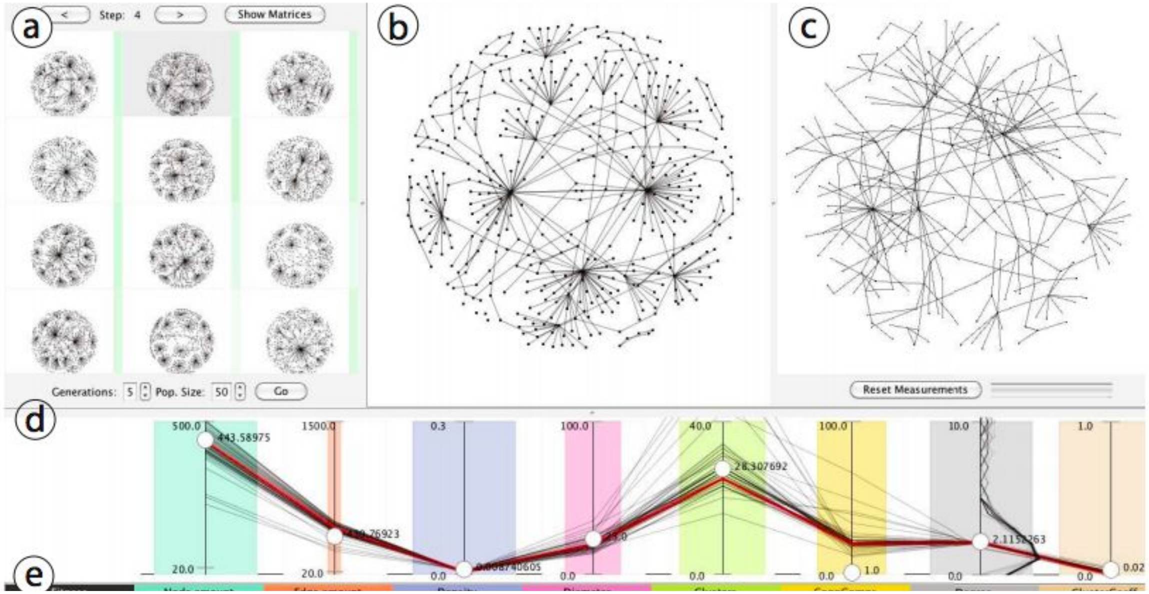
|V| = 9, SCC ∈ (0.75, 0.85), ACC ∈ (0.75, 0.8), r ∈ (-0.3, -0.2), Rt ∈ (0.35, 0.45), and we can find graphs in 6 out of 8 buckets for *connectivity*

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
4. U. Soni, Y. Lu, B. Hansen, H. Purchase, S. Kobourov, R. Maciejewski, The perception of graph properties in graph layouts, EuroVis, 2018



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?